

# Machine Learning-Driven Approach to Predict Compound Accumulation

Alex Mortenson, Princy Kaitharan, Shangheng Zhong

## Abstract

Developing new drugs against *P. aeruginosa* is a major component of the fight against antibiotic resistance. However, mechanisms of intrinsic resistance, including outer membrane permeation and efflux, remain poorly understood. Herein we assess the ability of over 300 compounds to accumulate in *P. aeruginosa*. We have developed a set of machine learning models designed to predict accumulation in *Pseudomonas*. After screening through various models, Random Forest Regression is found to be the best model to predict the accumulation value in *P. aeruginosa*. This model yielded results closely aligned with experimental evidence that adding hydrophilic properties to the molecule promotes the accumulation of antibiotics in *P. aeruginosa*.

## 1. Background

The emergence of antibiotic resistance poses a serious threat to human health worldwide. In 2019, the World Health Organization estimated that bacterial antimicrobial resistance was responsible for 1.27 million global deaths and contributed to approximately 5 million more. Within the past five years, almost 40% of antibiotics applied for gut, urinary tract, and blood-based infections lost effectiveness. Furthermore, antibiotic resistance is projected to cost over one trillion dollars in additional healthcare expenses by 2050<sup>12</sup>.

Most clinically relevant antibiotic-resistant infections have been attributed to a select group of species – *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumonia*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and species within the *Enterobacteriales* group – collectively referred to as the “ESKAPE” pathogens. These pathogens are all common in clinical environments, can cause life-threatening infections if untreated, and are highly capable of escaping treatment.<sup>3</sup> Among others, *Pseudomonas aeruginosa* is one of the leading causes of hospital-acquired infections. This bacterium is especially threatening towards immunocompromised patients, demonstrating a high mortality rate. It represents the second most common cause of pneumonia, the third most common cause of UTIs, and is among the 10 most frequently identified pathogens in bloodstream infections. *Pseudomonas* can survive under a wide range of conditions, including within medical implants and other instruments. Critically, *pseudomonas* resistance towards key antibiotics, including “last line of defense” drugs, has steeply increased within the past decade<sup>4</sup>. Therefore, the development of novel compounds to target *Pseudomonas* resistance has become an important objective in antimicrobials discovery.

*Pseudomonas* sports a diverse array of mechanisms that make it resistant towards a wide range of antimicrobial agents. These mechanisms can be summarized into four primary categories: the chemical alteration and/or destruction of drugs, the structural modification of the vital molecular structures targeted by the drugs, genetic adaptation to the drugs, and efflux and

permeability<sup>5</sup>. Efflux and permeability are considered the bacteria's two primary mechanisms of intrinsic resistance. Like other Gram-negative bacteria, *P. aeruginosa* includes an additional outer membrane (OM), a selective barrier of outwardly facing Lipopolysaccharide (LPS) that complicates the entry of polar drugs into the cell<sup>6,7</sup>. Many of these drugs must navigate through porins, protein channels in the OM that further selectivity<sup>8</sup>. Moreover, drugs able to permeate into bacterial cells are often susceptible to the activity multidrug efflux pumps<sup>7</sup>. These large transporter complexes bind a wide range of substrates, including drug molecules, detergents, metabolites, and simple solvents<sup>9</sup>. Using either proton motive force, ATP binding and hydrolysis, or the electrochemical gradient of Na<sup>+</sup> ions as an energy source, efflux pumps extrude foreign small molecules from the bacterial cell back into the extracellular environment<sup>10</sup>. The bacterial outer membrane and multidrug efflux pumps create an additional obstacle for potential antimicrobial agents targeting *Pseudomonas*. Successful drugs must inhibit their targets within the cell and accumulate in the cell at sufficiently high concentrations to be potent.

