JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

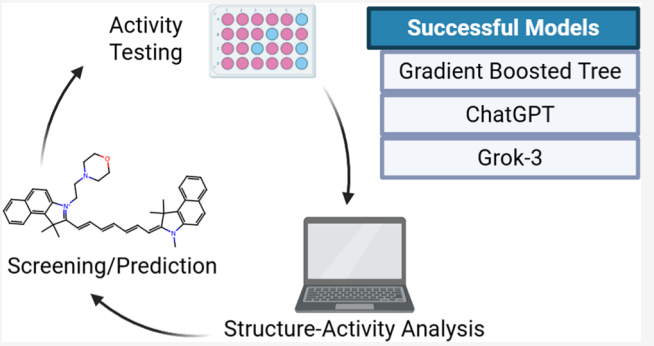
Identifying Structure–Activity Relationships for Cyanine-Derived Antibiotics Using Machine Learning and Commercial Large Language Models

Alexander Lathem, Angela Medvedeva, Ana Luisa L. Mendes dos Santos, Bowen Li, Tengda Si, Anatoly B. Kolomeisky, and James M. Tour*

Cite This: https://doi.org/10.1021/acs.jcim.5c01321 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Understanding the structure–activity relationship (SAR) of antibiotic scaffolds is crucial for the development of antibiotics to counter the growing crisis of antimicrobial resistant bacteria. However, an overwhelming space of structural features impairs a comprehensive understanding of the mechanism of action for potential antibiotic candidates. In this study, antibacterial data of a set of newly synthesized cyanine molecules are analyzed with both traditional machine learning (ML) and commercially available large language models (LLMs) to elucidate the SAR. Some LLMs, particularly Grok-3 Think and ChatGPT o1, outperform the traditional ML classifiers, and both approaches highlight positive charges and lipophilicity as key properties for effective cyanine antibiotics.



INTRODUCTION

Since their discovery in the early 20th century, antibiotics have transformed modern medicine and saved countless lives, yet this achievement is increasingly threatened by the global rise of antimicrobial resistance (AMR). Today, AMR represents one of the most urgent global health crises, with drug-resistant bacterial infections causing over 1 million deaths annually.[1] Without effective interventions, this burden is projected to reach between 2 and 10 million deaths per year by 2050,[1,2] with 39 million cumulative deaths directly attributable to AMR between 2025 and 2030. The economic impact is equally devastating, with AMR-related healthcare costs exceeding $55 billion annually in the United States alone.[3] Despite this pressing need, antibiotic development remains prohibitively resource-intensive, typically requiring 10–15 years and over $1 billion per approved compound, with success rates below 1% compared to 10–15% in other therapeutic areas.[4]

The discovery of new structural classes of antibiotics is particularly challenging, as evidenced by the 38 year gap between the introduction of fluoroquinolones in 1962 and oxazolidinones in 2000.[4] One strategy to accelerate antibiotic discovery is to revisit historically overlooked molecular scaffolds using modern computational tools. Cyanine dyes represent a compelling example of these overlooked scaffolds. Known primarily for their optical properties and applications in near-infrared biomedical imaging, exemplified by the Food and Drug Administration-approved dye Indocyanine Green (ICG),[5] cyanines were first studied for antimicrobial activity in the 1920s but were largely abandoned as therapeutic candidates due to nonspecific mechanisms and limited understanding of the structure–activity relationship.[6−8]

By leveraging machine learning (ML) and artificial intelligence (AI), researchers can now explore these overlooked scaffolds with high efficiency. AI approaches have demonstrated remarkable success in drug discovery, with AI-discovered molecules showing 80–90% success rates in Phase I clinical trials—substantially higher than historical industry averages.[9,10] These technologies enable rapid virtual screening of millions of compounds, dramatically reducing time and resources compared to traditional experimental methods. AI has transformed drug discovery across therapeutic areas, including cancer (optimizing kinase inhibitors with specific selectivity profiles),[11] neurodegenerative diseases (identifying compounds that cross the blood–brain barrier),[12] and antiviral development (designing molecules targeting viral proteases).[13] In the antibiotics discovery field, deep learning models have identified new structures such as halicin by screening over 107 million molecules.[14] While deep learning approaches have

Received: July 24, 2025
Revised: October 27, 2025
Accepted: November 3, 2025

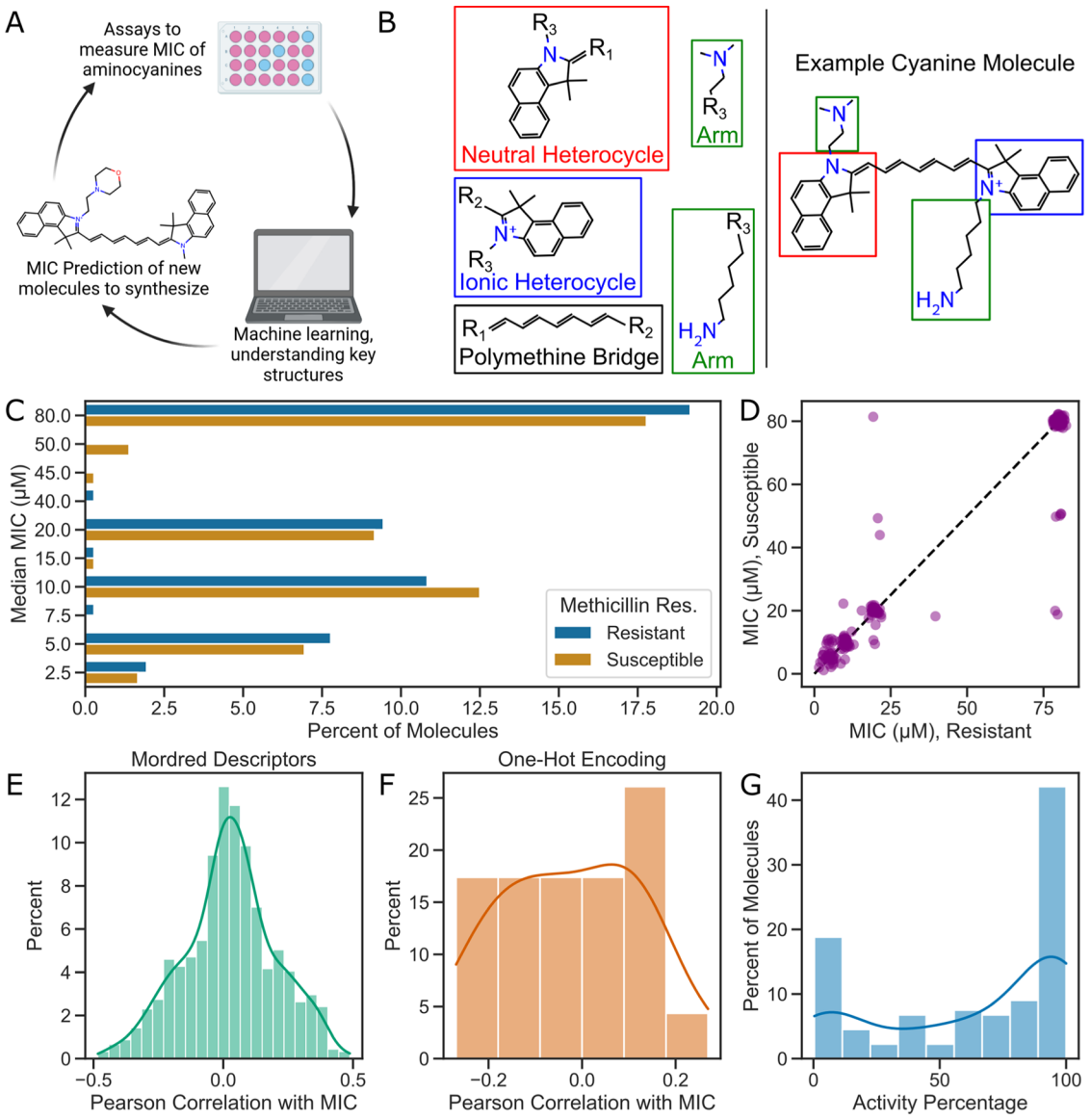


Figure 1. Overview of the study workflow and data set characterization. (A) Schematic overview of the project pipeline, including synthesis of cyanine-derived molecules, minimum inhibitory concentration (MIC) testing, ML and LLM-based prediction of activity, suggestion of new candidate molecules, and iterative refinement for future experimental validation. (B) Illustration of the major components of an example cyanine molecule: two heterocycles, a polymethine bridge, and the functionalized carbon chain (arm) extending from the nitrogen atom or ion. (C) Histogram of all MIC values observed across the data set, disaggregated by the methicillin resistance status of the *S. aureus* strains (resistant versus susceptible). (D) Scatterplot comparing the median MIC of each molecule against methicillin-resistant versus methicillin-susceptible *Staphylococcus aureus* strains. (E) Distribution of Pearson correlation coefficients between each of the 1133 Mordred descriptors and median MIC. (F) Distribution of Pearson correlation coefficients between each one-hot encoded structural subfeature and median MIC. (G) Histogram of activity percentage for all molecules, based on the binary classification threshold of MIC <80 μM.

enabled the identification of possible therapeutic candidates, these methods typically require specialized expertise in computational chemistry and complex feature engineering, limiting their accessibility to many research groups.

In this study, we developed an understandable, accessible approach for screening and optimizing cyanine-based antimicrobials against *Staphylococcus aureus*, a pathogen responsible for approximately 119,000 bloodstream infections and 20,000 deaths annually in the United States.[15] Methicillin-resistant *S. aureus* (MRSA) is particularly concerning, with resistance rates exceeding 50% in many regions.[16]

We investigated whether commercially available large language model (LLM)-based AI systems can predict the antibacterial activity of cyanine derivatives without requiring specialized computational infrastructure or expertise. We compared the performance of ChatGPT o1, o3 mini, o3 mini high; Claude 3.0 Opus, 3.5, Sonnet; Grok-2, Grok-3, Grok-3 Think; Google Gemini 2.0 Flash Thinking (FT); and Deepseek R1, each prompted using only the simplified molecular-input line-entry system (SMILES) notation,[17] against a gradient boosting classifier trained on 24 simple one-hot (a method of representing categorical data as numerical values, suitable for use in ML models) encoded features. For our analysis, we used a data set of 143 newly synthesized cyanine derivatives, each comprehensively tested against a panel of 163 *S. aureus* clinical isolates, including 110

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

methicillin-resistant strains, representing one of the most extensive structure−activity relationship (SAR) studies of this antimicrobial scaffold to date.

Using a traditional ML framework, we identified key SARs associated with antimicrobial activity. These included the presence of dimethylamine moieties on flexible linkers, extended polymethine bridges, and the absence of bulky substituents, which were found to diminish the activity. A gradient-boosted tree classifier trained on these features demonstrated robust predictive performance ($F_1 \approx 0.8$ across 1000 bootstrap iterations), underscoring the capacity of relatively simple molecular descriptors to capture the nuanced SAR within this compound class.

Furthermore, LLM-based systems Grok-3 Think and ChatGPT o1 achieved F_1 scores above 0.8 when classifying compounds as active or inactive, exceeding the F_1 score of the gradient-boosted classifier. Notably, the LLMs identified the same key SAR found by the ML framework including amines, aromatic heterocycles, and unmodified conjugated polymethine chains. The LLMs provided natural language explanations of the SAR that were consistent with the experimental validation. Importantly, several of the identified cyanine derivatives demonstrated broad-spectrum activity against not only *S. aureus* but also other clinically important Gram-positive pathogens, including multiple *Streptococcus* species.

