JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

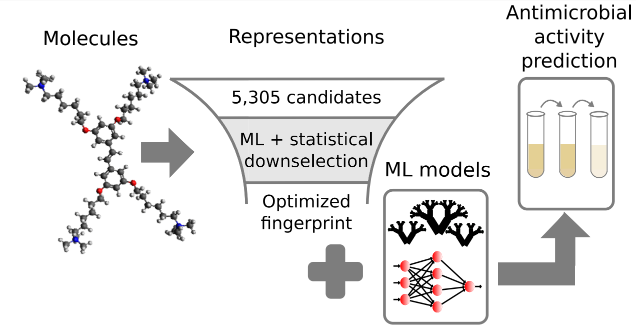
pubs.acs.org/JACS Article

Predicting Antimicrobial Activity of Conjugated Oligoelectrolyte Molecules via Machine Learning

Armi Tiihonen,* Sarah J. Cox-Vazquez,* Qiaohao Liang, Mohamed Ragab, Zekun Ren, Noor Titan Putri Hartono, Zhe Liu, Shijing Sun, Cheng Zhou, Nathan C. Incandela, Jakkarin Limwongyut, Alex S. Moreland, Senthilnath Jayavelu, Guillermo C. Bazan,* and Tonio Buonassisi*

Cite This: *J. Am. Chem. Soc.* 2021, 143, 18917−18931 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: New antibiotics are needed to battle growing antibiotic resistance, but the development process from hit, to lead, and ultimately to a useful drug takes decades. Although progress in molecular property prediction using machine-learning methods has opened up new pathways for aiding the antibiotics development process, many existing solutions rely on large data sets and finding structural similarities to existing antibiotics. Challenges remain in modeling unconventional antibiotic classes that are drawing increasing research attention. In response, we developed an antimicrobial activity prediction model for conjugated oligoelectrolyte molecules, a new class of antibiotics that lacks extensive prior structure−activity relationship studies. Our approach enables us to predict the minimum inhibitory concentration for *E. coli* K12, with 21 molecular descriptors selected by recursive elimination from a set of 5305 descriptors. This predictive model achieves an R^2 of 0.65 with no prior knowledge of the underlying mechanism. We find the molecular representation optimum for the domain is the key to good predictions of antimicrobial activity. In the case of conjugated oligoelectrolytes, a representation reflecting the three-dimensional shape of the molecules is most critical. Although it is demonstrated with a specific example of conjugated oligoelectrolytes, our proposed approach for creating the predictive model can be readily adapted to other novel antibiotic candidate domains.

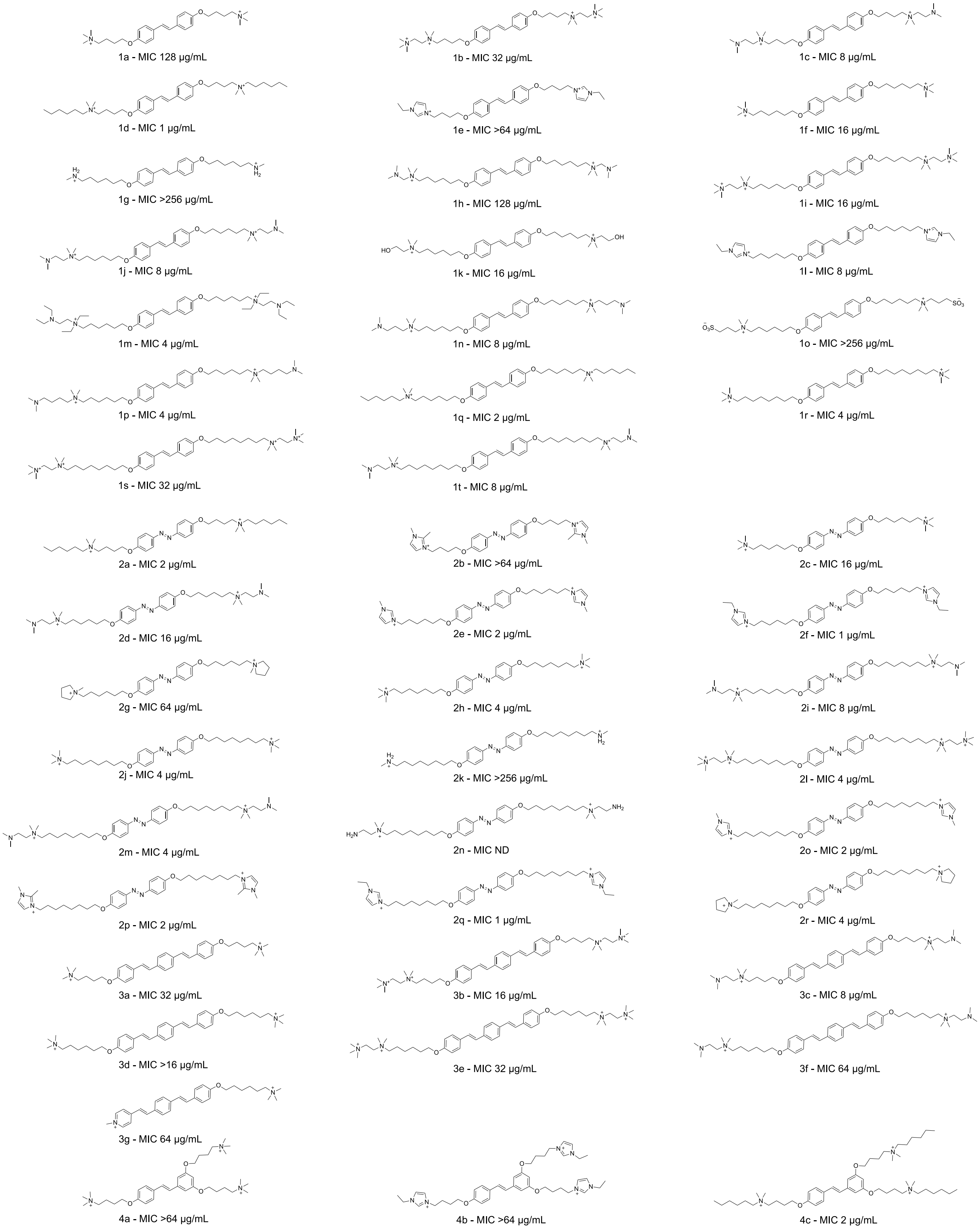
1. INTRODUCTION

Antibiotic resistance is on the rise globally and confronts us with the potential of up to 10 million deaths per year by 2050 if no actions are taken.[1] Even though new antibiotic candidates are in preclinical development pipelines,[2,3] the development cycle remains slow, typically taking 10−15 years. Moreover, microbes inevitably build resistance to new chemical structures, thereby requiring multiple new classes of antibiotics to be continuously discovered. The successful development of truly new antibiotics will likely be challenged by minimal mechanistic insight on the candidates and few structural analogues available. One may invest in the detailed biocentric traditional approach that identifies the point of action. Alternatively, a series of structural variations in the molecules are generated with the anticipation of drawing an empirical structure−activity relationship. Within the context of the latter, chemical intuition and inference are brought together to develop a more successful structure. Such is the situation with conjugated oligoelectrolyte (COE) molecules, a class of novel antibiotics candidates. In the case of COEs, the challenge is further exacerbated by their complex structure, absence of analogues, and Gram-type specificity, to name a few. Machine learning (ML) may offer an alternative approach to streamline development, whereby the essential elements in the molecular structure are identified that most strongly correlate with antibiotic activity. However, thousands of descriptors may be generated for each COE molecule, thereby necessitating the development of a principled downselection process in order to develop an operational model. We solve this problem and propose an ML model for predicting antimicrobial activity of COEs.

COEs are a distinctive class of molecules unified by their hydrophobic π-electron conjugated core and pendant groups bearing ionic functionalities;[4,5] see Figure 1 for the examples used in this work. The tunable and molecular COE framework offers a practical advantage for developing a new class of