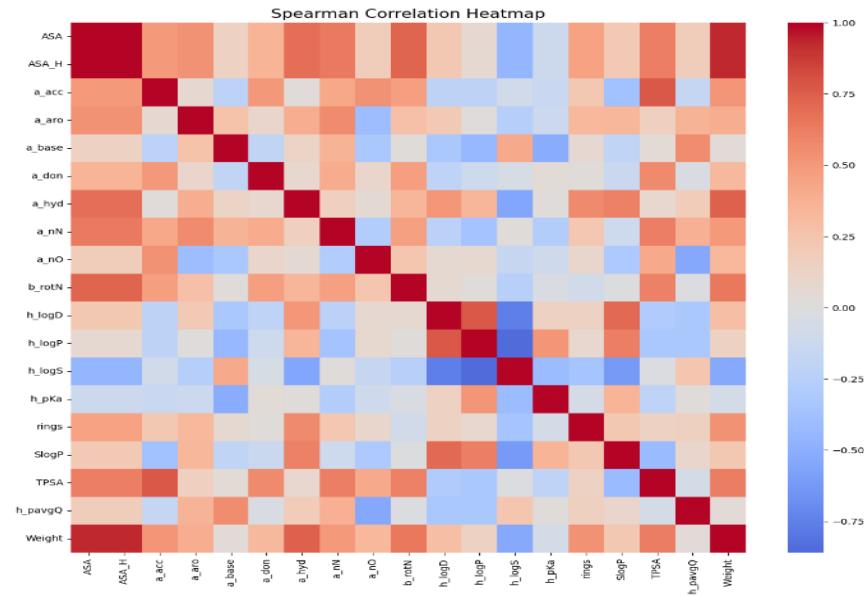
However, the physiochemical properties that promote accumulation within bacteria remain poorly understood. Previous mechanism-based studies have relied predominantly on known antibiotics. Such studies have identified that effective antibiotics tend to maintain lower molecular weights and high polarity<sup>11</sup>. However, effective drugs fail to meet these criteria, suggesting that other properties must also contribute. These retrospective analyses have also been skewed by the over-representation of certain drug classes, such as beta-lactams.<sup>12-15</sup>

In the following study, we leveraged statistical learning tools to build a predictive model for *P. aeruginosa* accumulation. A library of 322 compounds with corresponding LC-MS/MS data was leveraged, to train several models, and from these models we have compiled a set of guidelines for small-molecule accumulation in *Pseudomonas*.