This work demonstrates that LLMs can efficiently guide the exploration of chemical spaces and accelerate the discovery of antimicrobial compounds without requiring specialized computational expertise. The approach enables rapid iteration between virtual screening and experimental validation (Figure 1A), reducing the time and resources required for early-stage drug discovery. Beyond antimicrobials, this methodology may be applicable to a wide range of therapeutic areas where SARs are complex, and traditional computational approaches remain inaccessible to many researchers.

RESULTS AND DISCUSSION

Data Set Generation and Feature Representation. To evaluate the ability of traditional ML and LLM-based AI chatbots to predict antibiotic activity, we constructed a data set of 143 cyanine-derived molecules synthesized in our laboratory. These cyanine molecules share key structural motifs (Figure 1B): all have an extended polymethine bridge capped by a nitrogen (one neutral and one ionic). Each nitrogen is contained in a heterocyclic group, and a functionalized carbon chain extends from each nitrogen as well. This carbon chain, or "arm," can take various forms, from a methyl group to a 13-atom chain functionalized by *tert*-butyl carbamate. The symmetry of the cyanines varies dramatically, with some being completely symmetric except for the cationic nitrogen and others having entirely different heterocyclic groups.

Each molecule was tested against a panel of over 160 *S. aureus* strains including both methicillin-susceptible and methicillin-resistant clinical isolates. Antibacterial activity was assessed using a standard broth microdilution assay to determine the MIC, defined as the lowest concentration of a compound that inhibited bacterial growth after 18−24 h of incubation. Because the assay employs a 2-fold serial dilution series, the MIC values are inherently discrete, typically falling within a fixed set of concentrations, in this case: 0.625, 1.25, 2.5, 5, 10, 20, 40, and 80+ μM, denoting that the MIC is higher than the highest compound concentration tested of 80 μM.

To evaluate whether the methicillin resistance status of *S. aureus* strains influenced the activity of the tested compounds, we compared MIC distributions for each molecule across methicillin-susceptible and methicillin-resistant strains (Figure 1C). The Wilcoxon signed-rank test,[18] which is more appropriate than Student's t test for non-normal distributions, yielded a value of 157.0 ($p = 0.44$), indicating no statistically significant difference between susceptible and resistant MIC distributions. Figure 1D displays the median MIC values for each compound in both strain categories with a 5% jitter applied to improve visualization. The Pearson correlation coefficient for a compound's median MIC against susceptible versus resistant strains was 0.958 ($p = 1.3 \times 10^{-98}$).

Both statistical tests suggest that methicillin resistance has a minimal influence on the efficacy of the tested compounds, supporting the hypothesis that their mechanism of action differs from that of β-lactam antibiotics such as methicillin.

We then explored three complementary strategies for molecular feature representation: RDKit-based substructure fingerprints, Mordred physicochemical descriptors, and a manually curated one-hot encoding of key substructures. These representations reflect a range of chemical abstraction levels, from molecular fragments to computed chemical properties.

RDKit substructure fingerprints consist of 2048 bit binary vectors capturing the presence or absence of common molecular fragments. Mordred descriptors, in contrast, provide 1613 real-valued physicochemical and topological features derived from SMILES strings. After filtering out non-numeric and missing values, 1133 Mordred features were retained for analysis.

Next, we evaluated the extent to which individual Mordred descriptors correlate with the antimicrobial activity. Figure 1E shows the distribution of Pearson's correlation coefficients between each of the 1133 descriptors and MIC values. None exceeded an absolute correlation of 0.5, indicating that no single physicochemical property accounted for the antibacterial potency across this chemical class.

To test a more interpretable and lower-dimensional feature space, we manually selected 24 substructures based on relevance to the cyanine scaffold and encoded them as binary features (present = 1, absent = 0). These substructures (shown in Figure S1) were variants of the key functional groups that comprise a cyanine: six possible polymethine bridges, four possible heterocyclic groups on each end (one ionic and one neutral), and the ten most common alkyl "arms" that could be attached to the nitrogen on either end of the bridge. This simplified one-hot encoding offers a balance of interpretability and dimensionality control, especially valuable for small data sets. Pearson's correlation analysis again revealed that none of these binary features strongly correlated with MIC either (Figure 1F), suggesting that activity is not attributable to individual features but likely depends on nonlinear and combinatorial relationships of different molecular features.

Activity labels were assigned based on the median MIC across all tested strains: molecules with a median MIC below 80 μM were labeled "active," while those at or above this threshold were considered "inactive." This binarization yielded a roughly balanced data set of 60% active and 40% inactive compounds (Figure 1G).

To assess the diversity of the chemical space captured by the library, we calculated the pairwise Tanimoto similarity using the RDKit Tanimoto similarity function,[19] where each

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

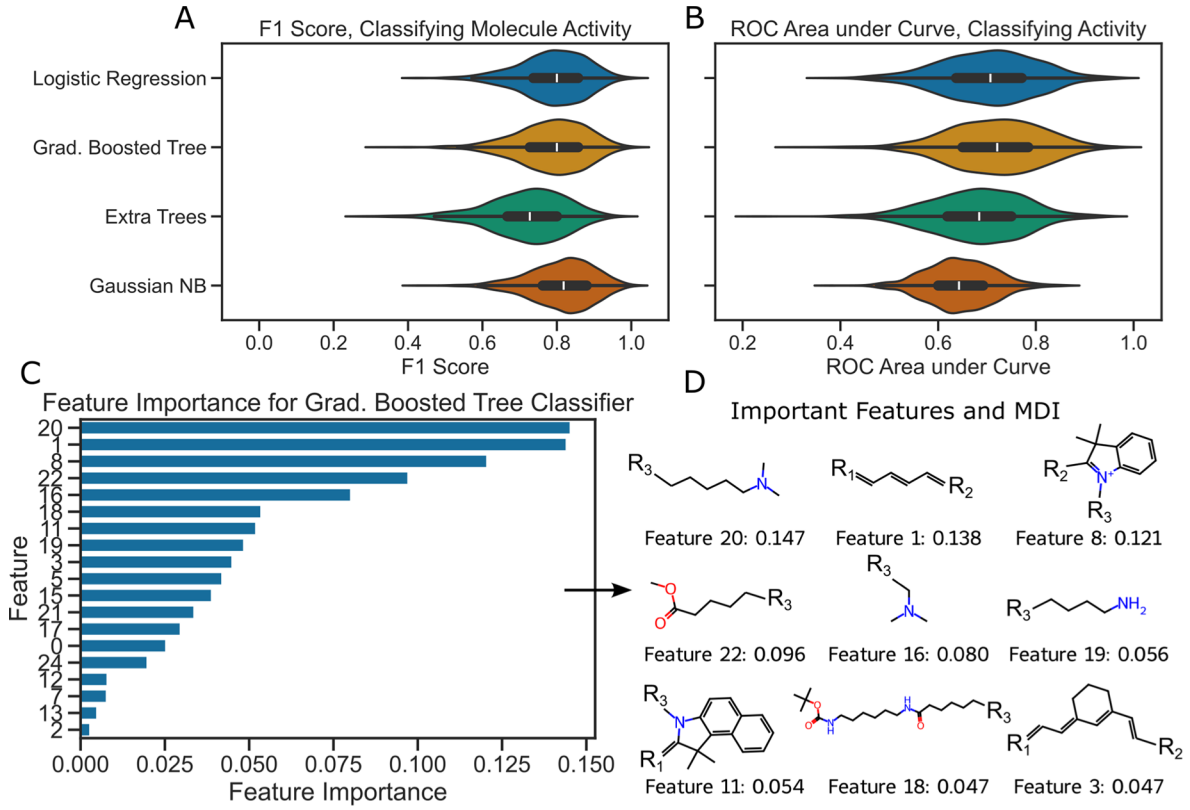


Figure 2. Performance metrics for traditional ML classifiers and feature importance analysis. (A) Distribution of F_1 scores across 1000 independent train/test splits for four classifiers: Gradient Boosted Trees, Gaussian Naive Bayes, Logistic Regression, and Extra Trees. Each model was trained on 90% of the data set and evaluated on a held-out set of 14 molecules (10%). (B) Distribution of the area under the receiver operating characteristic (ROC) curve (ROC AUC) for each classifier across the same iterations, capturing the balance between true-positive and false-positive rates. (C) Top 20 one-hot encoded molecular substructures ranked by their importance in the Gradient Boosted Tree classifier, as measured by normalized mean decrease impurity (MDI), reflecting the reduction in class impurity at decision tree nodes. (D) Chemical structures of the nine most influential substructures shown in panel (C), illustrating the key functional motifs associated with antibacterial activity.

structure was converted to a 2048 bit Morgan fingerprint vector with radius 2. The distribution of pairwise Tanimoto similarities is shown in Figure S2; the mean similarity was 0.389, and the standard deviation was 0.158. These Tanimoto similarity values are comparable to those reported for broad screening collections (0.27−0.57),[20] demonstrating that our data set is structurally diverse and suitable for evaluating model generalization.

Figure S3 shows the two-component principal component analysis (PCA) plot of the cyanine structures colored by the activity label according to the two feature sets: Mordred descriptors and one-hot encoding. While no strict clustering was observed, both PCA plots show enough concentration of active structures in one part of the space to suggest the importance of structure on activity. As with Figure 1E,F, the distribution of structures in the PCA space indicates a complex SAR that ML may elucidate.

ML Binary Classification. Given the weak individual correlations between single features and MIC values (Figure 1E,F), we hypothesized that antimicrobial activity might instead emerge from nonlinear or combinatorial interactions among features. To capture these more complex relationships, we framed the task as a binary classification problem: predicting whether a given molecule is "active" (median MIC < 80 μM) or "inactive" (median MIC ≥ 80 μM). The distribution of MICs across the data set is bimodal (Figure 1C), and most molecules were either consistently active or inactive across the strain panel (Figure 1G). As such, these observations justify a binary labeling approach and support the use of supervised classification techniques.

We trained and evaluated four widely used binary classifiers on the data set: Logistic Regression, Gradient Boosted Trees, Extra Trees, and Gaussian Naive Bayes. A detailed explanation of these classifiers is included in the Methods section, and the default hyperparameters were used for each (see Supporting Information, "Machine Learning Classifier Hyperparameters and Tuning Ranges").

Each classifier was evaluated using repeated train/test splitting: 90% (129) of the molecules were used for training and 10% (14 molecules) were held out for testing. This process was repeated 1000 times to estimate variability. The numbers of true positives (truly active molecules labeled "active"), false positives (truly inactive molecules labeled "active"), true negatives (truly inactive molecules labeled "inactive"), and false negatives (truly inactive molecules labeled "active") were recorded. These numbers were then used to calculate three key metrics: precision, recall, and F_1 score.[21,22]

Precision is defined as

$$p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

where TP is the number of true positives, and FP is the number of false positives. Recall is defined as

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

where FN is the number of false negatives. Then, the F_1 score is defined as the harmonic mean of the precision and recall

$$F_1 = \frac{2rp}{(r + p)} = \frac{2\mathrm{TP}^2}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

Because the F_1 score is the harmonic mean of both precision and recall, only a model that can correctly label as many active molecules as possible while simultaneously rejecting inactive molecules will have a high F_1 score. This makes the metric robust against imbalanced data sets and prevents trivial classifiers (such as always labeling a molecule as "inactive") from scoring well, as can occur with an ordinary accuracy score.