Received: June 4, 2021
Published: November 5, 2021

https://doi.org/10.1021/jacs.1c05055



1a - MIC 128 μg/mL

1b - MIC 32 μg/mL

1c - MIC 8 μg/mL

1d - MIC 1 μg/mL

1e - MIC >64 μg/mL

1f - MIC 16 μg/mL

1g - MIC >256 μg/mL

1h - MIC 128 μg/mL

1i - MIC 16 μg/mL

1j - MIC 8 μg/mL

1k - MIC 16 μg/mL

1l - MIC 8 μg/mL

1m - MIC 4 μg/mL

1n - MIC 8 μg/mL

1o - MIC >256 μg/mL

1p - MIC 4 μg/mL

1q - MIC 2 μg/mL

1r - MIC 4 μg/mL

1s - MIC 32 μg/mL

1t - MIC 8 μg/mL

2a - MIC 2 μg/mL

2b - MIC >64 μg/mL

2c - MIC 16 μg/mL

2d - MIC 16 μg/mL

2e - MIC 2 μg/mL

2f - MIC 1 μg/mL

2g - MIC 64 μg/mL

2h - MIC 4 μg/mL

2i - MIC 8 μg/mL

2j - MIC 4 μg/mL

2k - MIC >256 μg/mL

2l - MIC 4 μg/mL

2m - MIC 4 μg/mL

2n - MIC ND

2o - MIC 2 μg/mL

2p - MIC 2 μg/mL

2q - MIC 1 μg/mL

2r - MIC 4 μg/mL

3a - MIC 32 μg/mL

3b - MIC 16 μg/mL

3c - MIC 8 μg/mL

3d - MIC >16 μg/mL

3e - MIC 32 μg/mL

3f - MIC 64 μg/mL

3g - MIC 64 μg/mL

4a - MIC >64 μg/mL

4b - MIC >64 μg/mL

4c - MIC 2 μg/mL

Figure 1. continued

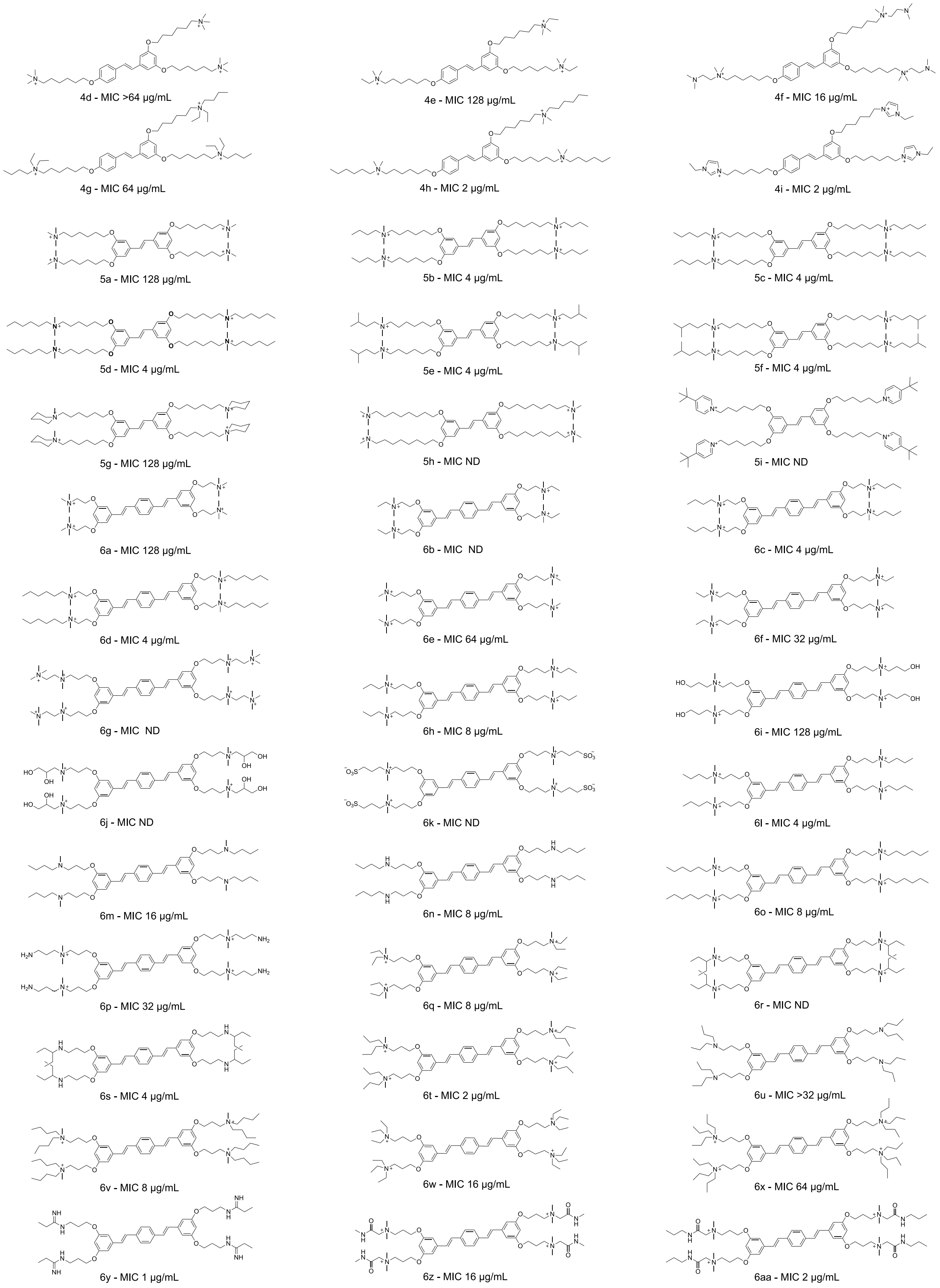
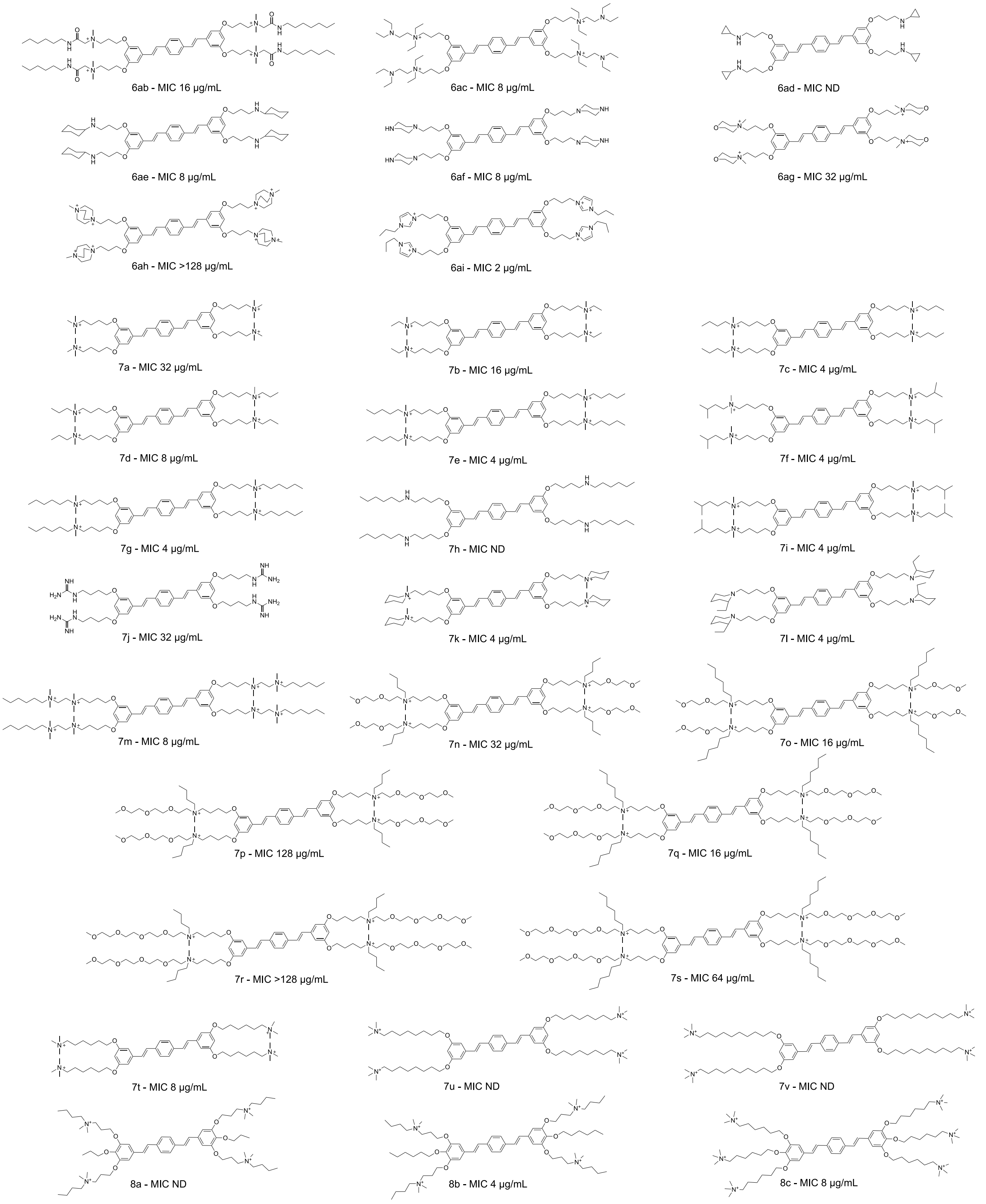