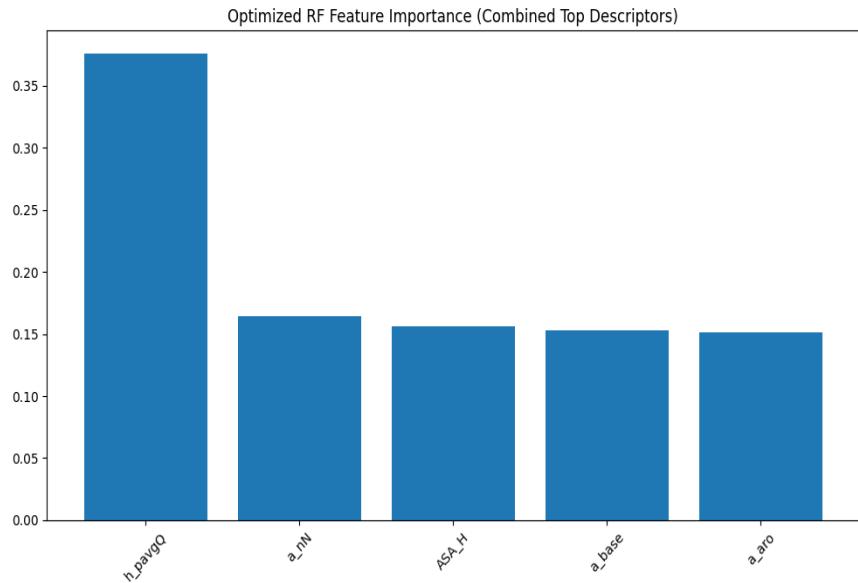
## 2. Methodology

### 2.1 Preprocessing and Feature Selection for Modeling *Pseudomonas aeruginosa* Accumulation

A systematic feature selection procedure was implemented to identify molecular descriptors most strongly associated with *Pseudomonas aeruginosa* accumulation. Initial correlation-based ranking indicated that descriptors such as h\_pavgQ, a\_base, a\_nO, and a\_aro exhibited the strongest relationships with the response variable, whereas ASA and ASA\_H were highly redundant. To mitigate multicollinearity within the models, ASA\_H was retained, and ASA was excluded, yielding an optimized descriptor set comprising nigh variables: h\_pavgQ, a\_base, a\_nO, a\_aro, a\_nN, h\_logP, h\_logS, ASA\_H, and h\_logP. Collectively, these descriptors represent the most informative physicochemical properties after redundancy removal. To further remove additional unimportant features to suppress overfitting, a smaller feature space with the least important feature removed is used to run random forest regression and evaluate its performance. This process is repeated until the performance of the random forest regression stops improving by eliminating features. After optimization, using only the top 5 features (h\_pavgQ, a\_nN, ASA\_H, a\_base, a\_aro) gives the best result for random forest regression.



**Figure 1. Spearman correlation matrix between features**



**Figure 2. Top 5 most important features for Pseudomonas aeruginosa accumulation**

In this project, a multi-feature regression framework was employed to obtain accurate predictions, using a suite of models that included Ridge regression, Kernel Ridge Regression (KRR), Support Vector Regression (SVR), and Random Forest Regression (RFR). These diverse algorithms were applied to the same descriptor set to systematically compare their ability to fit the

data and to identify the model that offered the best predictive performance. By evaluating each model using consistent metrics and validation procedures, the analysis determined which approach most effectively captured the relationships between molecular descriptors and *Pseudomonas aeruginosa* accumulation.

The Random Forest Regressor was selected as one of the models for comparative analysis. Random Forest is particularly effective for modeling nonlinear relationships between molecular descriptors and antibiotic accumulation, while also capturing complex higher-order interactions among features<sup>16</sup>. Its ensemble architecture, which aggregates predictions from multiple decision trees, helps reduce overfitting and improve generalization performance, making it well suited for datasets with diverse chemical properties. Moreover, Random Forest yields interpretable feature importance scores, providing mechanistic insight into the molecular determinants that govern antibiotic accumulation<sup>17</sup>.

To ensure that model evaluation remains unbiased and statistically valid, the workflow begins.

a) Train/Test Split for Unbiased Evaluation

The dataset is partitioned into a training set (80%) and a test set (20%) prior to model development. The training set is used for model fitting and cross-validation, whereas the test set remains completely isolated until the final evaluation stage. This separation ensures that model performance is assessed on unseen data, providing an unbiased estimate of the model's ability to generalize new compounds. Without an appropriate train-test split, models are prone to overfitting the data and yielding misleadingly optimistic accuracy estimates.

b) Scaling Inside Pipelines to Prevent Data Leakage

After splitting the dataset into training and test sets, the input features are standardized using a StandardScaler. The scaler is fit exclusively on the training data, and the resulting parameters are then applied to transform the test data, ensuring consistent scaling while preventing information leakage from the test set into the training process. Because many machine-learning algorithms require input features to lie on a comparable numerical scale, scaling must not use any information from the held-out data. Implementing this procedure within a Pipeline guarantees that the scaler is fit only on the training folds during cross-validation and that the same transformation is subsequently applied to the test data without refitting. This strategy effectively prevents data leakage, whereby information from the test set inadvertently influences model training and produces overly optimistic performance estimates.