Figure 2A shows the distribution of F_1 scores for each classifier across all iterations. Among the models tested, the Gradient Boosted Tree and Gaussian Naive Bayes classifiers exhibited the strongest performance, achieving mean F_1 scores exceeding 0.8 with lower variance compared with the Logistic Regression and Extra Trees classifiers.

The ROC is an alternative to the F_1 score for binary classification that tracks how the false-positive rate and true-positive rate relate to each other as a classifier becomes more sensitive to true positives. The plots of the mean and standard deviation for all 1000 ROC curves for each classifier are shown in Figure S4, and the distribution of area under the ROC curve is shown in Figure 2B. The Gradient Boosted Tree classifier had the highest ROC area under the curve on average, while the Gaussian Naive Bayes classifier had the lowest. These results suggest that the Gradient Boosted Tree offers the most reliable predictive performance across classification metrics.

The relatively small size of the data set (143 molecules) likely contributed to model performance variability, since the presence or absence of key molecular motifs in the training data set could greatly affect the classifier's ability to generalize. Nevertheless, the Gradient Boosted Tree model demonstrated robust performance, even when some critical molecules were missing from the training data.

The Matthews correlation coefficient (MCC) can be used as an additional scoring metric for binary classification which is robust under mild class imbalance,[23,24] as in the case of the 40−60% split in this data set. Figure S5 shows the distribution of MCC for the same four classifiers shown in Figure 2A,B. While the MCC is generally lower for all classifiers than the F_1 score, the relative MCC of the classifiers is similar to the relative F_1 score. Despite the class imbalance, the Gradient Boosted Tree classifier remains the best overall classifier using all three performance metrics.

Hyperparameter tuning was attempted for all four classifiers using the *BayesSearchCV* function in scikit-optimize, but it did not improve the performance and was therefore not pursued further. The failure of hyperparameter tuning across different models is explained by the size of the data set. Bayesian optimization methods such as *BayesSearchCV* rely on iterative sampling of the hyperparameter space, constructing a probabilistic surrogate model (typically a Gaussian process) to estimate the performance landscape and balance exploration versus exploitation.[25,26] This optimization algorithm generally requires numerous evaluations to model the response surface accurately. With limited or noisy data, the surrogate model may fail to converge toward an optimum.[27,28] In small data sets, cross-validation introduces high variance in performance estimates, which introduces the kind of noise that interferes with the surrogate model's convergence.[29,30] Given our modest data set (n = 143 molecules) and a 90/10 train/test split, each cross-validation fold contained very few samples (out of the 129 training-set molecules used for hyperparameter optimization, ∼14 molecules in the test set using 10-fold), leading to high variance and poor convergence of the surrogate model.

Because of the superior classification performance of the Gradient Boosted Tree, we investigated which features contributed most to its decision-making. Feature importance was measured using MDI, a built-in metric for tree-based models that quantifies how often and effectively each feature splits the data. The MDI is the average decrease in Gini impurity (G) when a given feature is used to bifurcate the data set. The Gini impurity[31] is defined as

$$G = \sum_{i=1}^{C} p(i)(1 - p(i)) \tag{4}$$

where C is the total number of classes, and $p(i)$ is the probability of selecting a point from the data set with class i. Figure 2C shows the top 20 features ranked by importance, assessed as the MDI, in the Gradient Boosted Tree classifier. The structures of the top nine features and their corresponding importance are shown in Figure 2D.

A high MDI score indicates that a feature plays an important role in classification but does not reveal whether the feature is associated with increased or decreased activity. In other words, a feature may contribute positively or negatively to activity, depending on its structural context. To clarify this directionality, we calculated the Pearson correlation between each important feature and MIC (Figure 1E). The top-ranked features and their correlations with the MIC are shown in Figure S6.

The most important features included a dimethylamine group attached to an unmodified, long alkyl chain (Feature 20), a five-carbon polymethine bridge (Feature 1), and an indole heterocycle. Several other top-ranking features represented charged substituents located at the termini of long side arms (e.g., Features 8, 22, and 19). These were typically associated with lower MIC values, suggesting that extended conjugation and localized positive charge are critical for the activity. In contrast, sterically bulky groups, particularly a large, branched arm (Feature 18), were strongly associated with inactivity. Some features, such as a benzoindole ring (Feature 11), had a more ambiguous relationship with activity, possibly due to context-dependent effects on the molecular scaffold. These findings support a rational design framework for future compound optimization: molecules should preserve charge (e.g., through cationic dimethylamine groups), maintain polymethine bridge length to maximize cation delocalization, and avoid sterically hindered substituents that can interfere with the cell target interaction or binding.

The key structural features identified by our Gradient Boosted Tree model, notably the importance of localized positive charge, extended conjugation of the cationic methine bridge, and conformational flexibility, strongly suggest a membrane-targeting mechanism of action. Cationic amphiphilic compounds, such as the active cyanine derivatives used in this study, typically interact electrostatically with the negatively charged phospholipid bilayer of Gram-positive bacteria such as *S. aureus*. The presence of dimethylamine

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Overview of the 14 Commercial LLMs Evaluated for Molecular Classification in This Study[a]

service	model	company	access	date used	success?
ChatGPT	4o	OpenAI	free[b]	2025/02/05	no
ChatGPT	o1	OpenAI	subscription	2025/02/12	**yes**
ChatGPT	o3 mini	OpenAI	subscription	2025/02/12	**yes**
ChatGPT	o3 mini-high	OpenAI	subscription	2025/02/12	**yes**
Meta AI		Meta	free	2025/02/05	no
Gemini	2.0 Flash	Google	free	2025/02/12	no
Gemini	2.0 Flash Thinking (FT)	Google	free	2025/02/12	**yes**
Grok	Grok-2	xAI	deprecated	2025/02/12	**yes**
Grok	Grok-3	xAI	free	2025/02/19	**yes**
Grok	Grok-3 Think	xAI	subscription	2025/02/19	**yes**
Grok	Grok-3 DeepSearch	xAI	subscription[b]	2025/02/19	no
Deepseek	R1	Perplexity/Deepseek	subscription[c]	2025/02/12	**yes**
Claude	3.5 Sonnet	Anthropic	subscription[b]	2025/02/12	**yes**
Claude	3 Opus	Anthropic	subscription[b]	2025/02/12	**yes**

[a]The "success" column indicates whether the model was able to complete the activity classification task successfully using the provided molecular input and prompt. [b]While this model is available with a free account, a subscription appreciably increases the usage allowed per session. [c]This model is free with some services but is only available on Perplexity with a subscription.

groups (Feature 20) and other positively charged moieties likely facilitate initial binding to anionic lipid headgroups, while the extended polymethine bridge (Feature 1) may enable membrane insertion and disruption through π-stacking and hydrophobic interactions. Conversely, bulky substituents (Feature 18) appear detrimental to activity, potentially by sterically hindering optimal membrane association.

This proposed mechanism is consistent with previous studies describing the behavior of cyanine dyes in biological membranes. It is known that, in eukaryotes, cationic cyanines, such as Cy3 and Cy5 derivatives, accumulate in mitochondria due to their positive charge and lipophilic nature, which enables electrostatic and hydrophobic interactions with the negatively charged mitochondrial membrane surface.[32−34] The mitochondrial membranes share several biophysical properties with bacterial membranes, including overall anionic lipid composition, which explains the analogous behavior of these dyes in bacterial systems. Structural modifications to cyanines that increase the cationic character, such as the incorporation of 1,4-diazabicyclo[2.2.2]octane (DABCO) moieties[35] or the introduction of lipophilic side chains (e.g., alkyl groups),[36] have been shown to increase membrane affinity and enhance cell penetration, while excessive steric bulk, such as *tert*-butyl substituents or double-conjugated fluorophores, has been associated with decreased cell uptake and targeting efficiency.[37] The presence of cationic moieties (e.g., arginine and lysine) is critical for initial electrostatic interactions of antimicrobial peptides with negatively charged bacterial membranes.[38] In the case of quaternary ammonium compounds such as benzalkonium chloride, the balance between hydrophobic alkyl chains and cationic headgroups is critical for membrane insertion.[39] In contrast, cyanines with reduced steric bulk exhibit greater flexibility, enabling deeper insertion into lipid bilayers.[40] These observations are consistent with the structural features that we identified in our model, reinforcing the hypothesis that cyanine-derived antibiotics likely exert their effects by association with and disruption of the cell membrane.

Our results extend and deepen the previous analysis of the SAR of cyanine-based antimicrobials. While cyanine dyes have been extensively characterized in terms of their photophysical properties and imaging, their antimicrobial potential, particularly beyond their use as photosensitizers, remains poorly understood. Mohamed and AbuEl-Hamd (2016) investigated a small number of bis-coumarin cyanine dyes and tested them for their antimicrobial activity against a limited number of organisms.[41] They focused primarily on the influence of specific metal coordination complexes and chromophore variants rather than on a systematic SAR approach. Similarly, Prakash et al. (2023) synthesized cyclohexene-based heptamethine−cyanine dyes containing sulfur and selenium atoms and evaluated their antimicrobial photodynamic therapy (APDT) mainly against *S. aureus* and *Escherichia coli*.[42] Although the authors investigated modifications at the heterocyclic ends, their SAR conclusions were limited to a few variants and largely focused on their performance as near-IR light-activated photosensitizers. Here, we comprehensively analyze 143 structurally diverse cyanine derivatives by quantifying the contributions of terminal dimethylamine groups, revealing the negative influence of steric bulk and defining the optimal charge distribution across the cyanine scaffold. Moreover, our computational approach enables the SAR investigation on cyanines at a larger scale compared to previous studies, which were typically limited to a few analogs, and provides clearer molecular design principles for scaffold optimization.

Binary Classification by Commercial LLMs. Given the good performance of traditional ML models and their ability to identify key structural motifs relevant to antibacterial activity, we sought to explore whether general-purpose LLMs could replicate or even exceed these capabilities. LLMs are increasingly used in scientific applications because of their ability to interpret diverse inputs, including text, code, and formatted strings such as the SMILES, and to compute upon them using data gained from pretraining using massive collections of written text. The process of training LLMs to conduct binary classification in this way is known as "in-context learning" (ICL)[43−45] and has been demonstrated in a variety of fields including drug toxicity prediction[46] and tumor detection in medical imaging.[47] The accessibility, flexibility, and speed of LLMs make ICL an attractive tool for early-stage drug discovery, particularly for hypothesis generation and low-barrier molecule screening. In this study, we evaluated whether LLMs could (1) predict the antibacterial activity of different cyanines, (2) propose new candidate molecules, and (3) identify structural features that are important for antibiotic