Figure 1. continued



6ab - MIC 16 μg/mL

6ac - MIC 8 μg/mL

6ad - MIC ND

6ae - MIC 8 μg/mL

6af - MIC 8 μg/mL

6ag - MIC 32 μg/mL

6ah - MIC >128 μg/mL

6ai - MIC 2 μg/mL

7a - MIC 32 μg/mL

7b - MIC 16 μg/mL

7c - MIC 4 μg/mL

7d - MIC 8 μg/mL

7e - MIC 4 μg/mL

7f - MIC 4 μg/mL

7g - MIC 4 μg/mL

7h - MIC ND

7i - MIC 4 μg/mL

7j - MIC 32 μg/mL

7k - MIC 4 μg/mL

7l - MIC 4 μg/mL

7m - MIC 8 μg/mL

7n - MIC 32 μg/mL

7o - MIC 16 μg/mL

7p - MIC 128 μg/mL

7q - MIC 16 μg/mL

7r - MIC >128 μg/mL

7s - MIC 64 μg/mL

7t - MIC 8 μg/mL

7u - MIC ND

7v - MIC ND

8a - MIC ND

8b - MIC 4 μg/mL

8c - MIC 8 μg/mL

Figure 1. continued

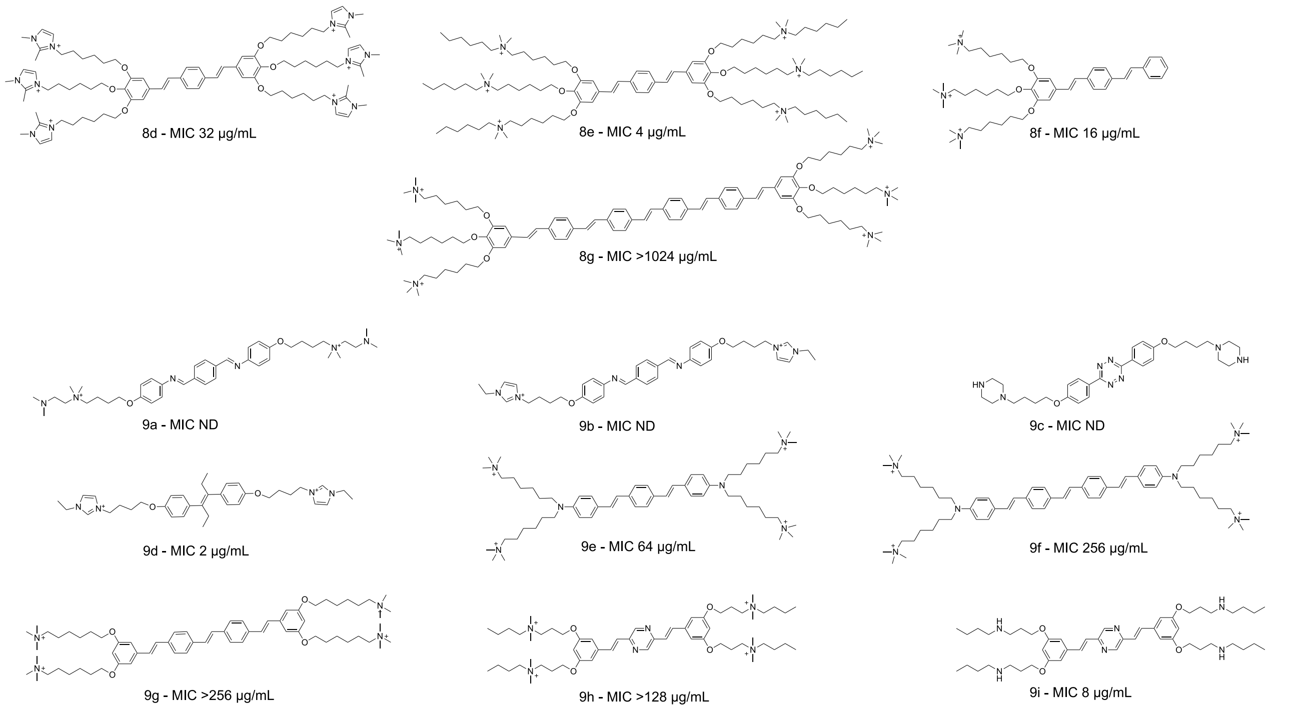


Figure 1. Molecular structures of COEs utilized in this work. Molecules are sorted into classes based on the core structure and ordered based on increasing length and complexity of the side chains. Counterions not shown. (1) disubstituted stilbenzene COEs, (2) disubstituted azobenzene COEs, (3) disubstituted styrylstilbenzene COEs, (4) trisubstituted stilbenzene COEs, (5) tetrasubstituted stilbenzene COEs, (6) tetrasubstituted styrylstilbenzene with ethyl/propylamine-based side chain COEs, (7) tetrasubstituted styrylstilbenzene with butylamine-based side chain COEs, (8) hexasubstituted styrylstilbenzene COEs, and (9) miscellaneous COEs that did not fall into the other categories. Experimentally determined MIC values against *E. coli* K12 are stated under each molecule (ND = not determined for that COE).

antibiotics, as the structures can easily be tailored to a desired physical property.[6−8] A common property of these compounds is their ability to intercalate and disrupt biological membranes, which is their presumed antibiotic mechanism of action.[9] This proposed mechanism is advantageous for developing new antibiotics, as bacteria are expected to be less likely to develop resistance to membrane disruptors, compared to specific receptors. However, there is an absence of a known specific binding site that may guide the design of chemical structures.

ML modeling of molecular properties is enabled by recent developments in applied ML, especially regarding deep learning and molecular representations as previously reviewed elsewhere.[10−1214] Molecular representations include string representations such as widely used SMILES (simplified molecular input line entry system)[15] or SELFIES (self-referencing embedded strings) designed to be used with ML methods.[16] Numeric vectors consisting of molecular descriptor values (also called features) have been utilized already before the era of ML in QSAR (quantitative structure−activity relationship) modeling[17] and have been formulated traditionally by feature engineering relying on domain knowledge or by attempting to form general fingerprints such as Morgan fingerprints.[18] Successful fingerprinting is a challenging task,[19] which is one of the reasons that molecular graphs networks and other learned representations have started to gain popularity over fingerprints. Molecular graph networks usually describe molecules with atoms as nodes and with bonds between the atoms as edges. Graph networks have resulted in comparative accuracy with fewer samples in some benchmarking tests,[20] but there are also opposing results.[21] ML model and molecular representation approaches can be closely tied together. For example, message passing neural networks are combinations of a molecular graph network and a neural network, in which information is shared between the near neighboring atoms by passing messages between them.[20,22,23] Tree-based models such as random forest regression (RF) or gradient boosting also remain as an effective approach. Progress in the model analysis also facilitates increasing use of applied ML methods. New ML model analysis methods alleviate the black box challenge of ML in cases when it is necessary to use models that are not directly interpretable. Shapley additive explanations (SHAP)[24−27] is one of such approaches utilized in this work. It is a game-theoretic approach that differs from traditional feature importance ranking within models by connecting optimal credit allocation from a model input feature with local explanations of the model.

These developments in molecular representations and ML approaches are beginning to impact antibiotics discovery. The progress has already led to the identification of a new antibiotic molecule: halicin had existed previously but was not known to act as an antibiotic before Stokes et al. identified it by using solely a combination of deep learning—directed message passing neural networks that are a combination of a feed-forward neural network and a molecule graph with directed message passing scheme[22]—and screening of large databases of thousands of existing molecules. Many of the approaches in ML-aided antibiotic property prediction have relied on large data sets and focus on finding new antibiotic candidates that resemble existing ones (i.e., have identical distributions). Such a setting is not readily available when considering new mechanisms of action or new molecular types as described above, while these approaches are gaining increasing attention in preclinical development.[2]

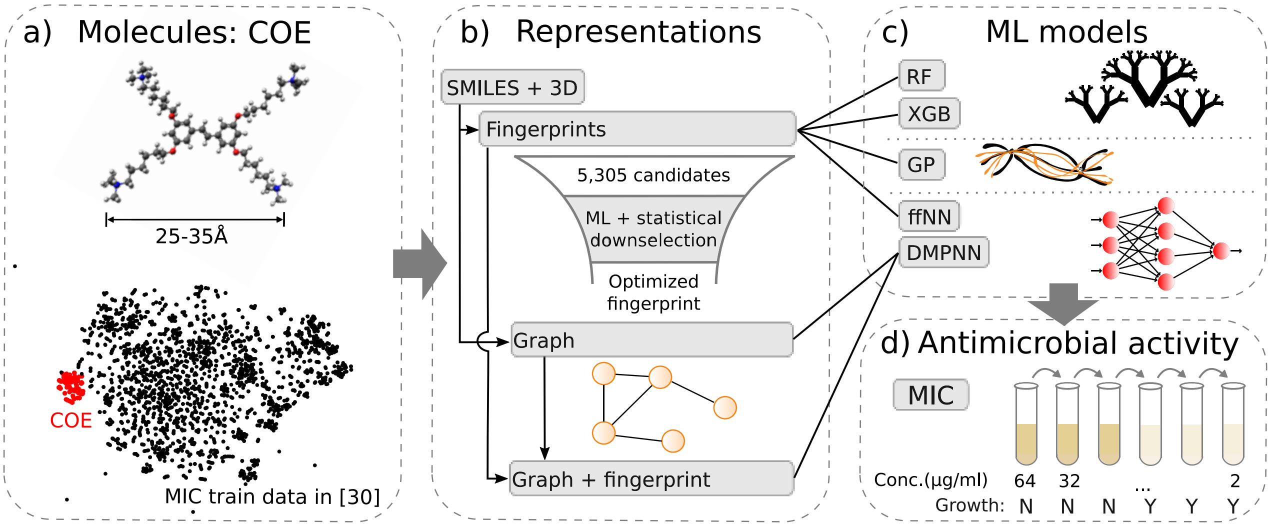


Figure 2. Predicting antimicrobial activity of conjugated oligoelectrolyte (COE) molecules involves optimizing the molecular representation for COE molecules and pairing that with a suitable machine learning (ML) model. (a) Example COE molecule (5a) and a similarity plot (t-distributed stochastic neighbor embedding with Morgan fingerprint and Jaccard similarity) with COEs (red) and an approximately 2000-molecule data set from Stokes et al.[30] (black) that they utilized for training a model to predict molecules with antimicrobial activity. (b) The molecular representation types investigated in this work. Molecular fingerprint candidates include a fingerprint optimized for COE by downselection from a base set of 5305 candidate descriptors. (c) ML models investigated in this work (definitions in the main text). (d) Minimum inhibitory concentration (MIC) for bacteria growth can be measured by diluting solutions until they cease to prevent *E. coli* K12 bacteria growth. Twofold dilution down to 1 μg/mL is one of the widely used approaches. The ML models are trained to predict the base-2 logarithm of the MIC.

Herein, we show an ML approach that may be applied to the challenge represented by new antibiotic classes, such as COEs. We present a framework for quickly establishing a predictive model of an antibiotic property. It consists of four components: (1) molecular representation, (2) feature downselection, (3) ML algorithm selection, and (4) molecular descriptor importance analysis. We apply this framework to conjugated oligoelectrolyte molecules, downselecting from 5305 features to 21 critical features governing antimicrobial activity. With only 136 compounds measured, we demonstrate antimicrobial activity prediction with an accuracy of $R^2 = 0.65$. This framework does not rely on prior domain knowledge and is therefore compatible with novel molecule domains. The trained model, together with molecular descriptor−importance analysis and domain expertise, could serve as a foundation for the accelerated development of novel COEs with enhanced antimicrobial activity.

2. RESULTS AND DISCUSSION

Synthesis of 136 Conjugated Oligoelectrolyte Molecules. An experimental data set covering 136 COE structures (Figure 1) was collected for this work. A total of 113 of the molecules are newly synthesized and presented for the first time in this work. The remaining molecules have been presented before[4,6−9,28] and are designated in the Supporting Information. The COEs were characterized by NMR or mass spectrometry (shown in Supplementary Section S14) to confirm the synthesized structures. The molecules were screened for antimicrobial activity against *E. coli* K12 bacteria by measuring minimum inhibitory concentration (MIC) for bacteria growth. The COEs are subdivided into nine categories based on the structure of the core and the number/composition of the pendant groups.