c) k-Fold Cross-Validation and Hyperparameter Tuning with GridSearchCV

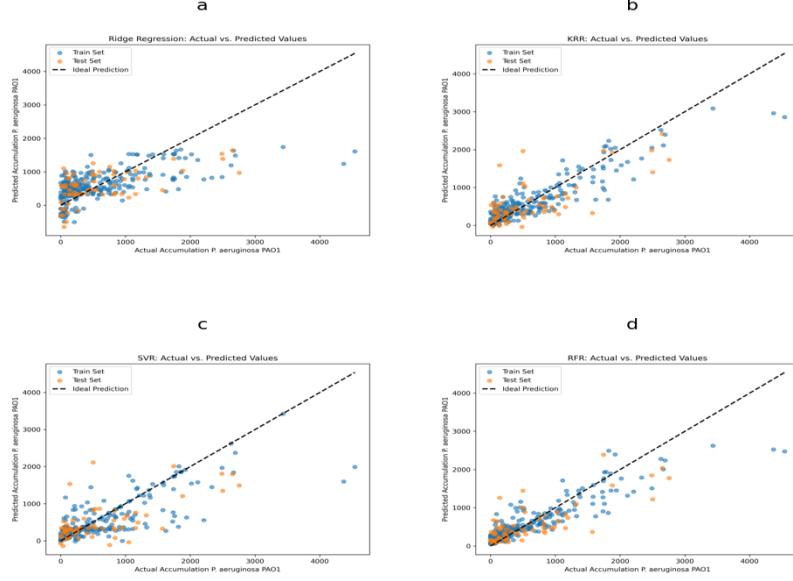
Cross-validation divides the training data into k subsets (folds). For each fold, one subset is used as the validation set while the remaining folds are used for training, and this procedure is repeated k times so that every data point serves in both training and validation. The average performance across all folds provides a robust estimate of generalization, reducing overfitting, stabilizing evaluation metrics, and ensuring efficient use of the available data. In this study, hyperparameter tuning for the RandomForestRegressor is performed using GridSearchCV with 5-fold cross-validation, cv=5. The parameter grid explores key Random Forest hyperparameters, including the number of trees ("n\_estimators": [100, 300, 500, 800]), maximum tree depth ("max\_depth": [None, 10, 20, 30]), minimum samples required for node splitting ("min\_samples\_split": [2, 5, 10]) and leaf nodes "min\_samples\_leaf": [1, 2,

4], and the fraction of features max\_features": ["sqrt", "log2", 0.5] considered at each split. For each hyperparameter combination, GridSearchCV conducts 5-fold cross-validation and selects the configuration that yields the highest validation  $R^2$ . GridSearchCV then trains a RandomForestRegressor for every combination in this grid using 5-fold cross-validation and selects the combination that yields the highest  $R^2$  score. This procedure ensures that the final Random Forest model is tuned within a structured and interpretable hyperparameter space, resulting in a model that is both well-fitted and optimally configured, with improved predictive accuracy and stability. Computation of Standard Regression Metrics ( $R^2$ , RMSE). In this study, a fixed random seed of 42 is used for all stochastic components of the modeling pipeline. Setting random\_state = 42 for operations such as data shuffling, train/test splitting, and Random Forest construction ensures that results are exactly reproducible across runs, model comparisons are fair, and plots, metrics, and selected features remain consistent. This controlled use of a random seed enhances the reproducibility and transparency of the analysis, which is critical in a scientific and pedagogical context.

d) Performance Evaluation

After the final model is trained, performance is evaluated using multiple regression metrics. The coefficient of determination  $R^2$  measures the proportion of variance in the target explained by the model; higher  $R^2$  values indicate greater explanatory power. RMSE (root mean squared error) is sensitive to large errors and quantifies how far predictions deviate, on average, from the observed values, while MAE (mean absolute error) measures the average magnitude of prediction errors without penalizing large outliers as strongly as RMSE. For the tuned Random Forest model, GridSearchCV identified the following optimal hyperparameters: 800 n\_estimators, 2 min\_sample\_split, 1 min\_sample\_leaf, log\_2 max\_feature, and no max\_depth. Under this configuration, the model achieved an  $R^2$  of **0.601** and an RMSE of **423.5** on the test set, indicating moderate explanatory power with prediction errors on the order of approximately 435 units on the response scale. Using  $R^2$ , RMSE, together provides a comprehensive assessment of model accuracy and robustness.

### 3. Results



**Figure 3. Actual vs predicted plot of (a): Ridge Regression; (b): Kernel Ridge Regression; (c): Support Vector Regression; (d): Random Forest Regression**

The parity plot compares predicted versus observed accumulation values for both the training and test sets. For the training data (blue points), observations are tightly clustered around the identity line, indicating that the Random Forest model achieves a strong fit and captures the underlying structure of the training set. For the test data (orange points), predictions also follow the identity line but exhibit greater dispersion, reflecting reasonable generalization with some prediction error and a few notable deviations, likely corresponding to outliers or regions where the model is less reliable.