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

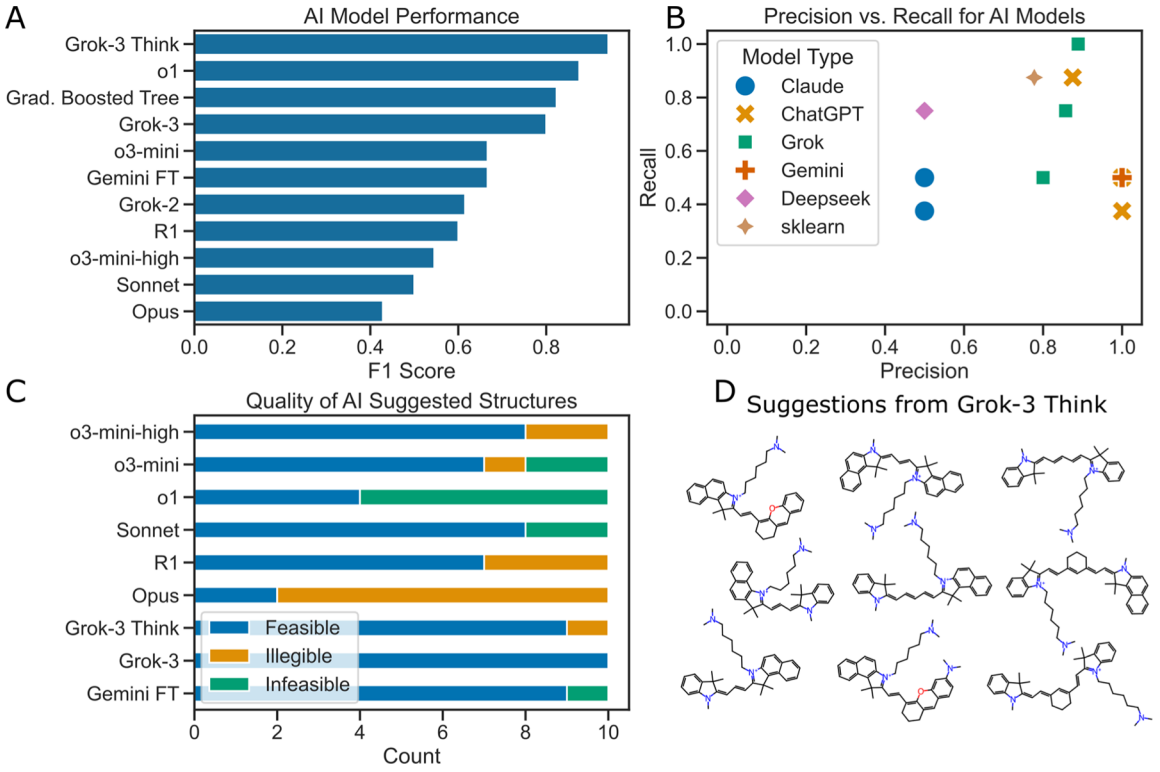


Figure 3. Evaluation of the commercial LLM performance in molecular classification and compound generation. (A) F_1 scores for all LLMs compared with the traditional Gradient Boosted Tree classifier on the task of predicting binary antibiotic activity labels ("active"/"inactive") for 14 unseen test compounds. (B) Precision versus recall for each model. Color and shape indicate model source or architecture (e.g., brand or framework). The scikit-learn Gradient Boosted Tree model serves as a traditional ML baseline. (C) Assessment of the chemical feasibility of LLM-generated molecules. Suggested SMILES were categorized as "feasible" (synthetically plausible), "infeasible" (implausible or chemically invalid), or "illegible" (unparseable due to incorrect formatting). (D) Chemical structures of the nine feasible molecules proposed by Grok-3 Think, the top-performing LLM.

activity. We also compare the prediction accuracy of these LLMs to the traditional Gradient Boosted Tree classifier, which requires far more training to implement and is therefore less accessible to nonexperts in the ML field.

To this end, we tested 14 distinct models from six AI companies during a two-week period in February 2025. These services undergo near-constant updates and expansions, meaning that the results and functionality of these models could change drastically over time.

All models received identical prompts according to a simple workflow illustrated in Figure S7. Each LLM was first presented with a training data set of SMILES-formatted molecular structures and the corresponding binary activity labels (i.e., "active" and "inactive"). The models were then asked to examine these training data and identify molecular features associated with activity. Next, they were given 14 previously unseen test molecules and asked to predict their activity classification (Figure S8). Finally, each model was prompted to design ten new molecules likely to be active based on insights gained from the previous analysis. The exact wording of the identical prompts given to each LLM is included in the Supporting Information, "LLM Binary Classification Scripts."

Table 1 shows the models with which binary classification was attempted. The table includes the name of the service, the specific model used, the company responsible for designing the model, the state of access for the model as of publication, the date the model was accessed, and whether the model could successfully complete the binary classification task.

Not all models were capable of the three tasks presented (predict, propose, and identify), even if the models were from the same service and company. This is likely because various services now have different specializations. Most LLM services have a generic conversational chatbot that is meant to synthesize responses to most questions using input data and Internet access. Some companies, however, have expanded their services to include "thinking" models that specialize in computation, coding, and logic (e.g., Grok-3 Think, Gemini FT, ChatGPT o1, o3 mini, and o3 mini-high). Other services specialize in deep research, focusing on finding answers to questions online and providing abundant sources (e.g., Grok-3 DeepSearch).

Ten of the 14 LLMs successfully completed the binary classification task. Figure 3A compares the F_1 scores of these LLMs to those of the traditional Gradient Boosted Tree classifier. The precision and recall for each model are shown in Figure 3B. Most models underperformed relative to the Gradient Boosted Tree (F_1 score of ~0.8), and only three LLMs exceeded an F_1 score of 0.7: Grok-3, ChatGPT o1, and Grok-3 Think. Of these, two LLMs, ChatGPT o1 and Grok-3 Think, exceeded the F_1 score of the Gradient Boosted Tree classifier. Grok-3 Think achieved the highest performance overall, with a recall of 1.0 and a precision of 0.9, while ChatGPT o1 achieved ~0.85 in both precision and recall. These metrics show that Grok-3 Think correctly labeled all active molecules in the test set, with ChatGPT o1 labeling one false negative. Both Grok-3 Think and ChatGPT o1 screened out all but one false positive.

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Structural Feature Attributions Reported by LLMs That Were Able to Complete the Binary Classification Task[a]

feature	o1	o3 mini	o3 mini-high	Gem-ini FT	Grok -2	Grok -3	Grok-3 Think	Perplexity R1	Claude 3.5 Sonnet	Claude 3 Opus
extended bridge	good	good	good	good	good	good	good	good	good	good
positive charge	good	good	good		good		good	good		good
arms (in general)	good				good					
bulky groups		bad	bad	bad	bad			bad	bad	bad
polarity/negative charge	bad	bad	bad	bad	bad	bad	bad	good	bad	
benzo-indoles				good		good	good			
arms with amines	good	good	good	good	good	good	good		good	
groups on bridge	bad			bad		bad	bad		bad	
modified rings	bad	bad	bad			bad	bad	bad		bad
neutral amides								bad		
symmetry									good	

[a]For each feature, a "good" label indicates that the model associated the feature with increased antibiotic activity, while "bad" indicates a negative association. Blank cells reflect features not mentioned by a specific model.

Because the test set of molecules in Figure S8 contained 6 inactive molecules and 8 active molecules, there was a slight class imbalance. The MCC was again calculated for the predictions of all LLM models and compared with the Gradient Boosted Tree classifier (Figure S9). The top-performing LLMs ranked the same under MCC as under F_1 score, but a few notable differences arise. Grok-3 had a slightly higher MCC than the Gradient Boosted Tree, and many of the lower-ranking models switched places in the MCC ranking. The Deepseek R1 model accessed through Perplexity was the eighth best by F_1 score (better than the three models) but worst under MCC.

All ten LLMs were also asked to generate ten new molecules that were likely to be active. The format of the suggestions provided by LLMs was always SMILES, although that was not explicitly requested in the prompt (see Supporting Information, "LLM Binary Classification Scripts"). This generative task required a form of creative synthesis and pattern abstraction that lies outside the capabilities of conventional ML models.

Each suggestion was categorized into one of three classes: (1) "feasible", denoting valid, synthetically plausible SMILES strings; (2) "infeasible", corresponding to syntactically valid SMILES that encoded molecules with unrealistic or chemically implausible features; and (3) "illegible", those with SMILES strings that were invalid or could not be interpreted using cheminformatics libraries such as RDKit. Figure 3C summarizes the performance of each model according to this classification scheme.

The quality of the suggested structures varied widely. The most successful generators were Grok-3 Think (9 feasible, 1 illegible), Gemini 2.0 FT (9 feasible, 1 infeasible), and Grok-3 (10 feasible). In contrast, ChatGPT o1 returned mostly infeasible structures. All LLM suggestions except those from Grok-3 Think are shown in Figures S10 and S11.

Grok-3 Think was not only the sole model to surpass the performance of the Gradient Boosted Tree under training/test validation but also the one most successful at suggesting feasible molecules for synthesis and testing. Figure 3D displays the nine valid molecules proposed by Grok-3 Think. These structures consistently featured a dimethylamine arm, a positively charged substructure also highlighted as important by both traditional ML models and other LLMs. Most of them adhered to cyanine or hemicyanine scaffolds, preserving the core design principles observed in the training data.

Each LLM was also asked to identify structural features that were beneficial or detrimental to antibacterial activity. Table 2 compiles the qualitative feature attributions across models. Each row in the table represents a feature mentioned by at least one model, with each column in that row labeled "good" if the corresponding model associated that feature with high antibacterial activity or "bad" if the model associated that feature with antibacterial inactivity.

Although the different LLMs varied in their predictive accuracy, several consistent SARs emerged among those that successfully completed the classification task (Table 2). All ten successful LLMs emphasized the importance of an extended conjugated polymethine bridge, with modifications to the bridge frequently associated with reduced activity. Seven models linked a high density of localized positive charge, often through terminal amines or dimethylamine arms, with increased antibacterial activity, while eight models noted that negatively charged substituents were detrimental. Dimethylamine groups were cited as beneficial by eight models. Steric bulk was also flagged as unfavorable: eight models identified long or bulky side arms as detrimental, and seven advised against additional substituents on the heterocyclic ring system, particularly halogens, such as chlorine or bromine. These qualitative attributions are remarkably consistent with the features identified by the Gradient Boosted Tree model (Figure 2D), strengthening confidence in the biological relevance of these design rules.

The consistency between feature attributions derived from LLMs and traditional ML models strengthens our understanding of the structural elements that contribute to the antimicrobial activity. Both approaches identified the same key structural determinants: positively charged groups (particularly dimethylamine groups), extended conjugation through the polymethine bridge, and the negative impact of bulky substituents. This agreement across different computational methods increases confidence in these SARs.