Principled Downselection Process. Our process for predicting the antimicrobial activity of COE molecules is illustrated in Figure 2. Figure 2a shows a similarity plot (t-distributed stochastic neighbor embedding with Morgan fingerprint and Jaccard similarity[29]) involving COEs and reference molecules that are approximately 2000 antibiotic candidate molecules previously tested for antimicrobial activity and used as a training data set by Stokes et al.[30] The data set was chosen as a reference due to its large size and diversity within the set. Analysis shows that COE molecules form a structurally distinct group of molecules compared to many other molecules that have previously been screened for antimicrobial activity, which warrants the development of a predictive ML model designed specifically for COEs. One of the differences is size. Most existing antibiotics and many reference molecules included into the similarity plot of Figure 2a are small compared to COEs (that are illustrated in Figure 1).

First, a molecular representation that is capable of capturing, and ideally even simplifying, the complexity of the structural and/or chemical factors driving the molecular activity is required. Second, this representation needs to be paired with a matching ML model. Here, we compare multiple molecular representation options: molecular fingerprints optimized for COEs by downselecting descriptors as well as existing fingerprints from the literature (described later), molecular graphs, and combinations of both (Figure 2b). In this work, we define a fingerprint as a molecular descriptor vector that can contain any numeric descriptors in addition to binary ones. The representations are paired with decision tree-based ML models, a kernel method, and neural networks (Figure 2c) to cover a range of widely used and advanced ML model options. We train the models to predict antimicrobial activity against *E. coli* K12 bacteria, which has been measured experimentally as MIC for bacteria growth (Figure 2d). The ML models are trained to predict antimicrobial activity in the form of $\log_2\left(\frac{\text{MIC}}{1\ \mu\text{g / mL}}\right)$ due to the base-2 exponential nature of MIC

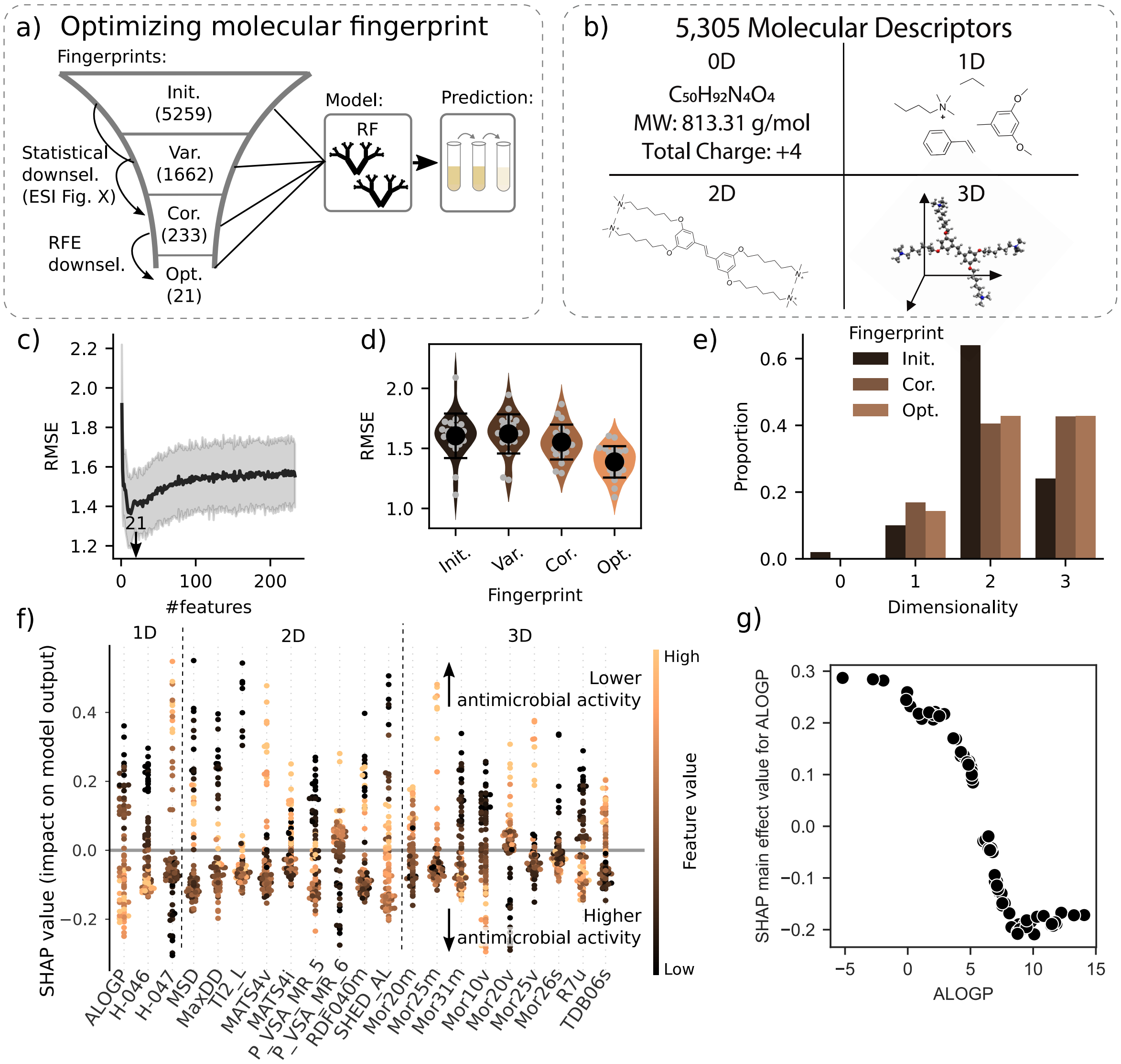


Figure 3. Optimization of the molecular fingerprint for COEs improves the accuracy of the random forest regression (RF) model in predicting antimicrobial activity. (a) Downselection process from a candidate descriptor set involving two stages of statistical downselection and one stage of ML downselection via recursive feature elimination (RFE). Fingerprints are defined in the main text, and the lengths are shown in parentheses. (b) Init. data set included descriptors related to zero- to three-dimensional shape of the molecules (explanations in the main text). (c) Root-mean-square error (RMSE; mean and standard deviation, std., are shown) during RFE shows a minimum error region centered at 21 descriptors, which are chosen as the Opt. fingerprint. RMSE is from cross-validation with 20 stratified subsampling repeats of the training data set. (d) RMSE for RF models trained with each fingerprint. Violins represent the distributions of the values; individual subsamples are shown in gray, and mean and std. of RMSE in black. (e) Proportion of descriptors related to the 1D−3D properties of the molecules in the specified fingerprints. (f) Shapley additive explanations (SHAP) analysis of the training data set molecules and the final RF model (trained with the whole training data set and Opt. fingerprint). Negative SHAP values push the model prediction toward low MIC (high antimicrobial activity), and vice versa. The AlvaDesc naming convention is used for the descriptors. (g) SHAP value for the Ghose−Crippen octanol−water partition coefficient (ALOGP) of the training set molecules as a function of descriptor value.

values. The lower the model output value is, the higher the predicted antimicrobial activity.

We applied data curation and dropped potential outliers (5 molecules) from the data as pretreatment steps (described in Methods and Supporting Information Section S13.1). These steps are crucial when treating experimental data sets, especially small data sets, because a few outliers might distort the model significantly. The molecules with experimentally measured MICs were split into 80%/20% train/test data sets. Throughout the results section, the performances of models and fingerprints are compared based on cross-validation with 20 stratified subsampling repeats of the training data set (described in Section 4).

Optimized Molecular Fingerprint for Conjugated Oligoelectrolytes. We began by optimizing a molecular fingerprint for COE. We aimed to avoid biasing our fingerprint toward any potential mechanism of antimicrobial activity by not using expert knowledge in the fingerprint-selection process. Instead, the process consists of multiple stages of automated molecular-descriptor downselection, starting from a broad set of candidate descriptors (see Methods for detailed descriptions of the steps and Figure 3a for process illustration). We initially generated 5305 molecular descriptor candidates with the chemistry analysis software AlvaDesc from the SMILES strings and 3D structures of each COE. The set is a diverse combination of descriptors related to molecule chemistry and structure (full list shown in ref 31 and more information in ref 32). Descriptors can be classified based on their dimensionality when considering how they approximate the three-dimensional (3D) shape of a molecule (Figure 3b). Our candidate descriptors include 0D (with no relation to shape, e.g., molecular weight), 1D (e.g., presence of certain active substructures within the molecule), 2D (e.g., molecular graph representations involving bonds between atoms but not bond lengths), and 3D (e.g., distances between certain atomic pairs in the molecule) ones. This broad set of descriptors is likely to contain information relevant for the antimicrobial action of a given COE. However, ML models are prone to overfitting with a large number of input parameters and small training data. This limits the options of viable ML models and is also likely to reduce the predictive accuracy of those options that remain to be viable. Therefore, we aimed for removing redundant and irrelevant descriptors that the 5305-descriptor data set would be anticipated to contain.

Of the 5305 total descriptors, 5259 are numerical descriptors, which we refer to as initial fingerprint (Init.). We applied two stages of statistical downselection and one stage of ML-aided descriptor downselection (Figure 3a): Constant or almost constant descriptors were filtered out (variance downselection with relative limit of 0.1, Var. fingerprint) using the whole COE data set. Next, the descriptors with high Spearman rank correlations between each other or a low correlation with MIC were dropped (correlation downselection with limits of 0.9 and 0.05, respectively, Cor. fingerprint). Conservative variance and correlation limits were chosen for statistical downselection. Nevertheless, they allow for drastic reduction in the number of descriptor candidates (from 5259 to 1662 and then to 233), while restricting the likelihood of losing relevant descriptors (the resulting correlation matrices are shown in Supplementary Figure S1). The final, ML-aided stage of downselection was recursive feature elimination (RFE),[33] which eliminates features based on feature importance rankings of the random forest regression. The same kind of approach could have been adopted also with other ML model types; however, RF is a good option due to its straightforward training and feature importances. Wrapper methods, such as RFE, ensure that the resulting fingerprint carries the most important information by actually training ML models. RFE should complement the statistical filtering stages that ensure the breadth of information among the descriptor candidates. RFE was applied on the Cor. fingerprint and training data set in a mode that drops one descriptor at each recursion; see Methods for detailed information. The root-mean-square error (RMSE) graph in Figure 3c shows that the optimum predictive accuracy of RF is reached in a region centered to a fingerprint length of 21 descriptors. These descriptors were consequently chosen as the final fingerprint optimized for COEs (Opt., see Supporting Information Section S2 for descriptors).