The close alignment of test points with the identity line suggests low systematic bias, whereas their broader spread relative to the training points indicates moderate variance, consistent with the behavior of ensemble methods such as Random Forest. Although the tighter clustering of training points compared with test points is indicative of mild overfitting, the overall concordance between training and test patterns suggests that the model retains acceptable generalization performance.

Ridge regression was implemented as a linear baseline model to assess how well a purely linear relationship between descriptors and *Pseudomonas aeruginosa* accumulation fits the data. All models were trained on standardized descriptors so that coefficients are directly comparable across features. For Ridge regression, the regularization strength  $\alpha$  was tuned over a logarithmically spaced grid using 5-fold cross-validation ( $k = 5$ ) on the training set. The  $\ell_2$  penalty shrinks all coefficients toward zero without eliminating them, providing stable estimates in the presence of multicollinearity among descriptors. After selecting the optimal  $\alpha$ , the final Ridge model was refit on the full training set and evaluated on the held-out test set using  $R^2$  and RMSE, establishing Ridge as a well-regularized linear benchmark.

For the Ridge regression model, the parity plot of predicted versus observed accumulation values indicates moderate model fit. In the training set (blue), points lie reasonably close to the identity line, with some dispersion, consistent with the effect of L2 regularization in preventing overfitting. In the test set (orange), points are more widely scattered, particularly at the extremes of the response range, indicating reduced accuracy on unseen data and moderate prediction error.

Ridge was tuned over a grid of  $\alpha \in \{0.01, 0.1, 1, 10, 100\}$ , with  $\alpha = 10$  selected as optimal. Under this setting, the model achieved  $R^2 = 0.41$  and RMSE = 539.14 on the training set, and  $R^2 = 0.34$  and RMSE = 543.98 on the test set. This pattern reflects the stabilizing influence of Ridge regularization but also suggests residual bias and underfitting of more complex, nonlinear structure in the data.

In this project, Support Vector Regression (SVR) was implemented as a nonlinear benchmark model for predicting *Pseudomonas aeruginosa* accumulation. All molecular descriptors were first standardized using StandardScaler. An  $\varepsilon$ -SVR with a radial basis function (RBF) kernel was then fitted, and its key hyperparameters, the penalty parameter  $C$ , the kernel width  $\gamma$ , and the  $\varepsilon$ -insensitive margin, were tuned using 5-fold cross-validation on the training set via grid (or randomized) search. The best-performing configuration (highest validation  $R^2$ ) was refit on the full training data and finally evaluated on the held-out test set using  $R^2$ , RMSE.

For the SVR model, the parity plot of predicted versus observed accumulation values indicates a good but imperfect fit. Hyperparameters were tuned using GridSearchCV with 5-fold cross-validation over the grid

$$\text{kernel} \in \{\text{rbf, linear, poly}\}, C \in \{0.1, 1, 10, 100\}, \gamma \in \{"\text{scale}", 10^{-3}, 10^{-2}, 10^{-1}\}, \varepsilon \\ \in \{0.01, 0.1, 0.2, 0.5\},$$

yielding the optimal configuration {'C': 10, 'epsilon': 0.01, 'gamma': 0.1, 'kernel': 'rbf'}. With these settings, SVR achieved an  $R^2$  of 0.6296 and an RMSE of 426.55 on the training set, and an  $R^2$  of 0.4027 and an RMSE of 518.02 on the test set. The parity plot is consistent with these metrics: training predictions cluster closely around the identity line, while test predictions remain aligned with the general trend but show increased dispersion, especially at extreme values. This behavior indicates that the RBF-SVR captures important nonlinear structure in the data and outperforms linear baselines, but exhibits some loss of accuracy and mild overfitting when generalized to unseen compounds.

Kernel Ridge Regression (KRR) was implemented as a nonlinear extension of Ridge regression to model smooth relationships between molecular descriptors and *Pseudomonas aeruginosa* accumulation. All input descriptors were first standardized using StandardScaler. KRR was then fit using a kernel function (radial basis function in this study), so that the response is represented as a weighted combination of kernel evaluations between data points rather than explicit high-dimensional features.