This convergence is notable because LLMs and traditional ML models, such as Gradient Boosted Trees, analyze molecular information differently: LLMs process patterns learned from text-based training data, while the Gradient Boosted Tree classifier operates purely on the basis of statistical relationships between molecular features and experimental data. The fact that both approaches reach similar conclusions might suggest that these structural patterns have

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

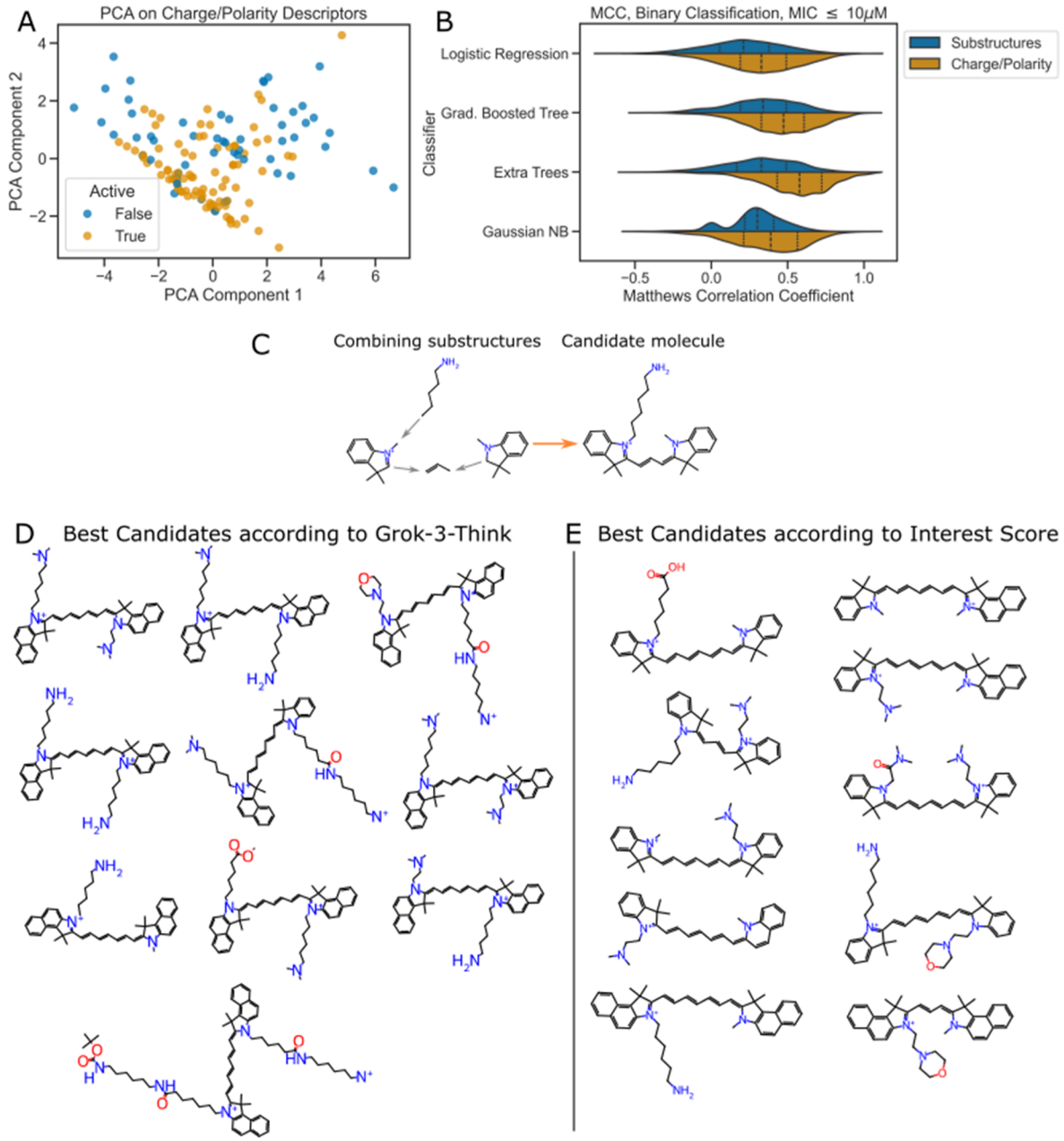


Figure 4. Identification of key properties and cyanine-derived antibiotic candidates using LLM- and ML-guided design. (A) PCA plot of the cyanine structures in the data set according to charge and polarity descriptors (see Figure S11), with the descriptors reduced to two components. (B) Violin plot of MCC for binary classification over 1000 repeated bootstrap samples for four classifiers. The top distribution shows the MCC under the original substructure encoding features, while the bottom distribution shows the MCC under the new LLM-guided charge and polarity descriptors. (C) Schematic illustrating the combinatorial assembly of new candidate molecules from five modular substructure classes: ionic heterocycles, neutral heterocycles, polymethine bridges, and two side arms (Arm A and Arm B). This framework enabled systematic enumeration of >13,000 theoretically possible molecules. (D) Top 10 candidate molecules selected by Grok-3 Think, the best-performing LLM, based on structural evaluation and predicted activity. These structures reflect design principles favored by the model, including extended conjugation and positively charged substituents. (E) Top 10 candidates ranked by the Gradient Boosted Tree classifier using a custom "interest score" that integrates model-predicted activity and structural similarity to the most active compounds in the data set. These Gradient Boosted Tree-prioritized candidates emphasize compact structures with indole cores and minimal steric bulk.

robust biological relevance rather than being artifacts of a particular modeling approach.

The convergence of findings from both traditional ML models and LLMs consistently highlighted cationic character, extended conjugation, and minimal steric hindrance as key predictors of antimicrobial activity. To further interrogate the physicochemical basis of these observations, we performed PCA on molecular descriptors associated with membrane interaction.

For all cyanine derivatives in our data set, we computed two classes of molecular features: (1) lipophilicity and permeability-related descriptors, including molecular weight, octanol−water partition coefficient (log P), total polar surface area (TPSA), total number of rings, and number of rotatable bonds and (2) charge-associated metrics, including number of amide

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

bonds, number of hydrogen donors and acceptors, and number of protonatable nitrogen atoms indicative of positive charge in aqueous environments.[48−50] Among these, TPSA and the number of positively charged sites showed the strongest inverse correlation with the median MIC values (Figure S12), consistent with model-derived feature importances.

Two-component PCA of these physicochemical descriptors (Figure 4A) revealed a notable pattern in which active molecules clustered distinctly from inactive ones, with active compounds forming an approximately linear arrangement in the PCA space. The first PCA component from physicochemical descriptors explains 41% of the variance, and the second explains 21%. This clustering demonstrates that antimicrobial activity is governed by specific combinations of charge and lipophilicity parameters rather than individual properties in isolation, consistent with the complex SAR identified by both the Gradient Boosted Tree model and the LLMs.

Additional evidence for the explanatory power of the physicochemical descriptors is provided in Figure 4B, which shows the distribution of the MCC for binary classification for the same four classifiers in Figure 2 but classifying at a much lower threshold of MIC $\leq$10 μM. A lower MIC threshold allows for screening out of molecules that have some activity but are still of low interest, with the cost of a more severe data set imbalance that can be difficult for a classifier to learn from. Figure 4B shows that the MCC using the LLM-guided descriptors was higher on average for all classifiers than substructure encoding, and for the Extra Trees classifier, nearly 75% of all samples had an MCC above 0.5. This shows that a classifier's performance can be significantly improved with feature selection assisted by LLMs.

This convergence of evidence across multiple approaches, especially feature importance from ML models, attribution analysis from LLMs, and now multivariate physicochemical analysis, provides compelling support for a membrane-targeting mechanism of action. The structural features consistently identified as critical for activity (cationic groups, extended conjugation, and conformational flexibility without bulky substituents) closely align with the known features of membrane-active molecules, including antimicrobial peptides, quaternary ammonium compounds, and other membrane-disruptive agents that bind to and destabilize bacterial phospholipid bilayers.[51−54]

Based on the insights gained from different approaches, we can prioritize the synthesis of derivatives with optimized charge distribution, appropriate hydrophobic/hydrophilic balance, and minimal steric hindrance for enhanced membrane interaction while potentially reducing the resource-intensive cycle of trial-and-error optimization.

Conceptualizing and Predicting New Molecular Candidates. While LLMs were able to independently suggest additional antibacterial molecules based on the cyanine scaffold, not all suggestions were chemically meaningful or synthetically feasible (Figure 3C). To address this limitation and expand the molecular design space more systematically, we developed a combinatorial approach to generate a comprehensive set of candidate molecules. This strategy is compatible with both traditional ML and LLM screening, allowing for a more targeted evaluation.

The substructures used to construct the cyanine-derived molecules, originally shown in Figure S1, were categorized into five functional groups: (1) ionic heterocycles, (2) neutral heterocycles, (3) the polymethine bridge, and two arms (4) Arm A and (5) Arm B, each capable of attaching to a nitrogen center on the heterocycle. A "blank" arm representing a methyl group was also included as an option. New candidate molecules were created by systematically combining one building block from each of the five categories, following the blueprint illustrated in Figure 4C. This process yielded 13,552 possible molecular permutations. The 86 redundant entries, corresponding to molecules already synthesized and tested, were filtered out, resulting in a refined library of 13,466 possible candidate structures.

To prioritize among the candidate molecules, we employed two screening strategies: one using Grok-3 Think (the top-performing LLM), and the other using the Gradient Boosted Tree classifier. For the LLM-based evaluation, we provided Grok-3 Think with a formatted prompt asking it to select the ten most likely active molecules from the candidate library (see Appendix 2). The structures selected by Grok-3 Think (Figure 4D) typically featured positively charged arms and long conjugated polymethine bridges, structural traits previously associated with high activity. However, some candidates also incorporated bulky arms (i.e., arms more than ten atoms in length), which could detract from activity due to steric hindrance.

In parallel, we developed a ML-based scoring function called the "interest score", which integrates both the Gradient Boosted Tree model's predicted probability of activity and molecular similarity to known potent molecules in the data set. This type of scoring to screen molecules is known as data fusion in cheminformatic literature,[55] and the technique can include the combination of structural similarity and ML results as done in this study.[56] The similarity between two molecules is calculated using the RDKit Tanimoto similarity function,[57] which outputs a number between 0 (least similar) and 1 (most similar). By multiplying the Tanimoto similarity scores with the probability of activity according to the classifier, we derived the *interest score* for the structure.

$$S(m_j) = P_{\text{active}}(m_j)\prod_{i=1}^{5} T(m_j, m_i) \tag{5}$$

where S is the interest score, P_{active} is the probability of activity according to the classifier, m_j is the candidate molecule, T is the Tanimoto similarity function, and m_i is one of the five most active molecules from the data set (Figure S13). We chose a multiplicative approach rather than additive or weighted averaging because it enforces that both prediction confidence and structural similarity must be high for a candidate to score well. This conservative approach is particularly valuable given the modest size of our training data set. The top five active molecules were chosen because those were the only five molecules in the data set with the lowest median MIC of 2.5 μM.

Of the entire candidate molecule set, we attempted to choose the 500 most chemically diverse representative candidates using sphere exclusion.[58,59] A minimum cluster centroid distance of 0.32432 yielded 501 representatives, which were then analyzed by the SA Score algorithm[60] to determine the feasibility of synthesis. Figure S14 shows the SA Score vs interest score for the 501 representatives. Ten molecules have a notably high interest score above 0.035, and although they have a relatively low SA Score compared to the majority of representative molecules, their SA Score is still

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

above 3.0 out of 5, meaning that they are not uncharacteristically difficult to synthesize.

The ten molecules with the highest interest scores (Figure 4E) were notably different from those selected by Grok-3 Think. Gradient Boosted Tree-prioritized candidates were generally smaller, more compact, and featured indole rather than benzoindole cores. Most of them contained a single arm, most often a dimethylamine or similar group, capable of bearing a positive charge in aqueous environments. This contrast between the two ranking approaches highlights the different inductive biases of LLMs versus traditional classifiers: while LLMs favored extrapolative design principles (e.g., maximizing positive charge or bridge length), the Gradient Boosted Tree leaned toward conservative optimization grounded in chemical similarity.

Together, the top 20 molecules across both approaches, specifically, derived from Grok-3 Think and the Gradient Boosted Tree model, represent compelling leads for experimental validation. Their structural diversity, overlapping design features, and complementary selection rationales increase the likelihood of identifying active compounds with desirable pharmacological properties.

Interestingly, our antibiotic cyanine derivatives are structurally related to the well-characterized voltage-sensitive carbocyanine dye DiSC3(5) (3,3′-dipropylthiadicarbocyanine iodide), which is widely used as a fluorescent probe for bacterial membrane potential.[61] DiSC3(5) itself is not used as an antimicrobial, but its uptake and fluorescence changes upon depolarization reflect its ability to interact with bacterial membranes and respond to changes in the transmembrane potential. It is notable, however, that DiSC3(5) is not reported as disrupting and destroying the membrane. The most similar molecule in our data set to DiSC3 is BL-545, an aminocyanine with one sulfur in each heterocyclic ring (Figure S15). BL-545 does not exhibit high activity against *S. aureus*, like other cyanines with sulfur in them, possibly due to a negative charge concentration in an aqueous environment that offsets the positive charge mechanism. The contribution of membrane potential changes to antimicrobial activity could be the subject of future work using assays analogous to those developed for DiSC3(5).