Next, RF models trained with the downselection fingerprints are investigated for the purposes of comparing the fingerprints. The cross-validation RMSE of the resulting RF models decreases along the downselection stages (Figure 3d), showing that the downselection process improves the fingerprints. The Opt. fingerprint provides average RMSE improved by 20% in comparison to the Init. fingerprint (additionally, the test data set predictions from the RF models trained with the full training data set are shown in Supplementary Figure S2). While the COE data set may be extensive from a synthetic viewpoint, it remains rather limited in size compared to typical data sets for ML. Thus, the consistency of descriptor downselection results needs to be evaluated carefully. We repeated the RFE with another tree-based ML model, gradient boosting regression (XGB). It resulted in a very similar downselected fingerprint (Supporting Information Section S4), thus suggesting that the results are not an arbitrary result of the chosen surrogate ML model.

Analysis of Molecular Fingerprint Optimized for Conjugated Oligoelectrolytes. The molecular fingerprint optimization process is useful not only because it optimizes predictive accuracy of the resulting ML model but also because the fingerprint content provides clues for understanding the underlying mechanisms of action for the molecules. Figure 3e illustrates that the Opt. fingerprint consists of 1D−3D descriptors, with a large proportion of them being a 3D. Cor. fingerprint, and also links directly to the MIC via statistical correlations and has similar proportions. A large proportion of 3D descriptors compared to the Init. fingerprint survives until the last stages of downselection, which suggests that the antimicrobial activity of COEs is influenced by 3D molecular shape. This supports the presumed antibiotic mechanism of action for COEs, that is, the disruption of microbial membranes.[9] In our work, 3D shapes of COEs have been determined by molecular dynamics simulations with a relatively fast method that can contain small differences of the actual shape of the molecule. This fact, together with the data set size, limits the extent of conclusions we can draw in this study based on individual 3D descriptors, but some observations seem consistent when analyzing the whole fingerprint.

Eight descriptors, more than one-third of the Opt. fingerprint, are 3D-MoRSE descriptors (molecular representation of structures based on electronic diffraction[34−36]). These descriptors sum up contributions from each atom of the molecule based on pairwise distances of the atom pairs. Here, MoRSE descriptors with scattering factors 31, 26, 25, 20, and 10 $Å^{-1}$ survived the downselection and are weighed by atomic mass (m), van der Waals volume (v), or intrinsic I-state (s). The high-order scattering factors suggest that subtle changes in the distances of atomic pairs (e.g., $2\pi/20$ Å $\approx$ 0.3 Å) and short-range interactions overall might affect the antimicrobial activity of COEs. The rest of the 3D descriptors in the Opt. fingerprint are a GETAWAY descriptor and a topological distance descriptor, both related to autocorrelations.[37]

From a molecular property prediction viewpoint, the large proportion of 3D descriptors is notable because 3D molecular representations are not commonly used in antimicrobial activity prediction. Additionally, most of the existing such works focus on peptides[38,39] instead of small molecules. There are several reasons for the present scarcity of the use of 3D representations. First, many of the existing known antibiotic molecules are small and may operate via mechanisms of action that are dependent on intermolecular interactions, e.g., via targeting certain enzymes.[40−42] Thus, 3D representations may not have as strong an effect when searching for molecules with this type of operation, since they relate more to intramolecular interactions. The use of 3D molecular representations may not be necessary in cases when molecules with similar operation to the predecessors are searched for. Second, the benefits of 3D

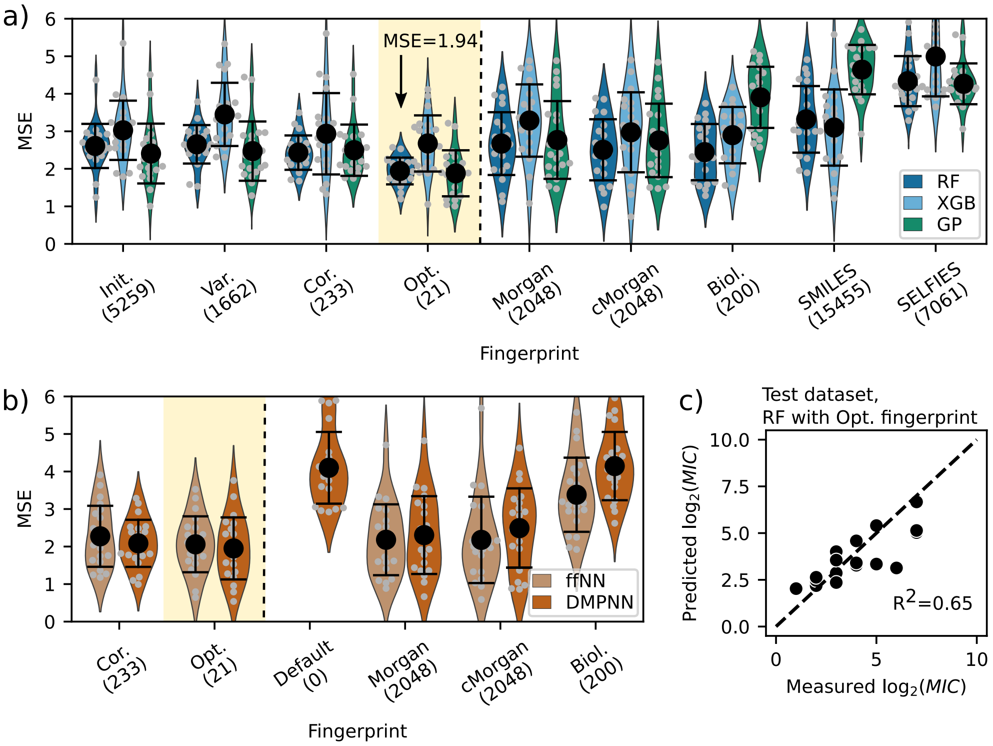


Figure 4. Comparison of different combinations of molecular representation and machine learning models, to predict antimicrobial activity of COEs. Mean-square error (MSE) for (a) random forest (RF), gradient boosting (XGB), and Gaussian process (GP) regression models, as well as (b) feed forward neural network (ffNN) and directed message passing neural network (DMPNN) models trained with each fingerprint option. Fingerprints (lengths in parentheses) and models are defined in the main text. MSE values are computed from 20 stratified subsampling repeats of the training data set. Violins represent the distributions of the subsampling results, mean and standard deviation of MSE are shown in black, and individual subsample results are in gray. The high-MSE tails of the distributions are clipped on MSE axes to highlight differences between the mean values. Full data are shown in the SI Figure S5) Test data set prediction with the highest-performing model, RF with the Opt. fingerprint, trained with the whole training data set (test data set withheld from training).

representations in molecular property prediction generally are currently not fully clear. For example, 3D molecular graph networks have resulted in a similar or even decreased accuracy with many prediction tasks focusing on small molecules[20,23] but nevertheless improved predictions of some, e.g., quantum mechanical molecule properties[20] derived directly from the physical properties of the molecules. Performance has also depended on the exact model−representation combination.[23] Third, generating 3D structures of molecules requires additional effort and computation time in comparison to representations that can be generated solely from, for example, SMILES strings. Our results show that the effort is worthwhile in the case of COEs. This result may generalize into other antibiotic candidate domains with larger molecular structures. Incorporating the 3D shape of the molecules should therefore be considered in these cases.

To gain chemical insight into the relative contributions of different features, we investigated the descriptors in the Opt. fingerprint set with SHAP analysis. The method itemizes the contributions from each molecular descriptor to the predicted antimicrobial activity of each molecule when the ML model makes predictions. In Figure 3f, SHAP is applied to analyzing predictions of the training data set molecules using the RF model trained with the Opt. fingerprint (which also coincides with the best-performing model−representation pair for COEs as we show later). There are no clear dominating descriptors, but the high antimicrobial activity builds up from multiple contributions. The impacts from different descriptors are not reaching consistently high or low within the same molecule, either. Instead, for most COEs, some of the descriptors reduce the model's predicted MIC value, while some other descriptors increase it. The complex feature interactions hinted by SHAP suggest why the antimicrobial activity of a COE is challenging to predict. Similar reasons may explain the complexity of predicting antimicrobial activity in other antibiotic candidate domains.

We highlight the behavior of one of the Opt. descriptors, the water−octanol partition coefficient (ALOGP), to suggest future direction for ML in antibiotics development. SHAP analysis shows that ALOGP should be high in order to reach high antimicrobial activity (Figure 3g). The observation is, again, in line with the presumed mechanism of action for a COE: a high ALOGP indicates high lipophilicity of the molecule, which is a major factor determining how easily the molecules penetrate lipid bilayers. The ML model trained only for antimicrobial activity could not, however, recognize that the guideline of high lipophilicity is in practice bound by other antibiotic development criteria,[43] e.g., cytotoxicity to human cells. This is because too high lipophilicity tends to lead to less targeted effects in the human body.[44] The case of ALOGP exemplifies that when investigated from the viewpoint of accelerating antibiotics development pipelines, ideal ML models do not only max out a single antibiotic development criterion but guide efforts toward candidates that fulfill several criteria adequately. Therefore, continued efforts to develop reliable multitask models[45−49] are valuable even though they require more training data than what is available for COEs at the moment.