The regularization parameter and kernel hyperparameters (e.g., the RBF width) were tuned using 5-fold cross-validation on the training set via grid or randomized search, and the combination yielding the highest validation  $R^2$  was selected. The final KRR model was refit on the full training data with these optimal hyperparameters and evaluated on the held-out test set

using  $R^2$ , RMSE, and MAE, providing a smooth nonlinear benchmark alongside SVR and Random Forest.

For the Kernel Ridge Regression (KRR) model, the parity plot of predicted versus observed accumulation values indicates a strong overall fit. In the training set (blue), points are tightly clustered around the identity line, suggesting that KRR has successfully captured complex, nonlinear relationships in the data. For the test set (orange), predictions remain reasonably close to the identity line but exhibit greater dispersion than in the training set, consistent with good generalization accompanied by moderate prediction error.

**KRR was tuned using GridSearchCV with the following parameter grid:**

$\text{kernel} \in \{\text{rbf}, \text{laplacian}, \text{poly}\}$ ,

$\alpha \in \{0.01, 0.1, 1, 10\}$ ,

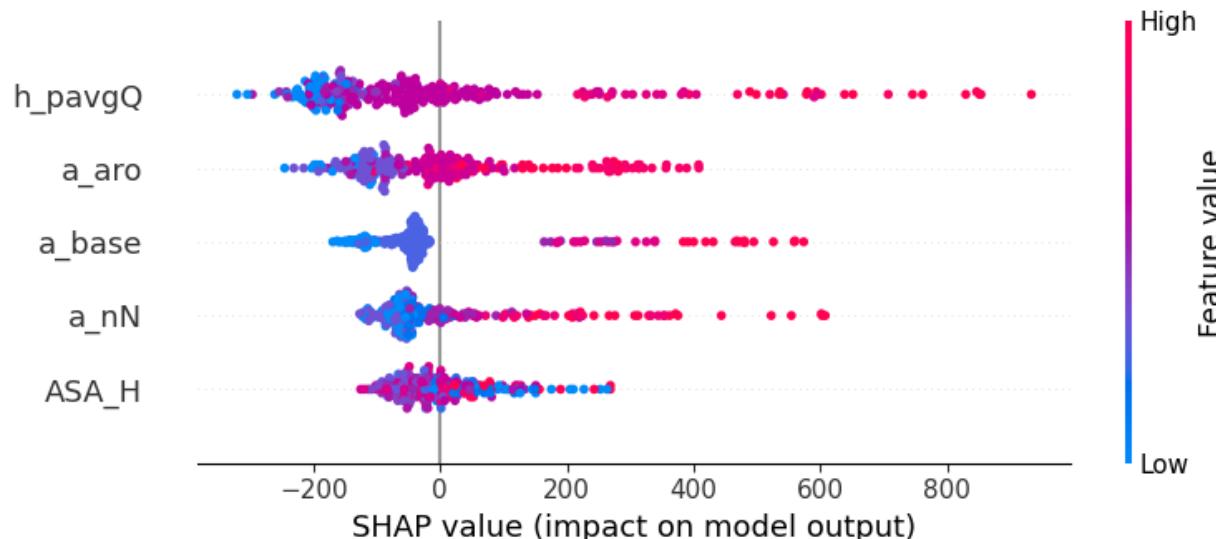
$\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ ,

yielding the optimal configuration  $\{\alpha=0.1, \gamma=0.1, \text{"kernel" = "laplacian"}\}$

With these settings, the model achieved  $R^2 = 0.808$  and RMSE = 307.5 on the training set, and  $R^2 = 0.524$  and RMSE = 462.3 on the test set.

This behavior reflects the combination of ridge regularization with kernel-based nonlinear mapping: the kernel effectively captures underlying structure while the regularization term limits overfitting. Compared with linear Ridge regression, KRR typically exhibits lower bias and higher variance; the observed spread in the test points is consistent with this trade-off and may indicate slight overfitting in regions with sparse data.

To investigate how each feature influences the prediction of the accumulation value, SHAP analysis is performed to indicate whether a feature is positively or negatively influencing the prediction value.