LLMs have improved considerably between 2022 and 2025, with new and more sophisticated models being released regularly.[62,63] Some models have also been discontinued, including Grok-2 and Grok-3 as well as ChatGPT 4. Any study of commercial LLMs is therefore time-sensitive, and newer models will likely show improvements in drug discovery tasks. For our classification task, Grok-4 Heavy, the most sophisticated xAI model, achieved a slightly higher MCC than earlier models (see Figure S16), but only because it had one false positive instead of one false negative. The increase in performance was marginal for a classification task at this scale, and therefore, conclusions about improvement in the Grok model are difficult to draw.

CONCLUSIONS

This study demonstrates that both traditional ML and LLMs can accurately predict the antibiotic activity of cyanine-derived molecules, achieving F_1 scores above 0.8 despite a modest data set ($n < 150$). Notably, both modeling approaches converged on similar SAR insights, identifying key features, such as localized positive charge and extended conjugation, as critical for activity, likely due to their role in bacterial membrane disruption. The hypothesis that the cyanines disrupt bacterial membranes due to charge concentration explains trends in the data, but future work could validate this experimentally. The top molecules suggested by the LLMs and Gradient Boosted Tree algorithm could be synthesized and tested, and the ones with higher charge concentration should have a lower MIC than the others.

The alignment between LLM-derived and ML-derived attributions strengthens the confidence in the underlying design rules and offers mechanistic interpretability. These features are consistent with a membrane-targeting mode of action, suggesting that highly charged, yet conformationally flexible, molecules could interact favorably with negatively charged phospholipid bilayers. Importantly, bulky or sterically hindered substituents appear to reduce activity, likely by impeding membrane association.[51−54]

To address concerns around reproducibility in LLM-based research, future work should explore the integration of open-source models such as LLaMA, Mistral, and BLOOM, which offer transparent architectures and version-controlled checkpoints.[64,65] Unlike proprietary systems that evolve unpredictably, these models can be frozen and self-hosted, enabling consistent replication of results across time and institutions.[64,66] Recent studies have demonstrated that open-source LLMs can match or exceed proprietary performance in scientific tasks while supporting ethical and reproducible workflows.[67] By leveraging these models, it is possible to build stable, auditable pipelines tailored to domain-specific applications, setting a foundation for long-term scientific validity and collaborative benchmarking.

Beyond prediction, ML and LLM querying were also leveraged to suggest new candidate molecules. LLMs such as Grok-3 Think provided creative, interpretable, and synthetically feasible designs, while the Gradient Boosted Tree model enabled a systematic, similarity-weighted ranking of likely active candidates. These complementary outputs produced a shortlist of testable molecules that can now be synthesized and validated, completing the cycle of design, evaluation, and iteration.

The success of LLMs in this context highlights their emerging role as accessible, low-barrier tools for medicinal chemistry. With minimal prompting and no fine-tuning, these models offered both an accurate classification and chemically plausible design suggestions. Their ability to reason over molecular structure using natural language interfaces facilitates broader use by nonspecialists, potentially democratizing early-stage drug discovery. Of special note is the large reduction in labor and time required when conducting analysis with LLMs instead of traditional ML programs, which, like other academic projects, can take months to complete.[68,69] The use of LLMs allowed for rapid analysis of the same SAR, with the only preparation beforehand being the choice of molecular structure formatting and careful prompt writing. The LLMs only take 1−3 min to finish their computations, even on a cell phone or cheap laptop, since the services are web-based and all response generation is physically conducted on remote servers. Thus, parallel studies could be conducted with the ten models from Table 1 within a two-week period.

Given the urgency of the AMR crisis, along with the high cost and failure rate of traditional antibiotic research and development, integrating LLMs into the molecular design pipeline offers a path to accelerate not only the discovery of new antibiotics but also the refinement of existing candidates.

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Beyond predicting the antibacterial activity, LLMs could be leveraged to suggest modifications that improve pharmacokinetic properties, reduce toxicity, and minimize the likelihood of resistance development. Future work should explore their application across larger data sets, more diverse chemical scaffolds, and experimental validation workflows, extending their utility across both antibiotic optimization and broader structure-based drug design efforts.

METHODS

Broth Microdilution Assay for Determining the Minimum Inhibitory Concentration. The antimicrobial activity of cyanine compounds was evaluated using a broth microdilution assay adapted from Clinical & Laboratory Standards Institute guidelines.[70] On day 1, *S. aureus* was streaked onto Luria−Bertani broth agar plates and incubated overnight at 37 °C for 16−20 h to obtain well-isolated colonies. On day 2, 3−5 colonies were picked using a sterile loop and suspended in sterile phosphate-buffered saline. This suspension was then diluted in fresh cation-adjusted Mueller−Hinton broth (CAMHB) to yield a final inoculum of approximately ∼1 × 10^5 CFU/mL. Sterile microtiter plates were prepared by dispensing 200 μL of the inoculated CAMHB into column 1 and 100 μL into columns 2 through 12.

Cyanine compounds were prepared as 16 mM stock solutions in 100% DMSO and stored at −20 °C. For MIC testing, 1 μL of the compound stock was added directly to column 1 of the inoculated plate to achieve a final test concentration of 80 μM. A 2-fold serial dilution was performed across columns 1−10 by sequentially transferring 100 μL and mixing thoroughly. Column 11 served as a growth control (inoculated CAMHB without a compound), and column 12 served as a sterility control (uninoculated CAMHB). Plates were covered and incubated statically at 37 °C for 16−20 h. MIC values were determined spectrophotometrically by measuring OD_{600}. The MIC was defined as the lowest compound concentration resulting in ≥90% reduction in absorbance relative to the growth control.

Synthesis of Cyanine Molecules. All molecules in this work were synthesized by B. Li and T. Si except for Cy7.5-amine, Cy7-amine, Cy5.5-amine, and Cy5-amine that were purchased from Lumiprobe Corp. (Maryland, USA). The synthesis method for aminocyanines is described in detail by Ayala-Orozco et al., Supporting Information, pages S60−S91.[71] The synthesis method for hemicyanines is described in detail in the Supporting Information.

Software Packages. All chemical structures analyzed in this work were represented in the SMILES format exported from ChemDraw representations. Data analysis and ML were conducted using the Python programming language[72] and the Jupyter interactive Python notebook[73] functionality provided by Visual Studio Code.

Several Python libraries were used for this study. The data were imported and interpreted using NumPy[74] and pandas,[75,76] and all plots were generated using Seaborn and Matplotlib. RDKit[77] was used to identify chemical features and substructures as well as to draw all chemical structures and substructures. The Mordred[78] feature calculator was used to determine Mordred descriptors. We used scikit-learn for all ML model training and cross-validation as well as for LLM model evaluation.[21]

ML Classifiers. The Logistic Regression model[79] is a linear classifier which predicts the probability of the positive class (active) according to the logistic equation

$$p(X_i) = \frac{1}{1 + \exp(-X_i w + w_0)} \tag{6}$$

where p is the probability of the molecule being active, w is the vector of weights for each feature, and w_0 is the vector of intercepts. The logistic regression model minimizes the cost function

$$\min_w \frac{1}{S}\sum_{i=1}^{n} s_i(-y_i \log(p(X_i)) - (1 - y_i)\log(1 - p(X_i))) \tag{7}$$

where s_i is the vector of user-defined weights (the default of 1 for each was used in this study), and S is the sum of all sample weights (equal to the number of features in this study).

The Gradient Boosted Tree classifier is based on decision tree classification. A decision tree[31] generates a tree of sequential feature queries to determine the right label for classification, similar to a flowchart. Each "decision" node in the tree is optimized to split the data set into classes with minimum Gini impurity,[31] and trees can vary in depth complexity. As an ensemble classifier, the Gradient Boosted Tree model generates multiple decision trees sequentially. Later trees use the error from earlier trees to adjust their decision structure according to the steepest gradient ascent algorithm.[80] The final classification is made by weighted consensus of the ensemble of trees, with later trees being weighed less than earlier trees.

The "extremely randomized trees" classifier, abbreviated as the extra trees classifier, is another model that classifies according to the vote of an ensemble of decision trees, but instead of sequentially generating trees based on the error of previous ones, the trees are randomized independently. The "extremely randomized" nature of the algorithm means that when each decision node of a tree is being constructed, the best threshold is chosen from a set of randomly selected thresholds (rather than an exhaustive search of all possible thresholds). The independently formed trees tend to be individually biased, and the classifier relies on the biases of many trees canceling out.

The Gaussian Naive Bayes classifier is a model that operates according to Bayes' theorem, which states that the probability of a molecule being of class y given features x_i is related to the converse conditional probability and independent probabilities of the class and features as follows:

$$P(y|x_1, ..., x_n) = \frac{P(x_1, ..., x_n|y)P(y)}{P(x_1, ..., x_n)} \tag{8}$$

All Naive Bayes classifiers make the "naive" assumption that the conditional probability of each feature given the class is independent,[81] and therefore that $P(x_1, ..., x_n|y) = \prod_{i=1}^{n} P(x_i|y)$. This key assumption simplifies Bayes' theorem to

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)} \tag{9}$$

Because the conditional probability distribution of the individual features is not known, each Naive Bayes classifier

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

assumes a different distribution. The Gaussian Naive Bayes classifier[82] assumes this distribution to be Gaussian:

$$P(y|x_i) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (10)$$

In the scikit-learn library, the parameters σ and μ are estimated using the maximum likelihood. Because $P(y)$ and $P(x_1, \cdots, x_n)$ can be directly observed by the classifier for a given data set, the classifier finally calculates the left-hand side of Bayes' theorem directly.

■ ASSOCIATED CONTENT

Data Availability Statement

Figure 1A and the graphic table of contents were created in BioRender. Lathem, A. (2025) https://BioRender.com/rodqtzo. The files have been deposited on Figshare using the following public link: https://figshare.com/projects/Identifying_Structure-Activity_Relationships_for_Cyanine-Derived_Antibiotics_Using_Machine_Learning_and_Commercial_Large_Language_Models/252515.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c01321.

SMILES strings for the studied compounds (XLSX)

Chemical structures, additional graphs, LLM prompts, scatter plots, bar graphs, synthesis methods, NMR characterization for hemicyanines, and other data (PDF)

■ AUTHOR INFORMATION

Corresponding Author

James M. Tour − *Smalley-Curl Institute, Department of Chemistry, Department of Computer Science, and Department of Materials Science and NanoEngineering, NanoCarbon Center, Rice Advanced Materials Institute, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0002-8479-9328; Email: tour@rice.edu

Authors

Alexander Lathem − *Smalley-Curl Institute and Department of Chemistry, Rice University, Houston, Texas 77005, United States*

Angela Medvedeva − *Department of Chemistry, Rice University, Houston, Texas 77005, United States*

Ana Luisa L. Mendes dos Santos − *Department of Chemistry, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0002-5450-9414

Bowen Li − *Department of Chemistry, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0003-4359-1712

Tengda Si − *Department of Chemistry, Rice University, Houston, Texas 77005, United States*

Anatoly B. Kolomeisky − *Department of Chemistry, Rice University, Houston, Texas 77005, United States;* orcid.org/0000-0001-5677-6690

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c01321

Author Contributions

Lathem: Conceptualization, programming, LLM interaction, figure design, writing. Medvedeva: Machine learning model architecture, statistical analysis, editing. Santos: Biological experiments, editing. Li: Molecule synthesis and characterization. Si: Molecule synthesis and characterization. Kolomeisky: Machine learning model architecture, editing. Tour: Conceptualization, writing, editing. All authors contributed to final edits and revisions.