Comparison of Molecular Representation−Machine Learning Model Pairs. While the preceding analysis focused on the optimized molecular fingerprint, and the RF model provided clues for factors contributing to COE activity, further comparisons of ML models and molecular representations are required to determine which combination has the highest predictive accuracy of antimicrobial activity for COEs. As illustrated in Figure 2, ML models tested include tree-based ML models (RF and gradient boosting regression, XGB), a kernel method (Gaussian process regression, GP), and neural network models (feed-forward neural network, ffNN, and directed message passing neural network, DMPNN). It should be noted that DMPNN is a combination of a molecule graph and ffNN; therefore all the combinations of a DMPNN and a fingerprint involve also a molecule graph.

The ML models combined with the downselection fingerprints described above show improved predictive accuracies with the progressing stages of downselection (fingerprints Init. to Opt. in Figure 4a and fingerprints Cor. and Opt. in Figure 4b) with the only exception of GP. The consistent improvement suggests that optimizing a fingerprint with an easily tunable model, here RF, and shifting to the final model of choice only after that, is a viable strategy to reduce model training difficulty. This is important because increasingly complex ML models, for example a range of deep-learning models, have been utilized in molecular property predictions lately. While they have provided convincing predictive accuracies, their performance may depend heavily on hyperparameter tuning. Thus, comparing and training them may be challenging,[12,50] in which cases using a different model for optimizing fingerprints is convenient. Init. and Var. fingerprints have been applied only to tree-based and kernel models (Figure 4a) and not to neural network models (Figure 4b) because of their length. It is unlikely that neural-network-based models with very long fingerprints would have been able to fully converge into proper neuron weights with the amount of COE training data available, and therefore the results would not have been quantitatively representative of how much the fingerprint had improved.

We compared ML models with multiple molecular fingerprints used in the literature: Morgan and Morgan count (cMorgan) fingerprints are extended connectivity fingerprints[18] that have been widely applied for small-molecule property prediction.[19] The fingerprint we denote as Biol. has been developed by Yang et al.[22] and utilized for biological targets, including successful application to antimicrobial activity prediction of mainly small molecules by Stokes et al.[30] SMILES and SELFIES fingerprints are one-hot-encoded molecule string representations of SMILES and SELFIES strings, respectively. These 2D fingerprints were tested also together with a molecular graph representation using DMPNN models. Finally, DMPNNs in Figure 4b were also trained using solely a 2D molecular graph representation without an additional fingerprint (Default), as illustrated in Figure 2. The molecular graph involves by default 10 different basic atomic and bond type descriptors that are utilized in the message-passing scheme of the DMPNN operation. It should be noted that Figure 4 is not meant for universally comparing the ML model−representation combinations, because the performance depends heavily on the domain and data set.

Figure 4a,b show that the best performance for a COE among all the combinations is achieved with the Opt. fingerprint combined with either the GP or RF model. RF results in lower variance in performance, and therefore it was chosen as the final model. RF with the Opt. fingerprint gives MSE = 1.94. To give a reference to experimental units, the MSE transforms to RMSE of approximately 2.6 μg/mL when the order of magnitude of the MIC is 2 μg/mL, or 330 μg/mL when the order of magnitude for the MIC is 256 μg/mL. With the final model (trained with the full training data set), a test data set prediction with $R^2 = 0.65$ is achieved (Figure 4c). The resulting model is capable of singling out most of the low-activity molecules that are not promising for further investigation ($\log_2(\text{MIC}) \geq 4$), which already facilitates synthesis work in a novel molecular domain. We further evaluated the ML model from the viewpoint of prospective use in predicting real-life data and as a part of a molecule optimization cycle, which are described in detail in Supporting Information Sections S11 and S12, respectively. We show that the ML modeling approach (1) performs well when the regressor is adapted as a classifier that aims to minimize false positives (so we do not expend experimental bandwidth on poor candidates) and (2) outperforms random sampling (e.g., when we know little about a new molecular system) when utilized in a pool-based active learning loop[51] utilizing Bayesian optimization algorithm.[52] These results provide further reason to believe that the ML model with the Opt. fingerprint is useful in the laboratory for the discovery of new antimicrobial COEs.

Figure 4 shows that RF performs consistently well with a range of different fingerprint options. GP provides narrowly the best MSE with Init. to Opt. fingerprints but results in larger variance. All the models result in lower MSE with the Opt. fingerprint compared to other fingerprint options (test data set predictions shown in Supplementary Figure S7). The performance of DMPNN and ffNN models is clearly improved by using the Opt. fingerprint: the reference fingerprints result in large variations in model performance within the repeated subsampling folds, many of which result in practically nonusable DMPNN and ffNN models with MSE > 4—a model that would always output training data set average MIC would result in an MSE in this range—whereas with Opt. they perform close to RF and GP models with an MSE = 1.95 and MSE = 2.06, respectively, albeit with clearly higher variance. This demonstrates that with an optimum fingerprint, a neural network-based model can be viable even with a relatively small data set. DMPNN might turn out to perform better than RF or GP in predicting COEs when the size of the data set increases in the future. DMPNN results in slightly lower MSE than ffNN even though the applied molecular graph message passing scheme is initially designed for small molecules.[22] In the DMPNN-Opt. model trained here, messages are passed between the five nearest neighbor atoms in the molecule graph, which is a relatively large neighborhood area. DMPNN with a 2D molecular graph representation alone (Default) does not, however, seem to provide sufficient information to predict antimicrobial activity of COEs, but needs to be combined with a supporting fingerprint (Opt.).

Many of the fingerprint−model combinations in Figure 4 result in high prediction errors. This happens either because the molecule representation is too long compared to the size of the data, resulting in an under-determined problem for the model, or because the fingerprint does not contain enough relevant information. The Opt. fingerprint (alone or combined with a molecular graph) provides consistently the best result among the investigated fingerprints. Opt. is a short fingerprint optimized for COEs and includes many 3D descriptors in contrast to the reference fingerprints on the right-hand side of the dashed lines in Figure 4a,b, all of which are longer and more general fingerprints containing only 2D information. Figure 4 shows that the choice of molecular representation drives the predictive accuracy even more than the choice of ML model, which has been suggested also by Wu et al.[20] The results highlight the importance of adapting the molecular representations for the target domain, even when the mechanisms of action governing the antibiotic activity are not thoroughly known.

3. CONCLUSIONS

We developed a model that predicts antimicrobial activity of conjugated oligoelectrolyte molecules, which can be utilized as an advisory tool to guide the synthesis of newly predicted COEs. This model was enabled by a framework that consists of four parts: (1) molecular fingerprint representation, (2) feature downselection, (3) ML model pairing, and (4) descriptor importance analysis. We applied this framework to a set of 136 COEs, using an automated molecular descriptor downselection process that is agnostic to the molecule domain. The resulting fingerprint consisted of 21 molecular descriptors, over 40% of which are related to the three-dimensional shape of the molecules. This is in contrast to descriptors that would relate to molecular properties not dependent on shape or descriptors related to lower-dimensional simplifications of molecular shape, such as molecule bonds with no length information. This is consistent with the presumed mechanism of action for COEs, which is intercalation into the bacterial membrane. Our results suggest the connection of 3D shape to the antimicrobial activity of COEs should be investigated further with detailed mechanistic experiments and accompanying modeling.

By developing the predictive model, we have demonstrated that ML can indeed aid development of antibiotics even in novel domains, namely, with families of understudied candidates, where sparse information exists regarding the underlying mechanism and there is a limited availability of experimental data. We have shown that molecular representation drives predictive accuracy in both traditional and complex ML models: it needs to capture the critical information about the underlying mechanisms of activity of the antibiotics in order to achieve high predictive accuracy. Self-learning representations have provided impressive results in molecular property prediction lately. We suggest that fingerprinting remains a valid option or a complementary component to learned representations, provided that fingerprints are optimized for the molecular domain under investigation. This is especially true in situations where there is limited data.

In closing, we presented a domain-agnostic framework to downselect and analyze a molecular fingerprint quickly, to describe novel classes of antibiotics. We investigated only one class of molecules, COEs, in this work, but 3D descriptors being so dominant in the fingerprint downselection suggests that representations and models capable of capturing 3D may be worth exploring also in other antibiotic candidate domains, particularly when the molecules are large and/or the mechanisms of action connect to molecular shape as is suspected here.

4. METHODS

4.1. Synthesis and Characterization of Conjugated Oligoelectrolyte Molecules.

The synthetic procedure of COEs refers to previous literature.[6−8] Generally, the alkylation steps were accomplished by Williamson ether synthesis with a carbonate base. The conjugated backbones of COEs were composed via Horner−Wadsworth−Emmons or McMurry coupling reactions. After final quaternization of the terminal alkyl halide groups with excess trimethylamine or other amines, the targeting COEs were obtained. Intermediates and COE products were purified using multiple approaches (including liquid−liquid extraction, column chromatography, precipitation, solvent removal under vacuum, etc.) and then characterized by NMR spectroscopy or mass spectrometry (see the Supporting Information for details).

4.2. Forming the Data Set.

MIC values for COEs were experimentally determined against *E. coli* K12 (ATCC 47076) using the broth microdilution method as previously described in ref 48. COE structures were converted into SMILES strings using ChemDraw v 19.0. SMILES strings were used in Avogadro 1.2.0 to generate 3D models of each COE. Molecular geometries were then optimized using the MMFF94 force field in Avogadro. The energy-minimized 3D structures were exported as .cml files and uploaded into alvaDesc v1.0.20 for molecular descriptor calculations. A total of 5305 1D, 2D, and 3D molecular descriptors were calculated for each COE.