**Figure 4. SHAP analysis summary diagram**

As shown in Figure 4, h\_pavgQ (average total charge), a\_aro (number of aromatic ring), a\_base (number of basic atoms), and a\_nN (number of nitrogen atoms) have a positive influence on the

prediction. In other words, a higher value on those features leads to a higher prediction value. ASA\_H (total hydrophobic surface area) has a negative influence on the prediction value. This result indicates that a molecule with higher hydrophilicity and flatter geometry might be crucial for accumulation in *P. aeruginosa*.

## 4. Discussion

Random Forest Regression (RFR) operates as a nonlinear ensemble model and achieves an excellent fit on the training set, with predictions tightly clustered around the identity line. On the test set, it shows good generalization, with only a slight increase in spread, indicating low bias but moderate variance and a small risk of overfitting. However, its ensemble structure limits interpretability, and feature effects are not easily expressed in simple coefficients, although it can capture complex interactions and nonlinear relationships.

Ridge regression, by contrast, is a linear model with L2 regularization. It provides a good but less precise fit on the training data, with moderate alignment to the identity line and noticeable spread, and its test performance shows wider dispersion and indications of underfitting. This reflects moderate bias and low variance: the model is stable and well regularized but restricted to linear trends in the descriptors. Its main advantage is high interpretability, as coefficients directly quantify the linear contribution of each feature.

Kernel Ridge Regression (KRR) combines L2 regularization with kernel methods to yield a nonlinear regression model. Like RFR, KRR attains an excellent fit on the training set, with points tightly clustered near the identity line. On the test set, it generalizes well and typically outperforms linear Ridge, though with slightly increased spread and a modest risk of overfitting in regions with limited data. KRR exhibits low bias and moderate to high variance due to its flexible kernel mapping. Its interpretability is intermediate: regularization stabilizes the solution, but the kernel representation makes direct interpretation of feature effects less straightforward. Nonetheless, KRR effectively captures nonlinear relationships that are not accessible to purely linear models.

In our prior tests, we identified that a Random Forest Regression (RFR) model gave us the best predictions of accumulation value in *P. aeruginosa*. Notably, through this model, we also identified several features that correspond with improved accumulation. Of these features, “h\_pavgQ” – the average total charge – was the overwhelming most important. This aligns closely with other similarly performed analyses, in which a large positive polar surface area was found to correlate with increased *P. aeruginosa* accumulation<sup>18</sup>. Additionally, other important features identified included “a\_nN”, the number of Nitrogen atoms, and “a\_base”, the number of basic atoms. These features both align closely with prior observations for strong compound accumulation in *E. coli*. Namely, the presence of an ionizable Nitrogen, such as a primary amine, can substantially improve a compound’s accumulation<sup>19</sup>. Additional studies suggest that these charged amines facilitate permeation across OM porins by via formation of favorable electrostatic interactions with the channel<sup>20</sup>. Thus, it is possible this remains true for porin types in *Pseudomonas*<sup>20</sup>. It is also plausible that other important features, such as the total hydrophobic surface area (ASA\_H) and number of aromatic atoms (a\_aro) result from unfavorable interactions between the compounds and *Pseudomonas* efflux pumps, thereby making them less susceptible to efflux. Ultimately, this model serves as a strong starting point for the prediction of *P. aeruginosa*.

accumulation. In the future, this model could be supplemented by additional compounds as well as the inclusion of more three-dimensional features.