Notes

The authors declare the following competing financial interest(s): Rice University owns intellectual property on the use of organic molecules for antibiotics. This intellectual property is currently unlicensed.

■ ACKNOWLEDGMENTS

This research was funded by the Welch Foundation (grants C-2017-20220331 and C-1559) and the Discovery Institute.

■ REFERENCES

(1) Murray, C. J. L.; Ikuta, K. S.; Sharara, F.; Swetschinski, L.; Robles Aguilar, G.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; Johnson, S. C.; Browne, A. J.; Chipeta, M. G.; Fell, F.; Hackett, S.; Haines-Woodhouse, G.; Kashef Hamadani, B. H.; Kumaran, E. A. P.; McManigal, B.; Achalapong, S.; Agarwal, R.; Akech, S.; Albertson, S.; Amuasi, J.; Andrews, J.; Aravkin, A.; Ashley, E.; Babin, F.-X.; Bailey, F.; Baker, S.; Basnyat, B.; Bekker, A.; Bender, R.; Berkley, J. A.; Bethou, A.; Bielicki, J.; Boonkasidecha, S.; Bukosia, J.; Carvalheiro, C.; Castañeda-Orjuela, C.; Chansamouth, V.; Chaurasia, S.; Chiurchiù, S.; Chowdhury, F.; Clotaire Donatien, R.; Cook, A. J.; Cooper, B.; Cressey, T. R.; Criollo-Mora, E.; Cunningham, M.; Darboe, S.; Day, N. P. J.; De Luca, M.; Dokova, K.; Dramowski, A.; Dunachie, S. J.; Duong Bich, T.; Eckmanns, T.; Eibach, D.; Emami, A.; Feasey, N.; Fisher-Pearson, N.; Forrest, K.; Garcia, C.; Garrett, D.; Gastmeier, P.; Giref, A. Z.; Greer, R. C.; Gupta, V.; Haller, S.; Haselbeck, A.; Hay, S. I.; Holm, M.; Hopkins, S.; Hsia, Y.; Iregbu, K. C.; Jacobs, J.; Jarovsky, D.; Javanmardi, F.; Jenney, A. W. J.; Khorana, M.; Khusuwan, S.; Kissoon, N.; Kobeissi, E.; Kostyanev, T.; Krapp, F.; Krumkamp, R.; Kumar, A.; Kyu, H. H.; Lim, C.; Lim, K.; Limmathurotsakul, D.; Loftus, M. J.; Lunn, M.; Ma, J.; Manoharan, A.; Marks, F.; May, J.; Mayxay, M.; Mturi, N.; Munera-Huertas, T.; Musicha, P.; Musila, L. A.; Mussi-Pinhata, M. M.; Naidu, R. N.; Nakamura, T.; Nanavati, R.; Nangia, S.; Newton, P.; Ngoun, C.; Novotney, A.; Nwakanma, D.; Obiero, C. W.; Ochoa, T. J.; Olivas-Martinez, A.; Olliaro, P.; Ooko, E.; Ortiz-Brizuela, E.; Ounchanum, P.; Pak, G. D.; Paredes, J. L.; Peleg, A. Y.; Perrone, C.; Phe, T.; Phommasone, K.; Plakkal, N.; Ponce-de-Leon, A.; Raad, M.; Ramdin, T.; Rattanavong, S.; Riddell, A.; Roberts, T.; Robotham, J. V.; Roca, A.; Rosenthal, V. D.; Rudd, K. E.; Russell, N.; Sader, H. S.; Saengchan, W.; Schnall, J.; Scott, J. A. G.; Seekaew, S.; Sharland, M.; Shivamallappa, M.; Sifuentes-Osornio, J.; Simpson, A. J.; Steenkeste, N.; Stewardson, A. J.; Stoeva, T.; Tasak, N.; Thaiprakong, A.; Thwaites, G.; Tigoi, C.; Turner, C.; Turner, P.; van DoornVelaphi, H. R. S.; Vongpradith, A.; Vongsouvath, M.; Vu, H.; Walsh, T.; Walson, J. L.; Waner, S.; Wangrangsimakul, T.; Wannapinij, P.; Wozniak, T.; Young Sharma, T. E. M. W.; Yu, K. C.; Zheng, P.; Sartorius, B.; Lopez, A. D.; Stergachis, A.; Moore, C.; Dolecek, C.; Naghavi, M. Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *Lancet* **2022**, *399* (10325), 629−655.

(2) O'Neill, J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*, 2016; p 84.

(3) Dadgostar, P. Antimicrobial Resistance: Implications and Costs. *Infect. Drug Resist.* **2019**, *12*, 3903−3910.

(4) Conly, J.; Johnston, B. Where Are All the New Antibiotics? The New Antibiotic Paradox. *Can. J. Infect. Dis. Med. Microbiol.* **2005**, *16* (3), 159−160.

(5) Alander, J. T.; Kaartinen, I.; Laakso, A.; Pätilä, T.; Spillmann, T.; Tuchin, V. V.; Venermo, M.; Välisuo, P. A Review of Indocyanine Green Fluorescent Imaging in Surgery. *Int. J. Biomed. Imaging* **2012**, *2012*, 940585.

(6) Browning, C. H.; Cohen, J. B.; Gulbransen, R. The Antiseptic Properties of Cyanine Dyes. *Br. Med. J.* **1922**, *1* (3196), 514−515.
(7) Wainwright, M.; Kristiansen, J. E. Quinoline and Cyanine Dyes–Putative Anti-MRSA Drugs. *Int. J. Antimicrob. Agents* **2003**, *22* (5), 479−486.
(8) Wainwright, M. Acridine—a Neglected Antibacterial Chromophore. *J. Antimicrob. Chemother.* **2001**, *47* (1), 1−13.
(9) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2019**, *20* (2), 273−286.
(10) Kp Jayatunga, M.; Ayers, M.; Bruens, L.; Jayanth, D.; Meier, C. How Successful Are AI-Discovered Drugs in Clinical Trials? A First Analysis and Emerging Lessons. *Drug Discovery Today* **2024**, *29* (6), 104009.
(11) Singha, M.; Pu, L.; Stanfield, B. A.; Uche, I. K.; Rider, P. J. F.; Kousoulas, K. G.; Ramanujam, J.; Brylinski, M. Artificial Intelligence to Guide Precision Anticancer Therapy with Multitargeted Kinase Inhibitors. *BMC Cancer* **2022**, *22* (1), 1211.
(12) Yu, T.-H.; Su, B.-H.; Battalora, L. C.; Liu, S.; Tseng, Y. J. Ensemble Modeling with Machine Learning and Deep Learning to Provide Interpretable Generalized Rules for Classifying CNS Drugs with High Prediction Power. *Briefings Bioinf.* **2022**, *23* (1), bbab377.
(13) Izmailyan, R.; Matevosyan, M.; Khachatryan, H.; Shavina, A.; Gevorgyan, S.; Ghazaryan, A.; Tirosyan, I.; Gabrielyan, Y.; Ayvazyan, M.; Martirosyan, B.; Harutyunyan, V.; Zakaryan, H. Discovery of New Antiviral Agents through Artificial Intelligence: *In Vitro* and *in Vivo* Results. *Antiviral Res.* **2024**, *222*, 105818.
(14) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688−702e13.
(15) Kourtis, A. P.; Hatfield, K.; Baggs, J.; Mu, Y.; See, I.; Epson, E.; Nadle, J.; Kainer, M. A.; Dumyati, G.; Petit, S.; et al. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus Aureus Bloodstream Infections — United States. *MMWR Morb. Mortal. Wkly. Rep.* **2019**, *68*, 214−219.
(16) Sá-Leão, R.; Santos Sanches, I.; Couto, I.; Alves, C. R.; de Lencastre, H. Low Prevalence of Methicillin-Resistant Strains among Staphylococcus Aureus Colonizing Young and Healthy Members of the Community in Portugal. *Microb. Drug Resist.* **2001**, *7* (3), 237−245.
(17) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(18) Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bull.* **1945**, *1* (6), 80−83.
(19) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; sriniker; Gedeck, P.; Jones, G.; NadineSchneider; Kawashima, E.; Nealschneider, D.; Dalke, A.; Swain, M.; Cole, B.; Turk, S.; Savelev, A.; tadhurst-cdd; Vaucher, A.; Wójcikowski, M.; Take, I.; Walker, R.; Scalfani, V. F.; Faara, H.; Ujihara, K.; Probst, D.; Lehtivarjo, J.; godin, g.; Pahl, A.; Monat, J. *Rdkit/Rdkit: 2025_03_1 (Q1 2025) Release*; Zenodo, 2025.
(20) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11* (23), 1046−1053.
(21) Rainio, O.; Teuho, J.; Klén, R. Evaluation Metrics and Statistical Tests for Machine Learning. *Sci. Rep.* **2024**, *14* (1), 6086.
(22) Hicks, S. A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M. A.; Halvorsen, P.; Parasa, S. On Evaluation Metrics for Medical Applications of Artificial Intelligence. *Sci. Rep.* **2022**, *12*, 5979.
(23) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21* (1), 6.
(24) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One* **2017**, *12* (6), No. e0177678.
(25) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2012; Vol. *25*.
(26) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104* (1), 148−175.
(27) Frazier, P. I. A Tutorial on Bayesian Optimization. *arXiv* **2018**, arXiv:1807.02811.
(28) Letham, B.; Karrer, B.; Ottoni, G.; Bakshy, E. Constrained Bayesian Optimization with Noisy Experiments. *Bayesian Anal.* **2019**, *14* (2), 495−519.
(29) Bengio, Y.; Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* **2004**, *5* (Sep), 1089−1105.
(30) Krstajic, D.; Buturovic, L. J.; Leahy, D. E.; Thomas, S. Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. *J. Cheminf.* **2014**, *6* (1), 10.
(31) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman and Hall/CRC: New York, 2017.
(32) Kejík, Z.; Hajduch, J.; Abramenko, N.; Vellieux, F.; Veselá, K.; Fialová, J. L.; Petrláková, K.; Kučnirová, K.; Kaplánek, R.; Tatar, A.; Skaličková, M.; Masařík, M.; Babula, P.; Dytrych, P.; Hoskovec, D.; Martásek, P.; Jakubek, M. Cyanine Dyes in the Mitochondria-Targeting Photodynamic and Photothermal Therapy. *Commun. Chem.* **2024**, *7* (1), 1−39.
(33) Saha, P. C.; Chatterjee, T.; Pattanayak, R.; Das, R. S.; Mukherjee, A.; Bhattacharyya, M.; Guha, S. Targeting and Imaging of Mitochondria Using Near-Infrared Cyanine Dye and Its Application to Multicolor Imaging. *ACS Omega* **2019**, *4* (11), 14579−14588.
(34) Nödling, A. R.; Mills, E. M.; Li, X.; Cardella, D.; Sayers, J.; Wu, S. H.; Jones, A. T.; Luk, L. Y. P.; Tsai, Y. H.; L, Y.; Tsai, Y.-H. Cyanine dye mediated mitochondrial targeting enhances the anti-cancer activity of small-molecule cargoes. *Chem. Commun.* **2020**, *56* (34), 4672−4675.
(35) Kurutos, A.; Orehovec, I.; Saftić, D.; Horvat, L.; Crnolatac, I.; Piantanida, I.; Deligeorgiev, T. Cell Penetrating, Mitochondria Targeting Multiply Charged DABCO-Cyanine Dyes. *Dyes Pigm.* **2018**, *158*, 517−525.
(36) Schwechheimer, C.; Rönicke, F.; Schepers, U.; Wagenknecht, H.-A. A New Structure−Activity Relationship for Cyanine Dyes to Improve Photostability and Fluorescence Properties for Live Cell Imaging. *Chem. Sci.* **2018**, *9* (31), 6557−6563.
(37) Okoh, O. A.; Lawrence, C. L.; Bisby, R. H.; Brennan, S. L.; Smith, R. B. Towards Structurally New Cyanine Dyes—Investigating the Photophysical and Potential Antifungal Properties of Linear Substituted Heptamethine Dyes. *Color. Technol.* **2025**, *141* (1), 20−25.
(38) Omardien, S.; Brul, S.; Zaat, S. A. J. Antimicrobial Activity of Cationic Antimicrobial Peptides against Gram-Positives: Current Progress Made in Understanding the Mode of Action and the Response of Bacteria. *Front. Cell Dev. Biol.* **2016**, *4*, 111.
(39) Alkhalifa, S.; Jennings, M. C.; Granata, D.; Klein, M.; Wuest, W. M.; Minbiole, K. P. C.; Carnevale, V. Analysis of the Destabilization of Bacterial Membranes by Quaternary Ammonium Compounds: A Combined Experimental and Computational Study. *ChemBioChem* **2020**, *21* (10), 1510−1516.
(40) Manzo, G.; Hind, C. K.; Ferguson, P. M.; Amison, R. T.; Hodgson-Casson, A. C.; Ciazynska, K. A.; Weller, B. J.; Clarke, M.; Lam, C.; Man, R. C. H.; Shaughnessy, B. G. O.; Clifford, M.; Bui, T. T.; Drake, A. F.; Atkinson, R. A.; Lam, J. K. W.; Pitchford, S. C.; Page, C. P.; Phoenix, D. A.; Lorenz, C. D.; Sutton, J. M.; Mason, A. J. A Pleurocidin Analogue with Greater Conformational Flexibility, Enhanced Antimicrobial Potency and in Vivo Therapeutic Efficacy. *Commun. Biol.* **2020**, *3* (1), 1−16.