4.3. Data Preprocessing.

The data set was preprocessed by dropping non-numeric descriptors. During the broth microdilution experiment, solutions of doubling concentrations are examined for bacterial growth. In cases where the end point of bacterial growth had not been met during the experiment, the MIC value was estimated to be the next largest available value (e.g., when experiment resulted in MIC > 256 μg/mL, the value was estimated as MIC = 512 μg/mL). A $\log_2()$ transformation was performed on MIC values since all the MIC values were multiples of two. All the graphs in this work show $\log_2\left(\frac{\text{MIC}}{1\ \mu\text{g / mL}}\right)$ values. It should be noted that due to the discrete nature of MIC results, the effective uncertainty for MIC measurement is typically one base-2 order of magnitude and arbitrarily high in the high end for those molecules that have open-ended MIC measurement results (MIC > X μg/mL). Finally, we examined the data for crude outliers by performing leave-one-out cross-validation for the whole data set using RF models with fixed hyperparameters. For each fold, we trained 10 RF models with different initial random states and evaluated the mean prediction error value. The molecule was dropped from the data set as an outlier if the prediction error was higher than 3.5 $\log_2\left(\frac{\text{MIC}}{1\ \mu\text{g / mL}}\right)$ units, indicating that MIC prediction was more than 10 times off. There were in total five such molecules in the original data set used for ML model development, which are detailed in Supporting Information Section S13.1. After the ML model had been developed, it was tested on new COE molecules (see Methods for the division of the data set into subsets). Three of these molecules were additionally flagged by the outlier detection method but not excluded.

4.4. Statistical Downselection of Molecule Descriptors.

We applied a multistage molecular descriptor filtering process to the data set. The starting point of the filtering process was the data set with all the numerical descriptors (Init.). We started by filtering out descriptors with variance less than 10% of the mean value across the whole data set (Var.). We continued by filtering out one descriptor of the descriptor pairs that had higher than 0.9 correlation with each other as well as descriptors that had lower than 0.05 correlation with $\log_2$(MIC) (Cor.). All the correlations were evaluated as Spearman rank correlations, which evaluate any monotonic relationships, whether they are linear or nonlinear ones. Correlation matrices for each fingerprint set are shown in Supporting Information Figure S1.

4.5. Division of the Molecule Data Set into Subsets.

The full data set is 136 molecules. It is divided into five parts: empty MIC, outliers, training, test, and new data to which the model is applied. At the time of initial model fitting, 33 of the molecules did not have measured MIC available and were excluded from the model fingerprint downselection and ML model training procedures. Molecules without MIC measured were still included into the data set collected in this work to provide a comprehensive picture of the COEs synthesized to date. A further five molecules were excluded during outlier detection (see Methods for data preprocessing). We designated 20% of the remaining molecules with valid MIC entry, totaling 20 molecules, into a test data set that was held out of the model training pipeline. The remaining 78 molecules were included in the training data set. For 15 of the molecules that did not initially have MIC measured, the measurement was performed during the ML development. Therefore, the real-life use of the fully optimized ML model in predicting any COEs chosen by a scientist was tested with these 15 new molecules as described in Supporting Information Section S11. The similarity of the outlier, training, test, and new data sets is evaluated using t-SNE similarities with the Opt. fingerprint in Supporting Information Section S10.

Throughout the whole ML development phase, we utilized cross-validation with 20 random but stratified subsampling repeats for training and estimating the accuracy of the model−representation pairs. Each subsample was divided into 20% validation set and 80% training set. The subsampling stratification for RF, XGB, and GP was performed based on high/intermediate/low MIC classes, and for DMPNN and ffNN models a similar balancing was implemented (both described in Supporting Information Section S13.2).

4.6. Machine Learning Models and Hyperparameter Optimizations for Them. RF and GP were implemented using Scikit-learn,[53] XGB using XGBoost,[54,55] and ffNN and DMPNN using Chemprop.[56] Hyperparameter optimizations were performed using the training data set on a high-performance computing server.[57] Hyperparameters for RF and XGB were optimized with Bayesian optimization with 20-fold cross-validation RMSE as a target. A Python package BayesianOptimization[58] was utilized. The GP surrogate model and expected improvement acquisition function were used for Bayesian optimization with 50 random initialization points, 300 iterations of optimization, and three restarts of the optimization. GP was implemented using Scikit-learn, and this implementation had internal hyperparameter optimization with a Newtonian kind of limited-memory BFGS algorithm.[59] GP had a Matern kernel and 50 restarts of the optimizer. Hyperparameters for the neural network models were optimized using internal implementation of chemprop, which also relies on Bayesian optimization. A total of 800 iterations and scaffold-balanced splits were utilized. The search spaces were expanded for the depth of the network (max. 12) and number of ffNN layers (max. 5). The optimized hyperparameter values are included in the GitHub repository linked to this work.

4.7. Recursive Feature Elimination. Recursive feature elimination was applied on hyperparameter-optimized RF with the Cor. molecular fingerprint. RF was chosen as the base model for RFE because it is fast to ramp up and can be used without scaling the inputs. We applied a shallow RFE with cross-validation; that is, RFE was first performed on the whole training data set with steps of recursively eliminating one descriptor at a time until only one descriptor was remaining. After that, each stage was evaluated with 20-fold repeated stratified subsampling. The optimized molecular fingerprint (Opt.) was determined as the optimum arising from bias-variance trade-off, i.e., as the descriptor set that was at the center of the valley that resulted in the lowest mean RMSE.

4.8. Molecular Fingerprints. Init., Var., Cor., and Opt. fingerprints were defined as described above. Mor., cMor., and Biol. fingerprints were computed with the chemprop Python package,[56] which internally runs the rdkit package.[60] SMILES were one-hot-encoded using our own implementation, and SELFIES fingerprints were computed using the selfies package.[61] Molecular fingerprints are used as such for RF and XGB, since these tree-based models do not require scaling of input features. GP, DMPNN, and ffNN were scaled to zero mean unit variance.

4.9. Shapley Additive Explanations. SHAP analysis was performed for the final RF model trained with the whole training data set using the SHAP Python package.[24,25,62] The molecules analyzed were the training data set. Prior to the analysis, it was confirmed that the molecular descriptors in the Opt. fingerprint did not have very high correlations with each other, which could distort the analysis (see Supplementary Figure S1). SHAP provides a more fine grained method for analyzing model operation compared to RF feature importance ranking (see Supplementary Figure S6) because the direction of the effects can be recognized.

4.10. Data and Code Availability. The codes for fingerprint optimization and running the comparisons between the molecule representation−ML model pairs are available in the GitHub repository: https://github.com/PV-Lab/MLforCOE. Details of the molecules synthesized are listed in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.1c05055.

> Additional experimental details, materials, and methods, including information on the descriptors at each downselection stage, test data set predictions of the fitted models, detailed description of outlier selection and subsampling processes, and experimentally measured minimum inhibitory concentrations as well as NMR or mass spectrometry spectra of the synthesized molecules (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Armi Tiihonen − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-9753-6802; Email: armi.tiihonen@gmail.com, tiihonen@mit.edu

Sarah J. Cox-Vazquez − *Departments of Chemistry and Chemical & Biomolecular Engineering, National University of Singapore, Singapore 119077, Singapore;* Email: sjcoxvazquez@nus.edu.sg

Guillermo C. Bazan − *Departments of Chemistry and Chemical & Biomolecular Engineering, National University of Singapore, Singapore 119077, Singapore; Center for Polymers and Organic Solids, Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, California 93106, United States;* orcid.org/0000-0002-2537-0310; Email: chmbgc@nus.edu.sg

Tonio Buonassisi − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Singapore MIT Alliance for Research and Technology, Singapore 138602, Singapore;* orcid.org/0000-0001-8345-4937; Email: buonassi@mit.edu

Authors

Qiaohao Liang − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Mohamed Ragab − *School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore*

Zekun Ren − *Singapore MIT Alliance for Research and Technology, Singapore 138602, Singapore*

Noor Titan Putri Hartono − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Zhe Liu − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* Present

Address: School of Materials Science and Engineering, Northwestern Polytechnical University (NPU), Xi'an, Shaanxi 710072, People's Republic of China

Shijing Sun − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Cheng Zhou − *Departments of Chemistry and Chemical & Biomolecular Engineering, National University of Singapore, Singapore 119077, Singapore*

Nathan C. Incandela − *Center for Polymers and Organic Solids, Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

Jakkarin Limwongyut − *Center for Polymers and Organic Solids, Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

Alex S. Moreland − *Center for Polymers and Organic Solids, Department of Chemistry and Biochemistry, University of California, Santa Barbara, Santa Barbara, California 93106, United States*

Senthilnath Jayavelu − *Institute for Infocomm Research, Artificial Intelligence, Analytics and Informatics, Agency for Science, Technology and Research, Singapore 138632, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.1c05055

Notes

The authors declare the following competing financial interest(s): The authors declare competing financial interests to include the filing of intellectual property regarding machine learning methods in general (but not the methods submitted with this paper, which are open-sourced), and start-ups concerning the use of said models and COEs.

■ ACKNOWLEDGMENTS

We thank Geraldine W. N. Chia and Chenyao Nie for performing MIC measurements. We thank Tian Xie, Allison Tam, and Regina Barzilay for providing the trained ML model from Stokes et al.[30] and for fruitful discussions. We thank Akshat Nigam, Daniel Flam-Shepherd, and Alán Aspuru-Guzik for fruitful discussions regarding path message passing neural network models. We thank Gemma E. Moran for fruitful discussions and providing minimum working example code related to DMPNN modifications. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper. Q.L. acknowledges generous funding from TOTAL S.A. for supporting his graduate research. Z.R. is supported by the National Research Foundation, Prime Minister's Office, Singapore, under its Campus for Research Excellence and Technological Enterprise (CREATE) program through the Singapore Massachusetts Institute of Technology (MIT) Alliance for Research and Technology's Low Energy Electronic Systems research program. A.T., N.H, Z.L., S.S., and T.B. acknowledge support from DARPA under Contract No. HR001118C0036, and TOTAL S.A. S.J. acknowledges funding from the Accelerated Materials Development for Manufacturing Program at A*STAR via the AME Programmatic Fund by the Agency for Science, Technology and Research under Grant No. A1898b0043. Work at NUS was supported through a Startup Grant to G.C.B. (R-143-000-A97-133). Work at UCSB was supported by the Institute for Collaborative Biotechnologies through grant W911NF-09-D-0001.

■ REFERENCES

(1) WHO. No Time to Wait: Securing the Future from Drug-Resistant Infections. *Report to the Secretary-General of the United Nations*; World Health Organization: Geneva, 2019.

(2) Theuretzbacher, U.; Outterson, K.; Engel, A.; Karlén, A. The global preclinical antibacterial pipeline. *Nat. Rev. Microbiol.* **2020**, *18*, 275−285.

(3) Coates, A. R.; Halls, G.; Hu, Y. Novel classes of antibiotics or more of the same? *Br. J. Pharmacol.* **2011**, *163*, 184−194.

(4) Zhou, C.; Chia, G. W. N.; Ho, J. C. S.; Moreland, A. S.; Seviour, T.; Liedberg, B.; Parikh, A. N.; Kjelleberg, S.; Hinks, J.; Bazan, G. C. A Chain-Elongated Oligophenylenevinylene Electrolyte Increases Microbial Membrane Stability. *Adv. Mater.* **2019**, *31*, 1808021.

(5) Wang, B.; Wang, M.; Mikhailovsky, A.; Wang, S.; Bazan, G. C. A Membrane-Intercalating Conjugated Oligoelectrolyte with High-Efficiency Photodynamic Antimicrobial Activity. *Angew. Chem., Int. Ed.* **2017**, *56*, 5031−5034.

(6) Limwongyut, J.; Nie, C.; Moreland, A. S.; Bazan, G. C. Molecular design of antimicrobial conjugated oligoelectrolytes with enhanced selectivity toward bacterial cells. *Chemical Science* **2020**, *11*, 8138−8144.

(7) Zhou, C.; Chia, G. W. N.; Ho, J. C. S.; Seviour, T.; Sailov, T.; Liedberg, B.; Kjelleberg, S.; Hinks, J.; Bazan, G. C. Informed Molecular Design of Conjugated Oligoelectrolytes To Increase Cell Affinity and Antimicrobial Activity. *Angew. Chem., Int. Ed.* **2018**, *57*, 8069−8072.

(8) Yan, H.; Rengert, Z. D.; Thomas, A. W.; Rehermann, C.; Hinks, J.; Bazan, G. C. Influence of molecular structure on the antimicrobial function of phenylenevinylene conjugated oligoelectrolytes. *Chemical Science* **2016**, *7*, 5714−5722.

(9) Hinks, J.; Wang, Y.; Poh, W. H.; Donose, B. C.; Thomas, A. W.; Wuertz, S.; Loo, S. C. J.; Bazan, G. C.; Kjelleberg, S.; Mu, Y.; Seviour, T. Modeling cell membrane perturbation by molecules designed for transmembrane electron transfer. *Langmuir* **2014**, *30*, 2429−2440.

(10) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263−270.

(11) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design - A review of the state of the art. *Molecular Systems Design and Engineering* **2019**, *4*, 828−849.

(12) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705−8722.

(13) Swann, E.; Sun, B.; Cleland, D. M.; Barnard, A. S. Representing molecular and materials data for unsupervised machine learning. *Mol. Simul.* **2018**, *44*, 905−920.

(14) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminf.* **2020**, *12*, 1−22.

(15) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.

(16) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024.

(17) Danishuddin; Khan, A. U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discovery Today* **2016**, *21*, 1291−1302.

(18) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(19) Capecchi, A.; Probst, D.; Reymond, J. L. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J. Cheminf.* **2020**, *12*, 43.

(20) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513−530.

(21) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021**, *13*, 12.

(22) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(23) Axelrod, S.; Gómez-Bombarelli, R. Molecular machine learning with conformer ensembles. *arXiv (Computer Science)* **2020**, *2012.08452*, (accessed Sept 9, 2021).

(24) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Information Processing Syst.* **2017**, *30*, 4765−4774.

(25) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2020**, *2*, 56−67.

(26) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *J. Med. Chem.* **2020**, *63*, 8761−8777.

(27) Hartono, N.; Thapa, J.; Tiihonen, A.; Oviedo, F.; Batali, C.; Yoo, J.; Liu, Z.; Li, R.; Marrón, D.; Bawendi, M.; Buonassisi, T.; Sun, S. How machine learning can help select capping layers to suppress perovskite degradation. *Nat. Commun.* **2020**, *11*, 4172.

(28) Yan, H.; Catania, C.; Bazan, G. C. Membrane-Intercalating Conjugated Oligoelectrolytes: Impact on Bioelectrochemical Systems. *Adv. Mater.* **2015**, *27*, 2958−2973.

(29) Van Der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Machine Learning Research* **2008**, *9*, 2579−2605.

(30) Stokes, J. M.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688−702.

(31) alvaDesc Molecular Descriptors - Alvascience. https://www.alvascience.com/alvadesc-descriptors/ (accessed Sept 9, 2021).

(32) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; John Wiley & Sons, 2008; Vol. *11*.

(33) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning* **2002**, *46*, 389−422.

(34) Schuur, J. H.; Selzer, P.; Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(35) Devinyak, O.; Havrylyuk, D.; Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graphics Modell.* **2014**, *54*, 194−203.

(36) Saíz-Urra, L.; González, M. P.; Teijeira, M. QSAR studies about cytotoxicity of benzophenazines with dual inhibition toward both topoisomerases I and II: 3D-MoRSE descriptors and statistical considerations about variable selection. *Bioorg. Med. Chem.* **2006**, *14*, 7347−7358.

(37) Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.

(38) Liu, S.; Bao, J.; Lao, X.; Zheng, H. Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides. *Sci. Rep.* **2018**, *8*, 11189.

(39) Khalil, T. E.; El-Dissouky, A.; Al-Wahaib, D.; Abrar, N. M.; El-Sayed, D. S. Synthesis, characterization, antimicrobial activity, 3D-QSAR, DFT, and molecular docking of some ciprofloxacin derivatives and their copper(II) complexes. *Appl. Organomet. Chem.* **2020**, *34*, No. e5998.

(40) Boehm, H. J.; Boehringer, M.; Bur, D.; Gmuender, H.; Huber, W.; Klaus, W.; Kostrewa, D.; Kuehne, H.; Luebbers, T.; Meunier-Keller, N.; Mueller, F. Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **2000**, *43*, 2664−2674.

(41) De Pascale, G.; Wright, G. D. Antibiotic Resistance by Enzyme Inactivation: From Mechanisms to Solutions. *ChemBioChem* **2010**, *11*, 1325−1334.

(42) Guo, Y.; Wang, J.; Niu, G.; Shui, W.; Sun, Y.; Zhou, H.; Zhang, Y.; Yang, C.; Lou, Z.; Rao, Z. A structural view of the antibiotic degradation enzyme NDM-1 from a superbug. *Protein Cell* **2011**, *2*, 384−394.

(43) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(44) Arslan, E.; Findik, B. K.; Aviyente, V. A blind SAMPL6 challenge: insight into the octanol-water partition coefficients of drug-like molecules via a DFT approach. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 463−470.

(45) Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; Hadsell, R. Progressive Neural Networks. *arXiv (Computer Science)* **2016**, *1606.04671* (accessed Sept 9, 2021).

(46) Sosnin, S.; Vashurina, M.; Withnall, M.; Karpov, P.; Fedorov, M.; Tetko, I. V. A Survey of Multi-task Learning Methods in Chemoinformatics. *Mol. Inf.* **2019**, *38*, 1800108.

(47) Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068−2076.

(48) European Committee for Antimicrobial Susceptibility Testing (EUCAST) of the European Society of Clinical Microbiology and Infectious Diseases (ESCMID), Determination of minimum inhibitory concentrations (MICs) of antibacterial agents by broth dilution. *Clin. Microbiol. Infect.* **2003**, *9*, ix−xv.

(49) Camacho, D. M.; Collins, K. M.; Powers, R. K.; Costello, J. C.; Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **2018**, *173*, 1581−1592.

(50) Maziarka, L.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule Attention Transformer. *arXiv (Computer Science)* **2020**, *2002.08264* (accessed Sept 9, 2021).

(51) Liang, Q.; Gongora, A. E.; Ren, Z.; Tiihonen, A.; Liu, Z.; Sun, S.; Deneault, J. R.; Bash, D.; Mekki-Berrada, F.; Khan, S. A.; Hippalgaonkar, K.; Maruyama, B.; Brown, K. A.; Fisher, J.; Buonassisi, T. Benchmarking the Performance of Bayesian Optimization across Multiple Experimental Materials Science Domains. *arXiv (Condensed Materials)* **2021**, *2106.01309* (accessed Sept 9, 2021).

(52) 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*, 148−175.

(53) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.

(54) XGBoost Documentation — xgboost 1.4.0-SNAPSHOT documentation. https://xgboost.readthedocs.io/en/latest/index.html (accessed Sept 9, 2021).

(55) Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785−794.

(56) Chemprop — chemprop 1.2.0 documentation. https://chemprop.readthedocs.io/en/latest/ (accessed Sept 9, 2021).

(57) Reuther, A.; et al. Interactive Supercomputing on 40,000 Cores for Machine Learning and Data Analysis. *2018 IEEE High Performance Extreme Computing Conference, HPEC 2018* 2018, 18290412.

(58) GitHub - fmfn/BayesianOptimization: A Python implementation of global optimization with Gaussian processes. https://github.com/fmfn/BayesianOptimization (accessed Sept 9, 2021).

(59) sklearn.gaussian_process.GaussianProcessRegressor — scikit-learn 0.24.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html (accessed Sept 9, 2021).

(60) RDKit: Open-source cheminformatics. http://www.rdkit.org/ (accessed Sept 9, 2021).

(61) GitHub - aspuru-guzik-group/selfies: Robust representation of semantically constrained graphs, in particular for molecules in chemistry. https://github.com/aspuru-guzik-group/selfies (accessed Sept 9, 2021).

(62) GitHub - slundberg/shap: A game theoretic approach to explain the output of any machine learning model. https://github.com/slundberg/shap (accessed Sept 9, 2021).