## References

- (1) *Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022*, 1st ed.; World Health Organization: Geneva, 2022.
- (2) World Health Organization. *Global Antimicrobial Resistance Surveillance System (GLASS) Report: Early Implementation 2016-2017*; World Health Organization: Geneva, 2017.
- (3) Miller, W. R.; Arias, C. A. ESKAPE Pathogens: Antimicrobial Resistance, Epidemiology, Clinical Impact and Therapeutics. *Nat. Rev. Microbiol.* **2024**, *22* (10), 598–616. <https://doi.org/10.1038/s41579-024-01054-w>.
- (4) Elfadadny, A.; Ragab, R. F.; AlHarbi, M.; Badshah, F.; Ibáñez-Arancibia, E.; Farag, A.; Hendawy, A. O.; De Los Ríos-Escalante, P. R.; Aboubakr, M.; Zakai, S. A.; Nageeb, W. M. Antimicrobial Resistance of *Pseudomonas Aeruginosa*: Navigating Clinical Impacts, Current Resistance Trends, and Innovations in Breaking Therapies. *Front. Microbiol.* **2024**, *15*, 1374466. <https://doi.org/10.3389/fmicb.2024.1374466>.
- (5) Munita, J. M.; Arias, C. A. Mechanisms of Antibiotic Resistance. *Microbiol. Spectr.* **2016**, *4* (2), 4.2.15. <https://doi.org/10.1128/microbiolspec.VMBF-0016-2015>.
- (6) Nikaido, H. Molecular Basis of Bacterial Outer Membrane Permeability Revisited. *Microbiol. Mol. Biol. Rev.* **2003**, *67* (4), 593–656. <https://doi.org/10.1128/MMBR.67.4.593-656.2003>.
- (7) Nikaido, H. Prevention of Drug Access to Bacterial Targets: Permeability Barriers and Active Efflux. *Science* **1994**, *264* (5157), 382–388. <https://doi.org/10.1126/science.8153625>.
- (8) Cowan, S. W.; Schirmer, T.; Rummel, G.; Steiert, M.; Ghosh, R.; Paupert, R. A.; Jansonius, J. N.; Rosenbusch, J. P. Crystal Structures Explain Functional Properties of Two *E. Coli* Porins. *Nature* **1992**, *358* (6389), 727–733. <https://doi.org/10.1038/358727a0>.
- (9) Nikaido, H. Multidrug Efflux Pumps of Gram-Negative Bacteria. *J. Bacteriol.* **1996**, *178* (20), 5853–5859. <https://doi.org/10.1128/jb.178.20.5853-5859.1996>.
- (10) Kobylka, J.; Kuth, M. S.; Müller, R. T.; Geertsma, E. R.; Pos, K. M. AcrB: A Mean, Keen, Drug Efflux Machine. *Ann. N. Y. Acad. Sci.* **2020**, *1459* (1), 38–68. <https://doi.org/10.1111/nyas.14239>.
- (11) O’Shea, R.; Moser, H. E. Physicochemical Properties of Antibacterial Compounds: Implications for Drug Discovery. *J. Med. Chem.* **2008**, *51* (10), 2871–2878. <https://doi.org/10.1021/jm700967e>.
- (12) Nikaido, H.; Rosenberg, E. Y.; Foulds, J. Porin Channels in *Escherichia Coli*: Studies with Beta-Lactams in Intact Cells. *J. Bacteriol.* **1983**, *153* (1), 232–240. <https://doi.org/10.1128/jb.153.1.232-240.1983>.
- (13) Richter, M. F.; Drown, B. S.; Riley, A. P.; Garcia, A.; Shirai, T.; Svec, R. L.; Hergenrother, P. J. Predictive Compound Accumulation Rules Yield a Broad-Spectrum Antibiotic. *Nature* **2017**, *545* (7654), 299–304. <https://doi.org/10.1038/nature22308>.
- (14) Brown, D. G.; May-Dracka, T. L.; Gagnon, M. M.; Tommasi, R. Trends and Exceptions of Physical Properties on Antibacterial Activity for Gram-Positive and Gram-Negative Pathogens. *J. Med. Chem.* **2014**, *57* (23), 10144–10161. <https://doi.org/10.1021/jm501552x>.
- (15) Bazile, S.; Moreau, N.; Bouzard, D.; Essiz, M. Relationships among Antibacterial Activity, Inhibition of DNA Gyrase, and Intracellular Accumulation of 11 Fluoroquinolones.

- Antimicrob. Agents Chemother.* **1992**, *36* (12), 2622–2627.  
<https://doi.org/10.1128/AAC.36.12.2622>.
- (16) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.  
<https://doi.org/10.1023/A:1010933404324>.

**Group Contribution:**

Alex: perform feature calculation, Ridge regression, and provide biological background

Princy: perform random forest regression, feature selection, and correlation analysis

Shangheng: perform Kernel Ridge Regression, Support Vector Regression, SHAP analysis, and feature space optimization.