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(41) Mohamed, N. S. E.-D.; AbuEl-Hamd, R. M. Synthesis, Spectroscopic and Antimicrobial Studies of Some Novel Cyanine Dyes Based on Bis-Coumarin Heterocycles Derivatives. *Eur. J. Chem.* **2016**, *7* (1), 66−72.
(42) Prakash, A. V.; Yazabak, F.; Hovor, I.; Nakonechny, F.; Kulyk, O.; Semenova, O.; Bazylevich, A.; Gellerman, G.; Patsenker, L. Highly Efficient Near-IR Cyclohexene Cyanine Photosensitizers for Antibacterial Photodynamic Therapy. *Dyes Pigm.* **2023**, *211*, 111053.
(43) Li, G.; Wang, P.; Liu, J.; Guo, Y.; Ji, K.; Shang, Z.; Xu, Z. Meta In-Context Learning Makes Large Language Models Better Zero and Few-Shot Relation Extractors. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024; Vol. 7, pp 6350−6358.
(44) Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; Sun, X.; Li, L.; Sui, Z. A Survey on In-Context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Miami, FL, USA, 2024; pp 1107−1128.
(45) Mohamed, A.; Rashid, M. E.; Shaalan, K. In-Context Learning in Large Language Models (LLMs): Mechanisms, Capabilities, and Implications for Advanced Knowledge Representation and Reasoning. *IEEE Access* **2025**, *13*, 95574−95593.
(46) Chen, Y.-Q.; Yu, T.; Song, Z.-Q.; Wang, C.-Y.; Luo, J.-T.; Xiao, Y.; Qiu, H.; Wang, Q.-Q.; Jin, H.-M. Application of Large Language Models in Drug-Induced Osteotoxicity Prediction. *J. Chem. Inf. Model.* **2025**, *65* (7), 3370−3379.
(47) Ferber, D.; Wölflein, G.; Wiest, I. C.; Ligero, M.; Sainath, S.; Ghaffari Laleh, N.; El Nahhas, O. S. M.; Müller-Franzes, G.; Jäger, D.; Truhn, D.; Kather, J. N. In-Context Learning Enables Multimodal Large Language Models to Classify Cancer Pathology Images. *Nat. Commun.* **2024**, *15*, 10104.
(48) Pires, C. L.; Moreno, M. J. Improving the Accuracy of Permeability Data to Gain Predictive Power: Assessing Sources of Variability in Assays Using Cell Monolayers. *Membranes* **2024**, *14* (7), 157.
(49) Venable, R. M.; Krämer, A.; Pastor, R. W. Molecular Dynamics Simulations of Membrane Permeability. *Chem. Rev.* **2019**, *119* (9), 5954−5997.
(50) Leung, S. S. F.; Mijalkovic, J.; Borrelli, K.; Jacobson, M. P. Testing Physical Models of Passive Membrane Permeation. *J. Chem. Inf. Model.* **2012**, *52* (6), 1621−1636.
(51) Lei, J.; Sun, L.; Huang, S.; Zhu, C.; Li, P.; He, J.; Mackey, V.; Coy, D. H.; He, Q. The Antimicrobial Peptides and Their Potential Clinical Applications. *Am. J. Transl. Res.* **2019**, *11* (7), 3919−3931.
(52) Arnold, W. A.; Blum, A.; Branyan, J.; Bruton, T. A.; Carignan, C. C.; Cortopassi, G.; Datta, S.; DeWitt, J.; Doherty, A.-C.; Halden, R. U.; Harari, H.; Hartmann, E. M.; Hrubec, T. C.; Iyer, S.; Kwiatkowski, C. F.; LaPier, J.; Li, D.; Li, L.; Muñiz Ortiz, J. G.; Salamova, A.; Schettler, T.; Seguin, R. P.; Soehl, A.; Sutton, R.; Xu, L.; Zheng, G. Quaternary Ammonium Compounds: A Chemical Class of Emerging Concern. *Environ. Sci. Technol.* **2023**, *57* (20), 7645−7665.
(53) Adams, E. The Antibacterial Action of Crystal Violet. *J. Pharm. Pharmacol.* **1967**, *19* (12), 821−826.
(54) Zhang, N.; Ma, S. Recent Development of Membrane-Active Molecules as Antibacterial Agents. *Eur. J. Med. Chem.* **2019**, *184*, 111743.
(55) Arif, S. M.; Holliday, J. D.; Willett, P. Chapter 5 - The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening. In *Advances in Mathematical Chemistry and Applications*; Basak, S. C., Restrepo, G., Villaveces, J. L., Eds.; Bentham Science Publishers, 2015; pp 92−112.
(56) Jeon, W.; Kim, D. FP2VEC: A New Molecular Featurizer for Learning Molecular Properties. *Bioinformatics* **2019**, *35* (23), 4979−4985.
(57) Muegge, I.; Mukherjee, P. An Overview of Molecular Fingerprint Similarity Search in Virtual Screening. *Expert Opin. Drug Discov.* **2016**, *11* (2), 137−148.
(58) Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J.; Osman, J. Parameter Based Methods for Compound Selection from Chemical Databases. *Quant. Struct.-Act. Relat.* **1996**, *15* (4), 285−289.
(59) Landrum, G. *Sphere Exclusion Clustering with the RDKit*; RDKit Blog. https://greglandrum.github.io/rdkit-blog/posts/2020-11-18-sphere-exclusion-clustering.html (accessed Sept 10, 2025).
(60) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1* (1), 8.
(61) te Winkel, J. D.; Gray, D. A.; Seistrup, K. H.; Hamoen, L. W.; Strahl, H. Analysis of Antimicrobial-Triggered Membrane Depolarization Using Voltage Sensitive Dyes. *Front. Cell Dev. Biol.* **2016**, *4*, 29.
(62) Maslej, N. *Artificial Intelligence Index Report 2025*; Stanford Institute for Human-Centered AI, 2025.
(63) Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey. *arXiv* **2025**, arXiv:2402.06196.
(64) Spirling, A. Why Open-Source Generative AI Models Are an Ethical Way Forward for Science. *Nature* **2023**, *616* (7957), 413.
(65) Manchanda, J.; Boettcher, L.; Westphalen, M.; Jasser, J. The Open Source Advantage in Large Language Models (LLMs). *arXiv* **2025**, arXiv:2412.12004.
(66) Schur, A.; Groenjes, S. Comparative Analysis for Open-Source Large Language Models. In *HCI International 2023—Late Breaking Posters*; Stephanidis, C., Antona, M., Ntoa, S., Salvendy, G., Eds.; Springer Nature Switzerland: Cham, 2024; pp 48−54.
(67) Staudinger, M.; Kern, B. M. J.; Miksa, T.; Arnhold, L.; Knees, P.; Rauber, A.; Hanbury, A. Mission Reproducibility: An Investigation on Reproducibility Issues in Machine Learning and Information Retrieval Research. In *2024 IEEE 20th International Conference on e-Science (e-Science)*, 2024; pp 1−9.
(68) Hecht, L. E. *Add it Up: How Long does a Machine Learning Deployment Take?*; The New Stack. https://thenewstack.io/add-it-up-how-long-does-a-machine-learning-deployment-take/ (accessed May 06, 2025).
(69) Tumin, D.; Brewer, K. L.; Cummings, D. M.; Keene, K. L.; Campbell, K. M. Estimating Clinical Research Project Duration from Idea to Publication. *J. Invest. Med.* **2022**, *70* (1), 108−109.
(70) CLSI. M07|Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria That Grow Aerobically. https://clsi.org/shop/standards/m07/ (accessed May 05, 2025).
(71) Ayala-Orozco, C.; Li, G.; Li, B.; Vardanyan, V.; Kolomeisky, A. B.; Tour, J. M. How to Build Plasmon-Driven Molecular Jackhammers That Disassemble Cell Membranes and Cytoskeletons in Cancer. *Adv. Mater.* **2024**, *36* (14), 2309910.
(72) Van Rossum, G.; Drake, F. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, 2009.
(73) Kluyver, T.; Ragan-Kelley, B.; Pé Rez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.; Willing, C.; Team, J. D. Jupyter Notebooks − a Publishing Format for Reproducible Computational Workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*; IOS Press, 2016; pp 87−90.
(74) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357−362.
(75) McKinney, W. *Data Structures for Statistical Computing in Python*, Austin, TX, 2010; pp 56−61.
(76) The Pandas Development Team. *Pandas-Dev/Pandas: Pandas*; Zenodo, 2024.
(77) Landrum, G. *RDKit: Open-Source Cheminformatics. 2006*. https://cir.nii.ac.jp/crid/1370004237630036224 (accessed May 01, 2025).
(78) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), 4.

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(79) McCullagh, P. *Generalized Linear Models*, 2nd ed.; Routledge: New York, 2019.
(80) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189−1232.
(81) Zhang, H. The Optimality of Naive Bayes. In *Proceedings of the Florida Artificial Intelligence Research Symposium*, 2004; Vol. *17*.
(82) 1.9. Naive Bayes. scikit-learn. https://scikit-learn.org/stable/modules/naive_bayes.html (accessed April 30, 2025).

https://doi.org/10.1021/acs.jcim.5c01321
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX