https://doi.org/10.1038/s42004-024-01161-y

Predicting permeation of compounds across the outer membrane of *P. aeruginosa* using molecular descriptors

Check for updates

Pedro D. Manrique [1], Inga V. Leus [2], César A. López[3], Jitender Mehla [2], Giuliano Malloci [4], Silvia Gervasoni [4], Attilio V. Vargiu[4], Rama K. Kinthada[5], Liam Herndon[3], Nicolas W. Hengartner[3], John K. Walker[5], Valentin V. Rybenkov [2], Paolo Ruggerone [4], Helen I. Zgurskaya [2] & S. Gnanakaran [3]

The ability Gram-negative pathogens have at adapting and protecting themselves against antibiotics has increasingly become a public health threat. Data-driven models identifying molecular properties that correlate with outer membrane (OM) permeation and growth inhibition while avoiding efflux could guide the discovery of novel classes of antibiotics. Here we evaluate 174 molecular descriptors in 1260 antimicrobial compounds and study their correlations with antibacterial activity in Gram-negative *Pseudomonas aeruginosa*. The descriptors are derived from traditional approaches quantifying the compounds' intrinsic physicochemical properties, together with, bacterium-specific from ensemble docking of compounds targeting specific MexB binding pockets, and all-atom molecular dynamics simulations in different subregions of the OM model. Using these descriptors and the measured inhibitory concentrations, we design a statistical protocol to identify predictors of OM permeation/inhibition. We find consistent rules across most of our data highlighting the role of the interaction between the compounds and the OM. An implementation of the rules uncovered in our study is shown, and it demonstrates the accuracy of our approach in a set of previously unseen compounds. Our analysis sheds new light on the key properties drug candidates need to effectively permeate/inhibit *P. aeruginosa*, and opens the gate to similar data-driven studies in other Gram-negative pathogens.

The emerging antibiotic resistance crises are driven by the indiscriminate use of existing antibiotics and the lagging discovery of new antibiotics[1,2]. This has fueled rise of bacterial resistance at unprecedented rates. According to the World Health Organization priority list, all three pathogens classified as critical (its most urgent category) are Gram-negative[3–13]. Yet no new major class of antibiotics has been approved to treat infections caused by this group of organisms since 1962[10,14]. Therefore, there is a critical need to find effective ways to bypass the biological and chemical challenges that hamper the discovery of new and effective antibacterial treatments.

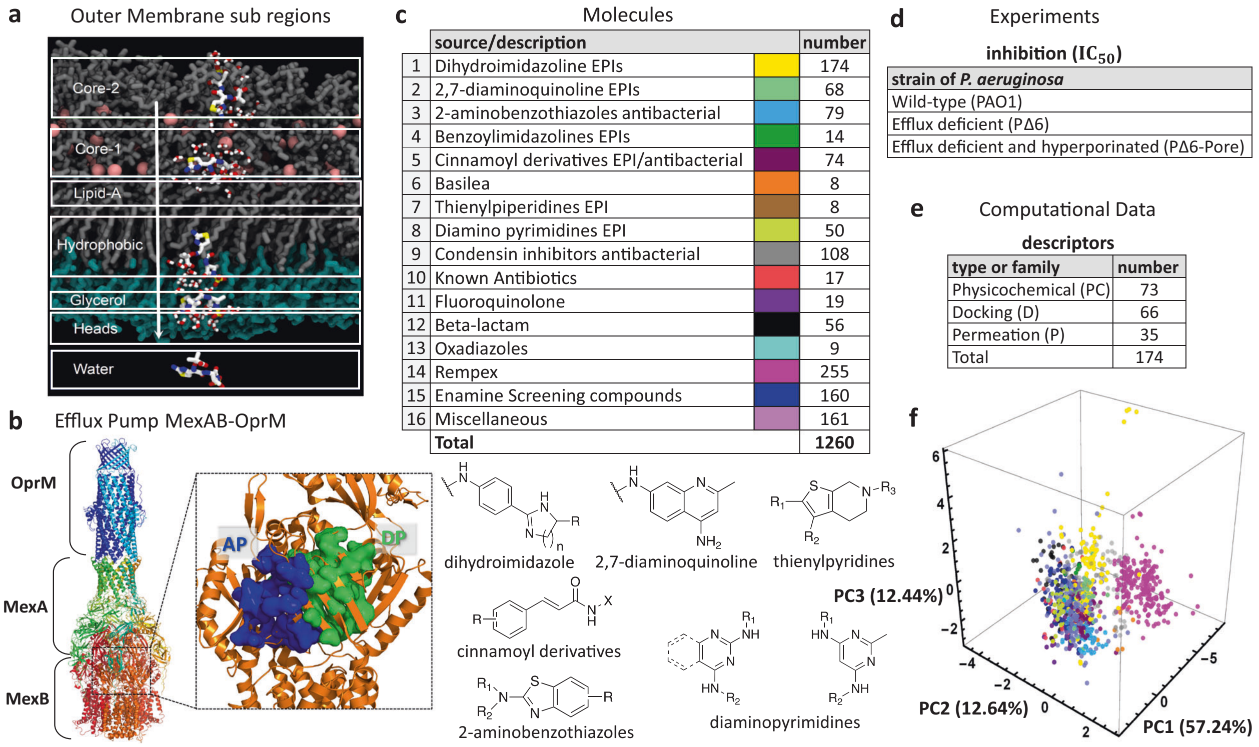
The major determinants of resistance in Gram-negative bacteria are (1) the low permeability barrier of the outer membrane (OM) that hinders diffusion of drug molecules across the membrane, and (2) the action of multidrug efflux pumps that expel drugs and other noxious compounds from the cytoplasm and periplasm back into the extracellular environment[4–8,15]. The synergistic relationship between slow permeation and efflux effectively prevents intracellular accumulation of antibiotics to reach critical concentration levels that inhibit bacteria growth. Mathematical modeling efforts have been able to quantify critical aspects of single-cell in/out flux dynamics[16–18] and their implications at the colony level[19,20]. However, the large complexity and diversity of the interactions occurring between the drugs and the determinants of antibiotic resistance at molecular scale makes it harder to develop predictive models for drug permeation, efflux avoidance and antibacterial activity. Therefore, there is a need to incorporate detailed molecular determinants from computational approaches that probe the bacterium-specific molecular-level interaction profiles.

[1]Physics Department, George Washington University, Washington 20052 DC, USA. [2]Department of Chemistry and Biochemistry, University of Oklahoma, Norman 73019 OK, USA. [3]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos 87545 NM, USA. [4]Department of Physics, University of Cagliari, Monserrato 20052 CA, Italy. [5]Department of Pharmacology and Physiology, Saint Louis University, St. Louis 63103 MO, USA.
e-mail: pmanriq@gmail.com; gnana@lanl.gov

Molecular dynamics (MD) simulation has emerged as a useful tool to provide mechanistic understanding of the structure-function relationship of complex macromolecules and how they behave and interact with particular environments at the molecular scale. In addition, MD is able to offer spatio-temporal information that can fill the gap between experimental resolution and modeling limitations. MD has been successfully applied in the field of drug discovery[21–24], and in particular at providing detailed information of the molecular structures responsible for the low permeability of Gram negative OM[25–29] and on the drug trafficking through efflux pumps[30]. Furthermore, it allows us to explore the complexity of the chemical environment by quantifying the interactions between a wide spectrum of compounds with specific proteins and compartments of a bacterial cell such as the OM and an efflux pump. The output of these MD simulations are often long multivariate time series describing the position of every atom over time, which are generally challenging to analyze. Traditional statistical techniques[31–34], network theory[35,36], and artificial intelligence[37] are among the most implemented and promising quantitative tools to help unravel complex patterns in these large multidimensional datasets.

In *P. aeruginosa*, the low permeability of the OM is mostly attributed to its particular composition. A combination of highly anionic Lipopolysaccharide (LPS) molecules, tightly complexed with divalent cations makes this membrane an almost impenetrable shield[7,25,38]. A single LPS molecule provides distinctive chemical environments across the OM of the Gram-negative bacteria: 1) long carbohydrate-enriched regions massively shield the exposure of the membrane towards the extracellular space; 2) phosphates and ionizable chemical groups are highly repellent to hydrophobic molecules and 3) a sheet of divalent cations provides strong coordination among the LPS of the outer layer in the OM. Thus, extracting the molecular determinants of the process governing the passive diffusion of molecules across this layer would be of tremendous aid in the design of new antibiotics. In order to achieve this, we have carried out massive MD calculations at atomic level, extracting specific properties during the assisted translocation of hundreds of compounds across the OM of *P. aeruginosa*. For each compound we have computed 35 permeability descriptors, which are extracted by instantiating MD trajectories from seven distinctive regions of the OM [Fig. 1a][37]. These regions were selected in order to have a full representation of all the chemical environments which directly affects the diffusion process. The approach is iteratively repeated until the entire set of compounds is fully covered. The permeability descriptors encompass a set of physical parameters which can directly impact the efficiency of molecular translocation: molecular interaction energy with the surrounding environment (Δh), number of hydrogen bonds with surrounding environment (HB), molecular lateral mean squared displacement (Δxy) and molecular entropy (Δs).

The major efflux pump of *P. aeruginosa* that contributes to clinical antibiotic resistance is MexAB-OprM, which extends across the inner and outer membrane aiding the organism to expel toxins from the intracellular and periplasmic region, directly into the extracellular space [Fig. 1b][39–41]. In this complex, MexB is a homotrimeric protein embedded into the inner membrane and belonging to the Resistance Nodulation cell Division (RND) superfamily. It is in charge of recognition, binding, and transport of diverse substrates[41–43], and it constitutes the main barrier that any compound in the intracellular region needs to overcome. Each monomer of the MexB trimer adopts three different conformations enabling access (A), binding (B), and extrusion (C) of substrates[44,45]. We quantify the interactions between each of the studied compounds and MexB via ensemble docking calculations, from which we collected all docking poses (600 per compound), average affinity binding, and identified the contacts made by each compound to every MexB



	source/description		number
1	Dihydroimidazoline EPIs	🟨	174
2	2,7-diaminoquinoline EPIs	🟩	68
3	2-aminobenzothiazoles antibacterial	🟦	79
4	Benzoylimidazoles EPIs	🟩	14
5	Cinnamoyl derivatives EPI/antibacterial	🟪	74
6	Basilea	🟧	8
7	Thienylpiperidines EPI	🟫	8
8	Diamino pyrimidines EPI	🟩	50
9	Condensin inhibitors antibacterial	⬜	108
10	Known Antibiotics	🟥	17
11	Fluoroquinolone	🟪	19
12	Beta-lactam	⬛	56
13	Oxadiazoles	🟦	9
14	Rempex	🟪	255
15	Enamine Screening compounds	🟦	160
16	Miscellaneous	🟪	161
	Total		**1260**

inhibition (IC_{50})

strain of *P. aeruginosa*
Wild-type (PAO1)
Efflux deficient (PΔ6)
Efflux deficient and hyperporinated (PΔ6-Pore)

descriptors

type or family	number
Physicochemical (PC)	73
Docking (D)	66
Permeation (P)	35
Total	174

Fig. 1 | Our library of compounds, experimental data, and molecular descriptors. **a** Computational representation of the outer membrane environment (OM) of *P. aeruginosa* detailing the seven sub-regions where MD simulations where the 35 descriptors listed in E were computed for each molecule. **b** Experimental structure of the tripartite efflux system MexAB-OprM (PDB ID: 6TA6 [10.1038/s41467-020-18770-5]). On the right: focus on the two MexB major binding pockets, AP and DP. **c** Assembled library of 1260 antimicrobial molecules classified into 16 distinct structural chemotypes as listed (top left), and some examples are shown in the bottom panel. **d** Each compound is characterized by its antimicrobial activity in three strains of *P. aeruginosa* by means of the 50% inhibitory concentration (IC_{50}). **e** Molecules in **c** are further characterized by 174 computationally-derived mechanistic descriptors classified as either docking (D), permeation (P), or physicochemical (PC). These are computed using QSAR methods, density functional calculations, ensemble docking and MD simulations in water and in the OM of *P. aeruginosa*. **f** Principal components third degree decomposition of the molecules following the color code shown in **c**.

residue. From the list of contacts, we selected a subset of residues of the access of the Loose monomer (AP) and deep (DP) of the Tight monomer substrate binding pockets based on known crystallographic data for AcrB from E. coli, homologous to MexB from *P. aeruginosa*[44,46]. These residues are generally considered to line/define the two pockets and are relevant for recognition/binding of compounds[45,47]. Some of them, in particular the PHE residues of the hydrophobic trap inside the DP, were found to be key for the interaction of the transporter with inhibitors[48,49]. Our computational analysis of MexB yielded 66 docking descriptors for each compound[37].

In this paper, we analyze the growth inhibitory activities of a unique library of 1260 antimicrobial molecules belonging to several structural classes of compounds including known antibiotics and efflux pump inhibitors [Fig. 1c] (data provided in the Supplementary Data 1). The antibacterial activities are measured in strategically designed strains of *P. aeruginosa* [Fig. 1d] that can isolate the effects of permeation and efflux avoidance[50]. We then use the antibacterial activity data to identify correlations with a large set of computationally-derived mechanistic descriptors described above [Fig. 1e]. These properties are subsequently characterized by means of their ranked correlations along with a hierarchical clustering algorithm to establish similarity relationships (linear and non-linear) among them. The resultant clusters are used as input parameters of a statistical model that, using the experimental 50% inhibition concentration (IC_{50}) data, identifies non-trivial relationships between different sets of descriptors and their ability to predict bacterial permeation. Unlike our previous study[37], which focused on efflux avoidance on a smaller set of molecules (290 Rempex compounds), our current analysis targets permeation/inhibition in a much broader and diverse library of compounds, considers non-linear relationships among the different types of descriptors, and provides explicit parameter ranges associated with permeation/inhibition of the pathogen. Our analysis identifies an optimal subset of nine relevant clusters containing the mechanistic markers yielding prediction accuracy scores of up to 96%. This is in contrast with other studies that, using chemical features, traditional physicochemical descriptors, or mass spectrometric measurements of accumulation, reach performance scores that are below 90%[51–53]. Our results highlight the role of the permeation descriptors quantifying the interactions between the compounds and the OM surface, the LPS lipid-A and oligosaccharide core 2 sub-regions of the OM. These features, combined with intrinsic properties of the compound like the hydrophobic surface area and the Randic index, show high correlations with permeation and growth inhibition information for a specified range of descriptor values. Our findings shed a new light into which specific molecular interactions are responsible for OM penetration and hindering of bacterial growth. Our approach and conclusions can impact the design of a new generation of antimicrobials.

Results

Following the protocol outlined in ref. 37, mechanistic descriptors are computed using variety of approaches for a much broader spectrum of molecules. We use traditional chemical/physical property evaluations, density functional theory calculations, and all-atom MD simulations of compounds in water. We refer to these 73 physicochemical (PC) descriptors that depend entirely on the compounds as QSAR, QM, and MD, respectively. In addition, we calculate an additional set of 101 descriptors that are generated based on the interaction of compounds with the bacterium-specific efflux pump and the OM and we call them mechanistic descriptors. Here, to account for influx, we consider descriptors calculated from the all-atom MD simulations of compounds interacting with the OM model [Fig. 1a], and, to account for efflux, we consider ensemble docking of compounds targeting specific binding pockets of MexB, the major efflux transporter of *P. aeruginosa*. We refer to them as permeation and docking descriptors, respectively. Our experimental data is obtained by analyzing inhibitory activity of an assembled library of 1260 compounds with antibacterial properties in two mutant derivatives of the wild-type *P. aeruginosa* (PAO1): the PΔ6 strain, which lacks six major efflux pumps (ΔMexAB-oprM, ΔMexCD-oprJ, ΔMexXY, ΔMexJKL, ΔMexEF-oprN, and ΔTriABC), and PΔ6-Pore, which is the hyperporinated version of the PΔ6 strain.

Assembly and properties of the compound library for analyses

For this study, we assembled a unique library of 1260 compounds with antibacterial and efflux inhibitory activities from several different sources [Fig. 1c]. The library included the two separate compound series developed by Basilea Pharmaceutica[54] (8 compounds) and Rempex Phamaceuticals[55] (255) which were culled from their respective efflux-pump inhibitor (EPI) projects. There were also 92 known antibiotics belonging to various structural classes, including Fluoroquinolone (19) and Beta-lactam (56), were also acquired to be included in this library. Several compound series in the collection were synthesized at Saint Louis University (SLU) as part of ongoing EPI and antibacterial projects. The largest source of compounds were a series of EPIs designed to inhibit the AcrAB-TolC pump in E. coli including dihydroimidazoline[56] (174), a related series of benzoylimidazolines[57] (14) and a chemical series of 2,7-diaminoquinoline[58] (68). The two sets of antibacterial compounds included a series of 2-aminobenzothiazoles (79) with an unknown antibacterial target and a series of quaternary amine compounds (108) that target a bacterial condensin enzyme[59]. A series of cinnamoyl derivatives[58] (73) which showed both antibacterial and EPI activity were also included in this set.

Several additional series identified in previous screening efforts were also obtained from commercial sources. These included a series of diamino substituted pyrimidines (50)[60], small series of thienylpiperidines and Oxadiazoles (8 compounds each), and a large set of diverse compounds (160) purchased from Enamine. Finally, a set of miscellaneous compounds (161) comprised synthetic intermediates, related analogs that did not belong to one of the aforementioned chemical series and various screening compounds from the NCI collection[61]. We use the 16 chemotype designations to broadly classify the compounds. An alternate classification of the compounds' 2D structures, using a complete Tanimoto similarity analysis[62], further subdivides the chemotypes shown in Fig. 1c yielding a total of 233 subgroups. This high-resolution classification of the chemical structures will become relevant to further analyze permeation predictability of a few key subgroups (see Section "Relationship between permeation predictability and chemical structure ofcompounds" and "Implementation of our statistical model and the permeation rules").

The analyzed 1260 compounds vary in molecular weight (MW) between 156 and 1260 Da, in the total charge between −2 and +5 and have $cLogD_{7.4}$ values between 11 and −11.3. To evaluate the physicochemical space occupied by the library, we carried out the principal component analysis of nine physicochemical properties of the compounds, which included the molecular weight, the number of hydrogen bond donors and acceptors, the total polar surface area (ASA_P), $clogD_{7.4}$, the topological surface area, the fraction of sp3 hybridized carbon atoms (Fsp3), the total charge, and the number of rotatable bonds for the analyzed compounds. The first three principal components (PC) [Fig. 1f] covered 82.3% of the explained variance. All nine properties almost equally contributed to the compound distribution in the PC1 coordinate, whereas the total charge, the number of hydrogen bond acceptors and the number of rotatable bonds were major contributors in the PC2 (see Supplementary Fig. S1 in the Supplementary Methods). Thus, the assembled library is unbiased in respect to one or more features and covers a broad physicochemical space. Additional details are provided in the Methods section and the Supplementary Methods.

Nonlinear relationships among descriptors

The diversity of the chemical space is reflected by the wide range of physicochemical properties of individual compounds, as well as, in their interactions with specific bacterial components such as the OM and the efflux pump. Finding the relevant properties that reliably correlate with a particularly desired behavior or process is challenging: among various descriptors of the compound and molecular descriptors of compound's interactions with bacterial components, some carry redundant information while others are uninformative. Therefore, reducing the number of descriptors is helpful in developing a robust predictive model. We achieve this by clustering the descriptors and grouping them into subsets that have similar co-variation across the 1260 compounds. Each cluster can be interpreted as a collection of nearly equally informative features, from which one can select a representative covariate to be used to predict an outcome. Such a reduction not only helps manage the complexity of predictive models, but also alleviate the experimental and computational efforts required to characterize each compound.

Traditional correlation coefficients (e.g., Pearson coefficient, C_{ij}), often implemented by clustering algorithms, quantify the strength of the linear relationships among random variables. Highly correlated variables are expected to belong to the same cluster, while variables with smaller correlation coefficients are placed on different clusters. It is well known that nonlinear transformations of a given variable, while containing the same information, can have a small correlation coefficient. Thus, clustering variables based on the correlation coefficient can have the undesirable property of separating into different clusters, variables that are non-linear transformations from one another. In our evaluation, we observe nonlinear relationships between features, and find that focusing only on linear relationships (e.g., using the standard correlation coefficient to cluster the variables) leads to poor predictive models. To address this problem, we consider rank correlations (R_{ij}), a generalization of the standard correlation, that captures both linear and (monotone) nonlinear relationships.

Figure 2 depicts key examples of the relationships that are found in the molecular descriptors computed on the studied compounds. Panel a shows the values of the compounds' cumulative entropy calculated in two different sub-regions of the OM, the lipid surface heads (y-axis) and the glycerol region (x-axis). They show a strong linear correlation captured by a Pearson coefficient and rank correlation values close to unity. In contrast, panel b shows that the relationship between the physicochemical descriptors associated to the rotational constant in the z coordinate (y-axis) and the isotropic polarization (x-axis), is monotonically decreasing. This pair of variables is characterized by a Pearson coefficient of −0.4787, which does not quantify the strong non-linear dependency shown. On the other hand, the rank correlation captures better the decreasing monotonic relationship shown by these two physicochemical descriptors with a coefficient value of −0.8851. These key changes in the correlation coefficients have greater effects when computing a hierarchical clustering algorithm of the full set of descriptors leading to the identification of 29 clusters using standard correlations, while the rank correlations identify 37 clusters [Supplementary Fig. S2]. In this case, the latter is able to better capture the wide diversity among the different families of descriptors, which has an ultimate key implication when identifying the optimal combination of descriptors (or clusters) that better correlate with the compounds' desired behavior. Finally, Fig. 2c shows the compounds' molecular weight against their resonantcount, resulting on values close to zero for both measures pointing to uncorrelated variables, which is in agreement with the shown dependency in the plot.

Non-trivial relationships among the different classes of descriptors

A hierarchical clustering characterization of the individual families of descriptors reveals two ways in which the grouping of these quantities occurs: first and most simple, descriptors that quantify properties associated with a single attribute gather together, and second, descriptors that are computed in neighboring locations of a specific molecular environment also tend to cluster together. An example of the former is the clustering of size-related intrinsic physicochemical quantities such as the molecular weight and the number of heavy atoms [Supplementary Figs. S3 and S4]. As for the latter, we find that the number of contacts a given compound makes with a specific residue in MexB (docking descriptor) is correlated to that of another residue, if the residues are close to each other within the same MexB monomer [Supplementary Fig. S5]. Combinations of these two cases are also found. For example, the cumulative entropy associated with a molecule when computed in neighboring sub-regions of the OM (permeation descriptors) are highly correlated among them [Supplementary Fig. S4]. A detailed quantitative analysis using hierarchical clustering algorithm on these individual families of descriptors is presented in the Supplementary Methods. The natural question that arises is how descriptors, belonging to different families, are correlated with each other, and what is the meaning of these relationships within the context of predicting bacteria permeation and growth inhibition.

Figure 3 shows a visual representation of the individual relationships among all 174 descriptors (definitions are provided in the Supplementary Data 2) together with the clusters that are identified by a hierarchical algorithm using the ranked correlation coefficients. The relationship between pairs of descriptors is quantified by means of a dissimilarity matrix (heat map), which is defined as the square-root of the unity minus the square of the (ranked) correlation coefficient associated with the pair, and ordered according to the clustering algorithm (see dendrogram). The colored regions in the dendrogram define the clusters, which are determined by the L-method[63] using the percentage of the variance explained as the critical

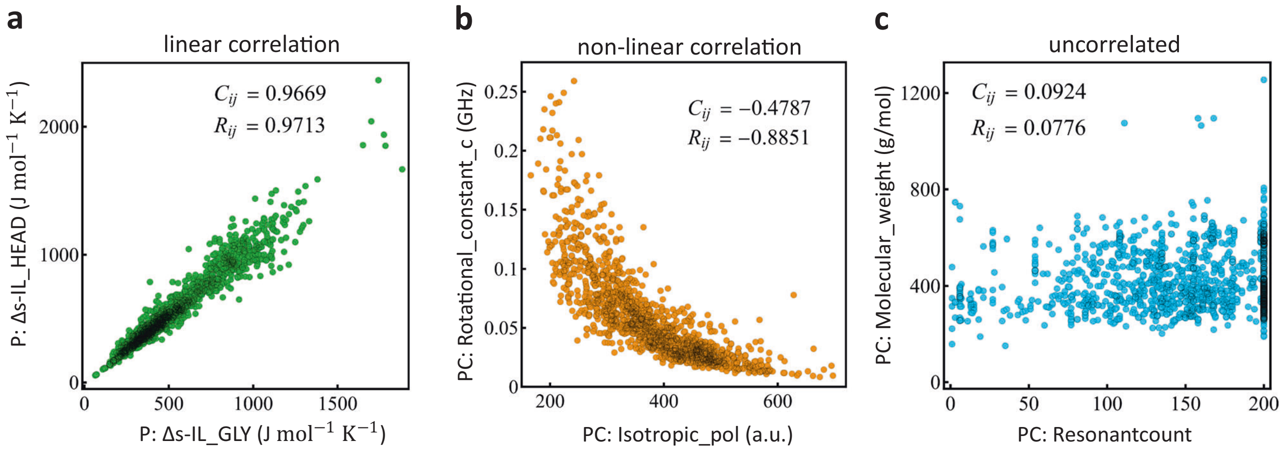


Fig. 2 | Basic relationships among the descriptors. Each data point represents a molecule, and it is projected on the two-dimensional space of two descriptors as shown. Some of the most common properties found among the descriptors are: **a** linear correlations, **b** non-linear correlations, and **c** uncorrelated. The numbers on each of the panels are computed by standard correlations (i.e., Pearson coefficient), C_{ij}, and rank correlations, R_{ij}. As shown, rank correlations capture better the non-linear relationship shown in the central panel.

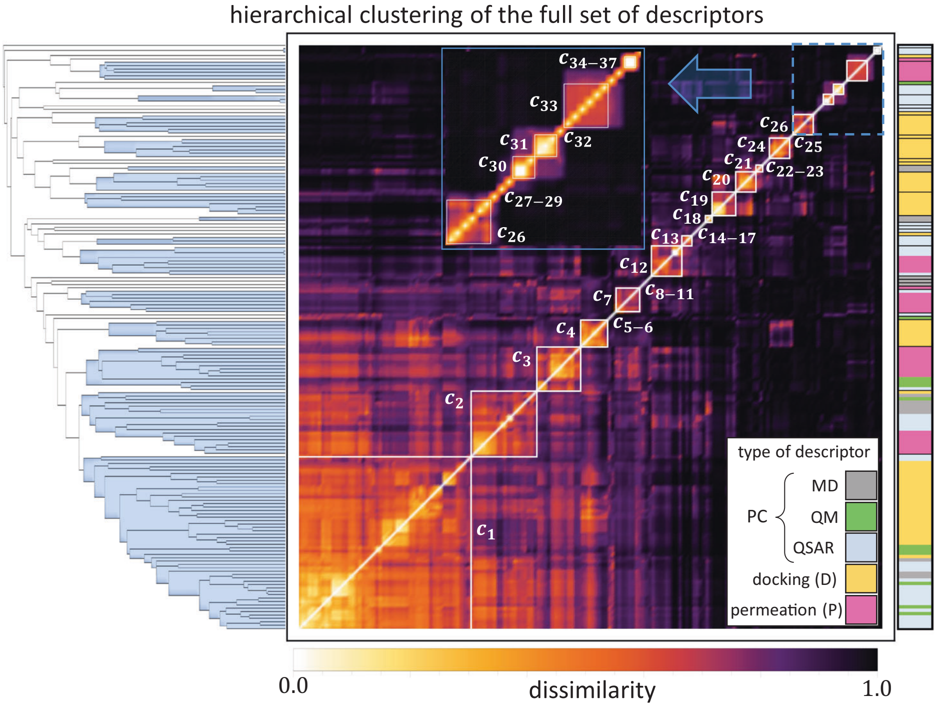


Fig. 3 | Data characterization of the 174 molecular descriptors by means of a hierarchical clustering algorithm using their associated rank correlations. The computation yields 37 dissimilar clusters of sizes ranging from single descriptor clusters (e.g., cluster 37) up to a large cluster of 52 descriptors (cluster 1). The dendrogram in the left-hand side depicts the individual as well as cluster level relationships among the descriptors (single line) and clusters (blue groups), respectively. It also permits the visualization of the cut defining the number of clusters, which was determined by the L-method (see Supplementary Methods). The heat map further highlights the different clusters as well as the relationships between themselves and between individual descriptors via a dissimilarity computation of their associated rank correlations. The type of descriptor is defined in the right-hand side by the color code shown in the legend.

parameter (see Supplementary Methods). The procedure identifies 37 clusters in total (blue groups in the dendrogram and white squares in the dissimilarity matrix), 32 of which are comprised by descriptors of a single type, while only 5 clusters are comprised by two or more types. The categorization of the descriptors is illustrated by the following color code: permeation descriptors are magenta, docking descriptors are orange, and for the physicochemical descriptors we further separate them into QSAR in light blue, QM in green, and MD in water in gray.

Among the single-type clusters found in the full set, there are similar grouping patterns than when performing hierarchical clustering on a single family of descriptors only [Supplementary Figs. S3–S5], which is expected given the wide spectrum of properties analyzed. However, we also find interesting differences. For example, a single-descriptor cluster in the single-family analysis (e.g., number of donors) becomes part of a larger cluster with descriptors belonging to a different type (e.g., number of hydrogen bonds in OM sub-regions). This provides helpful information to identify redundancy in the information carried by the data, which is desired in order to improve the performance of prediction models. Another interesting finding is that some large clusters formed when grouping single-family descriptors (e.g., permeation only), break when the full set of descriptors is considered. This is also helpful in order to identify outliers with a predictability power difficult to identify when they belong to a larger cluster. This is the case of cluster c_8, which contains one permeation descriptor (HB-MEM-INTER) that is later found to have a great predictability potential. This descriptor, when the clustering is carried out within the single family of permeation descriptors, is part of a larger cluster of other hydrogen bonds-related properties, that together hold a weaker predictability potential. Clusters c_{33} and c_{34}, both of them quantifying the lateral diffusion in sub-regions of the OM, is another example of this type of advantage of using the full set of molecular properties. They comprise a single cluster in the single-family analysis. According to our prediction analysis (presented in the next section), c_{34} has a higher potential of becoming a predictor than c_{33}, and hence, the resultant separation of these clusters in the full set analysis is critical.

When analyzing the full set of descriptors, the largest cluster (c_1) comprises 52 molecular quantities of all types except for permeation. The features grouped in this cluster are mostly related to intrinsic physicochemical properties of the molecules such as size (e.g., volume), graph topology (e.g., Szeged index), polarization (e.g., refractivity), and energy (e.g., thermal energy) of the molecules, together with docking information quantifying the binding energy at both of the studied binding sites of MexB (AP and DP). Interestingly, some additional docking descriptors quantifying the number of contacts between the molecule and residues in DP also comprise c_1. This is explained by a found statistical proximity between these docking descriptors in the DP with the binding energy in both AP and DP [Supplementary Fig. S5]. Indeed, the VINA scoring function[64] is additive and tends to favor larger compounds. Consistently, we found in cluster 1 both descriptors capturing molecular dimensions (e.g. Molecular weight, Atom Count, Heavy atoms, volume) and all the average docking scores (Aff APA 20%, Aff APA 30%, Aff DPB 20%).

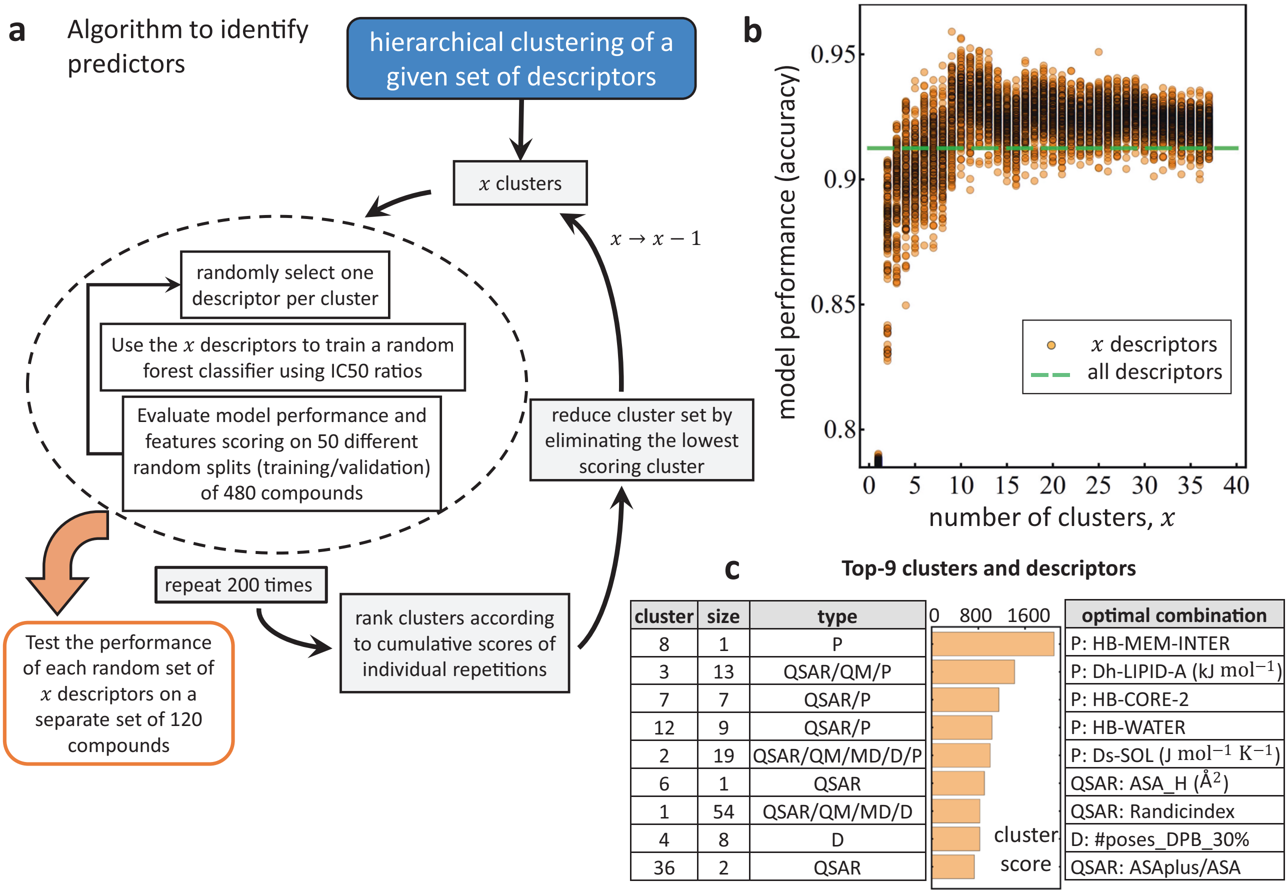
Among the clusters with more than one type of descriptors, c_2 is the only one that gathers properties from all types. These include the highly correlated entropy values found in the different sub-regions of the OM, descriptors quantifying flexibility (rotatable bonds), topology (chain and aliphatic atoms/bonds, rotational constant), and dynamical properties (MD fluctuations and minimal projection area) of the molecule. These properties are found to be correlated also with the number of contacts the molecule makes with residue Thr130 (Threonine) in the deep pocket of MexB. The knowledge of these non-trivial correlations is helpful when determining the predictability power of the cluster, and opens the gate to examine the extent of these relationships in larger families of compounds and the implication when analyzing particular interactions.

Identification of descriptors that predict permeation

The nonlinear relationships found among descriptors, together with the inability of traditional principal component analysis methods to distinguish between weak and strong permeators, lead us to design a framework that, accounting for these nonlinearities, determines the likelihood that a compound can be classified as be a good or a bad permeator. We aim at identifying a minimal set of molecular descriptors that better correlate with the ability of the compound to permeate/inhibit the pathogen. As many of these properties have strong linear and non-linear correlations among each other, our hierarchical clustering analysis that implements the ranked correlations serves to identify similarities in the descriptor space and hence becomes a good starting point to search for the minimal set of predictors. While descriptors that belong to different clusters are weakly correlated with each other, the information that they carry about the molecule is not redundant and it could point (from different angles) to its ability to permeate and inhibit the bacteria's growth. Similarly, descriptors within the same cluster carry correlated information about the molecule and not all of these values may be needed.

The target class (i.e., the measure of classification) in these calculations is determined by ratios of the IC_{50} values associated with each compound and extracted from inhibitory activity in two mutant derivatives of *P. aeruginosa* PAO1 (see Fig. 1d). For permeation we use the IC_{50} ratio PΔ6-Pore/PΔ6, which highlights the role of the OM. If the ratio tends to 1, the concentration of drug needed to inhibit 50% of the bacterial growth for both derivatives is very similar. This means that the OM barrier makes little to no difference in the action of such drug. Hence, a molecule with such ratio is classified as a strong permeator. Conversely, if the ratio tends to 0, the concentration of drug needed to inhibit 50% of the growth of the PΔ6 derivative is much greater than that needed to inhibit the hyperporinated PΔ6-Pore derivative. Hence, compounds with such ratios are classified as weak permeators. Our calculations show that using a threshold ratio of 0.5 to distinguish between the permeation classes optimizes the classification when compared to other threshold choices (see Supplementary Fig. S11 in the Supplementary Note 3), turning this analysis into a binary classification problem. In short, the target classes are defined as: strong permeators (i.e., class 1) having an IC_{50} ratio greater or equal to 0.5, and weak permeators (i.e., class 0) with an IC_{50} ratio smaller than 0.5. The total number of molecules with measurable inhibitory activity in these mutant pathogens is 600.

Figure 4a describes the algorithm designed to reduce the number of descriptors in order to identify an optimal set that are best associated with the molecule's ability to permeate the bacterial OM. Full details of the algorithm are provided in the Methods section. Here we provide a simplified description. The criteria for reduction is based on cluster performance. We randomly select *x* descriptors (one per cluster) to train and validate a random forest classifier[65] using 480 compounds and their respective IC_{50} ratios. For each random selection, we carry out 50 training/validation calculations (95:5 data proportion) where the data is scrambled at each iteration. The trained models are tested on the remaining 120 compounds, where we compute the model performance score and the Shannon entropy associated to each descriptor. The latter indicates by how much the descriptor reduces the uncertainty in the classification[66] and hence it becomes a measure of its importance. The average performance over the 50 iterations is shown in Fig. 4b (single orange circle) and the process is repeated for 200 random selections of *x* descriptors. By adding the importance score for each



cluster	size	type	optimal combination
8	1	P	P: HB-MEM-INTER
3	13	QSAR/QM/P	P: Dh-LIPID-A (kJ mol^{-1})
7	7	QSAR/P	P: HB-CORE-2
12	9	QSAR/P	P: HB-WATER
2	19	QSAR/QM/MD/D/P	P: Ds-SOL (J mol^{-1} K^{-1})
6	1	QSAR	QSAR: ASA_H ($Å^2$)
1	54	QSAR/QM/MD/D	QSAR: Randicindex
4	8	D	D: #poses_DPB_30%
36	2	QSAR	QSAR: ASAplus/ASA

Fig. 4 | Our data-driven model of predictors identification. a Hierarchical clustering algorithm is used to select different combinations of *x* descriptors. A random forest classifier is trained on the *x* descriptors alongside with IC_{50} ratios, and the descriptors performance are scored accordingly. Over the course of several random selections of *x* descriptors, the aggregated *x* scores are used to rank the clusters according to predictability. The lowest ranked cluster is eliminated and the value of *x* is reduced. In parallel, for each classification run, the fitted model is tested in a separate set of compounds and the evaluation metrics are stored. **b** Model performance accuracy for each cycle of the model. Individual circles represent the average accuracy score of a single random combination of *x* descriptors using a random forest classifier over 50 random training/validation splits. The dashed green line represents the average accuracy score for a random forest classifier using the full set of 174 descriptors. **c** Top-9 clusters ranked according to their testing performance. The table in the left panel distinguishes the cluster number, its size (number of descriptors comprising the cluster), and type of descriptors they contain. The central panel is the aggregated cluster score where all values add to 10^4, which is the total number of runs for a particular value of *x*. The right panel lists the top-9 optimal descriptors that produce a testing accuracy of 96.2%.

descriptor at each iteration, we construct the respective cluster importance score. The cluster with the lowest score is eliminated and the process is repeated using x-1 clusters until only one cluster is present. An example of the testing portion of the algorithm is illustrated in Fig. 4b for the evaluation metric of prediction accuracy (see Supplementary Fig. S6 in the Supplementary Note 1 for additional evaluation metrics). For a given value of x, it is shown how the random combination of descriptors coming from different clusters perform (orange circles), and how this metric is affected by reducing the number of clusters. It is also shown how the model performance compares with that of a baseline model consisting of simply running the classification algorithm in the full set of 174 descriptors. As illustrated, we find that many combinations of descriptors outperform the baseline model for values of x greater than 3. The maximum in the prediction accuracy is found for $x = 9$ clusters for some combinations of descriptors as noted in Fig. 4c, where the optimal combination of descriptors found is listed (additional details of this calculation are given in Supplementary Fig. S7 in the Supplementary Note 1). The full ranking of clusters is shown in Supplementary Table S1 in the Supplementary Note 1, where we also show additional details of the model performance for alternative selection of descriptors of these nine clusters [Supplementary Figs. S8 and S9]. Indeed, if we start with a different arrangement of compounds in the large and the small groups, we find changes in the combination of descriptors that maximize the accuracy. Interestingly, the top-9 clusters remain unchanged. The relevance of these nine clusters is preserved even when the large and small groups are randomly scrambled after each iteration (see Supplementary Table S2 in the Supplementary Note 1). Hence, the information provided by these clusters is crucial at determining the ability of a molecule to permeate and inhibit the pathogen.

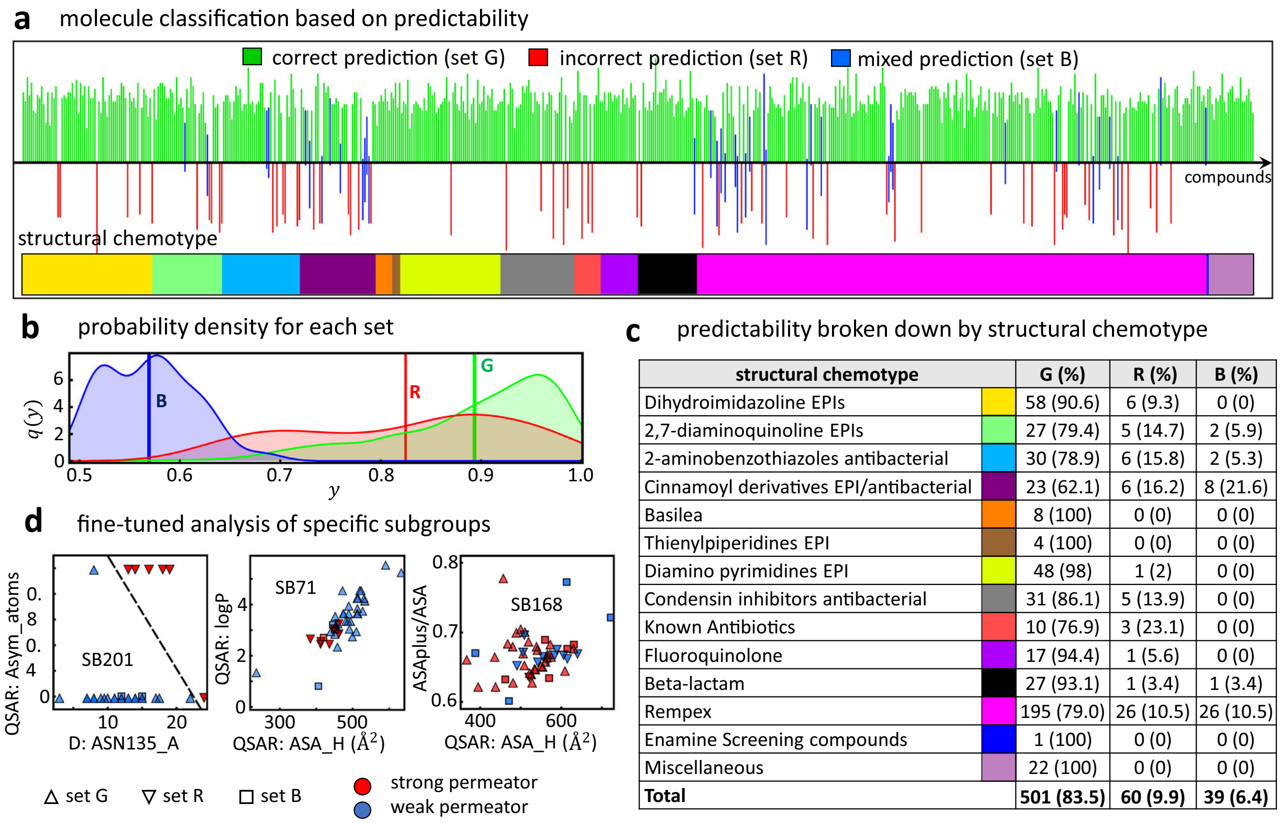
Figure 4c also shows the ranking of these 9 clusters by their cumulative scores (the sum of all values is 10^4), and the optimal combination of descriptors that yields a maximum prediction accuracy evaluated in the testing set equal to 96.2%. As shown, within the top predictors, the permeation descriptors associated with the interaction between the compound and the OM of *P. aeruginosa* in the external environment, in the lipid A, and the LPS core 2 sub-regions of the OM, along with the cumulative entropy and number of hydrogen bonds in the water-membrane interface at the inner leaflet of the OM, score the highest. Also, the physicochemical descriptors quantifying the hydrophobic surface area, the ratio between the solvent accessible surface area of all atoms with positive partial charge and the total water accessible surface area (ASAplus/ASA), and the number of docking poses in the DP of MexB complete the list of nine predictors. As mentioned above, the optimal combination of descriptors tends to change with the testing sample [Supplementary Table S3]. However, we note that this particular combination performs, on average, within 2.3% of the maximum score found for different random testing samples (see Supplementary Figs. S9 and S10 in the Supplementary Note 2). This is encouraging since this particular combination can be generalized across different random testing samples with a modest cost in the performance. Hence, the identification of these markers opens the gate for a deeper study of these key properties, which could guide the design of novel antimicrobials.

Relationship between permeation predictability and chemical structure of compounds

We analyze the robustness of our statistical model and classify the different active compounds in terms of their predictability. We carry out 5000 additional calculations on 100 random testing samples of 120 compounds (50 computations per sample where the training/validation group is scrambled at each computation) employing the classification algorithm over the identified molecular predictors (Fig. 4c). The aggregated results of the prediction for each batch of compounds were analyzed and used to classify the molecules in one out of three sets, as depicted in Fig. 5a. Compounds that were predicted correctly every time, are colored green (set G) and the bar is above the x-axis. This is regardless of the compound being a strong or weak permeator, here we are only examining the truthfulness of the prediction. Compounds that are always predicted incorrectly, are colored red (set R) and the bar is below the x-axis. Finally, compounds that for some simulations runs were predicted correctly and for some other were predicted incorrectly, are colored blue (set B) and portions of the bar are both above and below the x-axis. Interestingly, the largest fraction of the compounds (83.5% or 501 compounds) is predicted correctly every time pointing to a consistent relationship between the predictors' values and the permeation classes. This regularity is encouraging and talks about the existence of clear trends in the computational data that are strongly linked to the experiments and their consistency across a wide diversity of compounds. We devote the next section to unfold these trends for the most predictive descriptors. On the other hand, sets R and B, though much smaller in size, point to the limitations that a one-rule-fits-all approach have. We find that 9.9% of the compounds (60) were predicted incorrectly every time, pointing to a strong but incorrect signal from the computational data, while the remaining 39 compounds (6.4%) are found to give a mix prediction, which hints to a noisy and hence weak signal. Here we analyze these different sets via our model detailed output and examining the chemical structure of the compounds. Detailed metrics are listed in Supplementary Table S4.

To understand the nature of these results, we first analyze the probability densities associated with the identified sets of compounds. We define y as the dominant classification probability associated with each compound using the contributions of every estimator for all model realizations. For example, if for a given compound, out of the N_e estimators, n_0 of them choose class 0 while the remaining $n_1 = N_e - n_0$ class 1, we can define the probabilities for each class as $p_0 = n_0/N_e$ and $p_1 = n_1/N_e$. The dominant probability y of the compound is therefore defined as $y = \max\{p_0, p_1\}$. The probability density associated with each predictability set, $q(y)$, is illustrated in Fig. 5b. Compounds of the set B are characterized by having weak probability values with an average of $\bar{y} = 0.56$. This result is very close to the maximum uncertainty limit of 0.5 (i.e., a coin-flip classification) making the prediction highly unreliable, which is consistent with the mix signal shown in Fig. 5a. Compounds of the set R have greater values of y, but with a wider distribution and an average of 0.82. Finally, compounds of the set G hold a consistently higher average probability of $\bar{y} = 0.89$ with a median above 0.91. Therefore, having consistently high probability values help rule out compounds from the set B and most of those of the set R.

To explore further these differences, we look at the compounds' chemical structure and break down the different sets according to the 16 distinct structural chemotypes defined in Fig. 1c. As shown in Fig. 5c, five chemotypes have members in all three predictability sets, five chemotypes have members in two sets, and four chemotypes have members in only the set G. Hence, at this level of analysis, there is not a clear relationship between the structural chemotypes and the predictability sets. A sharper picture can be drawn when we examine the subdivisions of the distict chemotypes by means of a complete Tanimoto similarity analysis. As mentioned in Section "Assembly and properties of the compound library for analyses", this analysis finds a total of 233 subgroups. Interestingly, nearly 90% of the compounds in sets R and B are concentrated in just 10 Tanimoto subgroups, namely SB71, SB112, SB117, SB118, SB167-SB170, SB201, and SB223. Each of these subgroups is characterized by unique structural features as listed in Supplementary Table S2 in the supplementary section. We examine each of these subgroups individually using our model described in Fig. 4a but adjusting for the number compounds of each subgroup. In four of these subgroups (SB112, SB170, SB201 and SB223) there is a clear separation between the permeation classes using alternative descriptors to those identified for the full set of active compounds. An example of this finding is illustrated in the left panel of Fig. 5d for the subgroup SB201, which belongs to the structural chemotype 3 (i.e., 2-aminobenothiazoles). A combination of the docking descriptor quantifying the number of contacts between the molecule in question and the residue ASN135 in the access monomer in MexB, and the asymmetric atoms allow for a good separation of the permeation classes determined by our experiments in *P. aeruginosa*. This pair of descriptors did not show a wide relevance for the full set of active compounds, but they are found to be key for this specific subgroup of molecules. Our fine-tune analysis of this subgroup correctly classifies the eight



structural chemotype		G (%)	R (%)	B (%)
Dihydroimidazoline EPIs	🟨	58 (90.6)	6 (9.3)	0 (0)
2,7-diaminoquinoline EPIs	🟩	27 (79.4)	5 (14.7)	2 (5.9)
2-aminobenzothiazoles antibacterial	🟦	30 (78.9)	6 (15.8)	2 (5.3)
Cinnamoyl derivatives EPI/antibacterial	🟪	23 (62.1)	6 (16.2)	8 (21.6)
Basilea	🟧	8 (100)	0 (0)	0 (0)
Thienylpiperidines EPI	🟫	4 (100)	0 (0)	0 (0)
Diamino pyrimidines EPI	🟨	48 (98)	1 (2)	0 (0)
Condensin inhibitors antibacterial	⬜	31 (86.1)	5 (13.9)	0 (0)
Known Antibiotics	🟥	10 (76.9)	3 (23.1)	0 (0)
Fluoroquinolone	🟪	17 (94.4)	1 (5.6)	0 (0)
Beta-lactam	⬛	27 (93.1)	1 (3.4)	1 (3.4)
Rempex	🟪	195 (79.0)	26 (10.5)	26 (10.5)
Enamine Screening compounds	🟦	1 (100)	0 (0)	0 (0)
Miscellaneous	🟪	22 (100)	0 (0)	0 (0)
Total		**501 (83.5)**	**60 (9.9)**	**39 (6.4)**

Fig. 5 | Model prediction analysis. a Classification of compounds according to their predictability by our model. 100 random samples of 120 compounds each were tested on the remaining of the data. Compounds that were correctly predicted at each model realization are represented by a green bar pointing above the x-axis (set G). Compounds that were incorrectly predicted in every run are represented by a red bar pointing below the x-axis (set R). Compounds that in some runs were correctly predicted and in some other, incorrectly predicted, are represented by blue bars pointing both ways (set B). The color bar in the bottom indicates the structural chemotype a given compound belongs to as defined in Fig. 1. **b** Probability density $q(y)$ as a function of the probability value y associated with each category of descriptors (G, R, and B) for the dominant target class, i.e., $y = \max\{p_0, p_1\}$, where p_0 and p_1 are the probabilities of being a weak or a strong permeator, respectively. Vertical lines indicate the average probability $\bar{y}$ for each case. **c** Number of compounds and percentage of each set (G, R, and B as defined in **a**) for each structural chemotype following color scheme and ordering as **a**. **d** Analysis of three selected subgroups according to a complete Tanimoto similarity analysis that contain a relevant amount of compounds from the sets R (inverted triangles) and B (squares). Each panel shows the specific subgroups (SB201, SB71, and SB168) in the space of two descriptors identified by our model (Fig. 4a and compared to their respective experimental class: strong permeator (red) and weak permeator (blue). Dashed line in the left panel is produced by a support vector machine classification algorithm.

compounds of sets R and B that belong to chemotype 3. Similar results are found for subgroups SB112, SB170, and SB223 with different descriptors, as shown in the Supplementary Note 3 [Supplementary Fig. S15]. For subgroups SB71 and SB118, the alternative descriptors identified by the model separate better the permeation classes than those of the full set, but some overlap between the classes remains. This is shown in the central panel of Fig. 5d for the subgroup SB71 (see SB118 in Supplementary Fig. S15 in the supplementary Note 3). For the remaining four subgroups (SB117, SB167, SB168, and SB169), which belong to chemotype 14 (i.e., Rempex), a greater overlap between the classes persists pointing to complex nonlinearities between the permeation classes and their descriptors' values and hence the limitations of using descriptors to find clear trends able to distinguish the molecules according to their permeation class.

Focusing on the set G, which comprises the large majority of active compounds (501 molecules), we find clear trends in the values of the key descriptors that define parameter regions mostly associated to one of the two permation classes (Supplementary Figs. S12 and S13 in the Supplementary Note 3). For example, strong permeators are found at strong HB interaction values at the surface of the OM (HB-MEM-INTER) together with a strong negative enthalpy in the Lipid-A sub region of the OM. On the other hand, weak permeators fall into the opposite category with weak OM interaction and weak enthalpy [Supplementary Fig. S13]. In the next section we explore further these general trends and their mechanistic implications in OM permeation.

Descriptor values associated with strong and weak OM permeation

The consistency found in the predictability of the permeation class of the compounds in the set G, makes them a good batch to extract helpful rules that associate specific ranges of descriptor values with a particular target class, i.e., strong or weak OM permeator. To determine the approximate class boundaries and ranges for each individual descriptor, we train a traditional support vector machine (SVM) algorithm[67] for each descriptor of the top-9 clusters (112 descriptors) for the set G of active compounds (501 compounds). Density distribution of descriptor values across the range of selected individual descriptors associated with each permeation class given by their IC_{50} ratio, i.e., strong (red) or weak (blue) permeator, are depicted in Fig. 6. The vertical gray line indicates the binary class boundary identified by SVM. Strong permeators category highly correlates with parameters indicating stronger interactions with different OM subregions, as exemplified in Fig. 6a. They are able to stabilize a larger number of hydrogen bonds (HB) with different parts of the membrane (e.g. Core-2), while retaining a close water shell during the translocation process in the most hydrophobic regions (e.g. aliphatic tails). Consequently, this contributes with a more favorable enthalpy of interaction in specific parts of the membrane [Fig. 6b]. Interestingly, permeation highly correlates with the presence of more rotatable bonds as well as higher entropy, indicative of a more flexible molecular scaffold able to accommodate to the different spatial restrictions

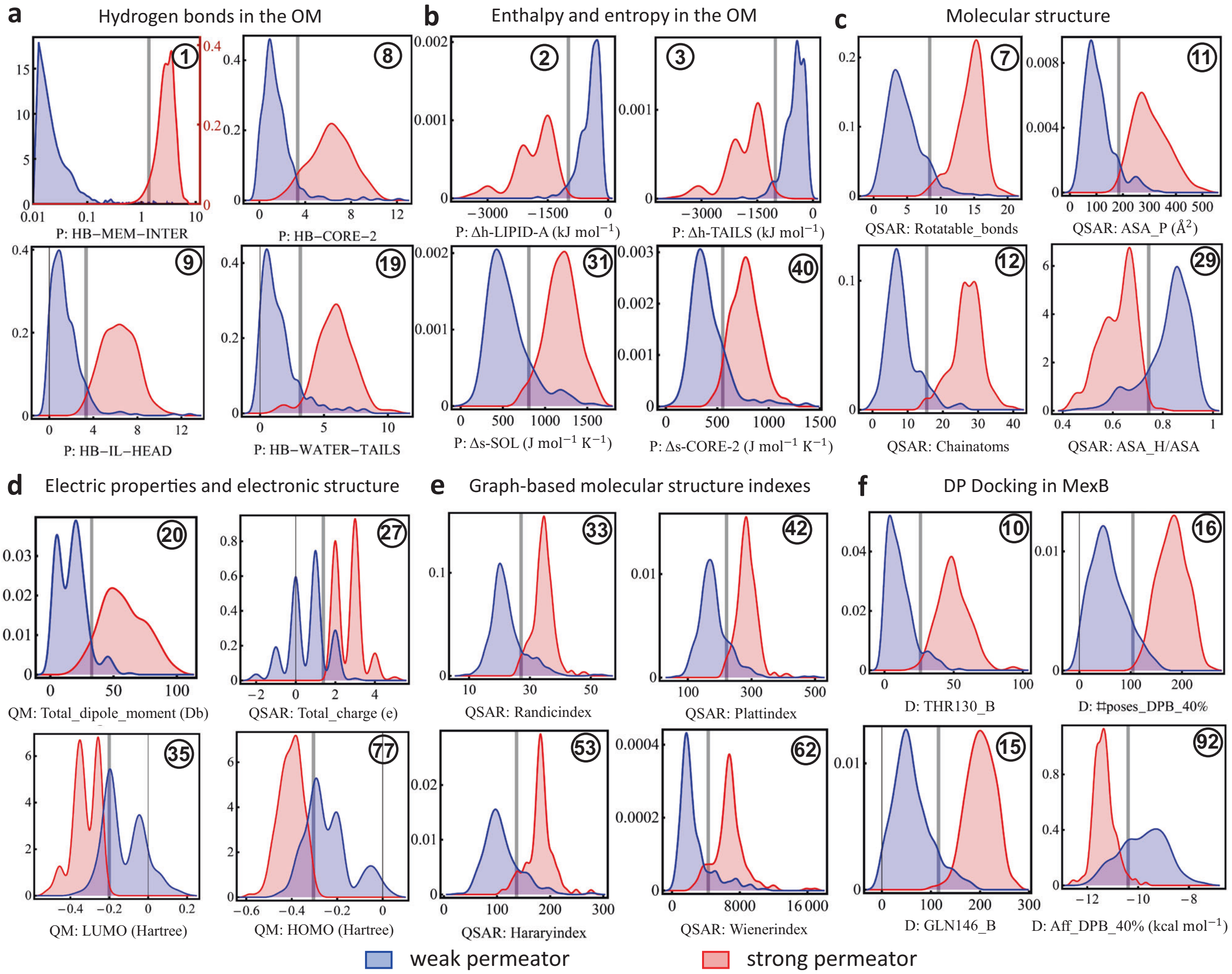


Fig. 6 | Single descriptor ranges according to class. Density distribution values across the range of selected individual descriptors associated with a particular target class given by their IC_{50} ratio, i.e., strong (red) or weak (blue) permeator, for the 501 compounds comprising the predictive group (Fig. 5). The vertical gray line indicates the class threshold estimated by an SVM algorithm. We considered all descriptors from the top 9 clusters from our predictive model (Fig. 4) The descriptors shown hold high predictability scores across general categories (see full list in Supplementary Tables S4 and S5) described as follows: **a**. Hydrogen bonds in the OM. Top panel uses two vertical scales and an horizontal logarithmic scale. The red vertical scale corresponds to strong permeators (red). All other panels use a single scale for both categories of compounds. **b** Enthalpy and entropy in the OM, **c** Molecular structure, **d** Electric properties and electronic structure, **e** Graph-based molecular structure indexes, and **f** DP docking in MexB. The circled number in each panel list the ranking according to their single-descriptor predictability scores (Supplementary Tables S4 and S5).

along the diffusion pathway. Counterintuitively, larger hydrophobic area does not favor the passage of molecules across the OM, a feature that correlates with the need for localized charges (+2*e* and higher) and stronger dipole moment. Graph-based molecular structure indicators such as Randić, Harary, Wiener, and Platt indexes are generally higher for strong permeators [Fig. 6e]. Finally, for docking descriptors, it is found that strong permeators hold higher number of poses inside the DP in contact with at least *z*% (*z* = 20, 30, 40) of the residues lining the pocket. This yields higher free energy bindings for strong permeators, and a consistently higher number of contacts to key residues inside the DP in MexB, than weak permeators [Fig. 6f].

The resultant cutoff values are then tested in the entire set of active compounds (600 compounds) and their associated evaluation metrics are computed for two sets, i.e., set G and the set of all active compounds. These implementations identify simple rules of permeation with very good accuracy in the entire set of active compounds. The circled number in each panel of Fig. 6 refers to the ranking of the descriptor in question according to accuracy for the entire set of active compounds. Supplementary Tables S4 and S5 list the complete ranking results of the individual descriptors. The resultant one-dimensional ranking is dominated by permeation descriptors. Among those of the top-10, 8 are permeation descriptors. Specifically, the hydrogen bonds and enthalpies computed across different sub-regions of the OM are, overall, the best individual descriptors at determining permeation. Accordingly, and in agreement with our findings at the cluster level, the persistence of hydrogen bonds (time-averaged over 20 ns) between the compound and inner leaflet of the OM at the membrane-water interface (HB-MEM-INTER) is the best single descriptor overall (i.e., best accuracy or a_0), and also the best one at detecting strong permeators (i.e., best positive predictive value or PPV). More than four fifths (0.836) of the active compounds analyzed, including 95.5% of the compounds in set G, that make (time-averaged) 1.36 or more HB with the surface of the outer membrane, are strong permeators.

Among the enthalpy descriptors, and again in agreement with the cluster-level analysis, the enthalpy calculated in the LIPID-A sub-region of the OM ranks the highest in accuracy and in determining strong permeators (i.e., high PPV). More than four fifths of the active compounds (0.815), including 93.7% of the set G, with an enthalpy value in the LIPID-A sub-region smaller than −988.85 kJ/mol, are strong permeators. Combining this information with knowledge of the entropy in the water neighboring the OM (Δs-SOL), increases the fraction of strong permeators correctly identified up to 0.854. Enthalpies associated with other OM sub-regions are next in the ranking of descriptors going from top-3 down to top-6, while the HB in the OM core 2 and heads, are top-7 and 9, respectively. 79.6% and 77.9% of the compounds making 3.34 (time-averaged) HB with the core 2 and heads subregion of the OM, respectively, are strong permeators. The QSAR descriptor quantifying the number of rotational bonds ranks eight and it is the best individual QSAR descriptor overall able to identify 89.2% and 78.06% of the weak and strong permeators, respectively. The best individual docking descriptor is the number of contacts between a compound and the residue THR130 in the DP of MexB. 88.9% of the compounds making less than 26 contacts with this residue are weak permeators. Ranking the compounds according to their negative predictive value (NPV), highlights descriptors that are good at detecting weak permeators. According to this ranking, the molecular orbitals and the number of donors, are the best descriptors at identifying this property. Between 90% and 91% of the active compounds with HOMO (LUMO) levels greater than −0.3045 (−0.2006) Hartree and/or less than four (3.88 on average) donors, are weak permeators.

Next, extending beyond the individual consideration of top-9 clusters (i.e. 1D), we analyze the descriptors for their joint behavior as duets (2D) and triplets of descriptors (3D) (see examples in Supplementary Figs. S16 and S17 in the Supplementary Note 4). Indeed, by carrying out higher order analyzes (i.e., two and more descriptors) it is possible to improve the classification metrics obtained with only one descriptor. For example, combining the permeation descriptor HB-MEM-INTER with the total polar surface area (i.e., QSAR: ASA_P($\AA^2$)), the classification accuracy increases up to 87.27% while the PPV to 86.68% when tested in the group of all compounds. Focusing on the set G, a combination of the total charge and the total polar surface area yield an accuracy of 99%, which is in agreement with recent studies via the whole-cell accumulation of 345 diverse compounds finding a strong correlation between these physicochemical properties with, *P. aeruginosa* accumulation[68]. Further, in a three-descriptor analysis, the highest accuracy score found is 88.4% when combining the enthalpy in the LPS sub-region core 2 (P: Δh-CORE-2), the hydrophobic surface ratio of the compound (QSAR: ASA_H/ASA), and the number of contacts that the compound makes with residue PHE615 in the DP of MexB. The PPV and NPV scores are 86.1% and 90.0%, respectively. Certainly, there are many more combinations that produce slightly lower but competitive scores [Supplementary Fig. S13]. We have listed the most important in the two- and three-descriptor analysis in the Supplementary Tables S7 and S8, respectively, in the Supplementary Note 4. According to the analysis in Section "Relationship between permeation predictability and chemical structure ofcompounds", the metric values found for the three-descriptors case lie at the ceiling of the evaluation metric given the behavior of the compounds in the sets R and B. Hence, we do not expect further improvements when going to higher dimensions without breaking down the groups of compounds as we did in the previous section. This is in agreement with the reduction algorithm of the Section "Non-trivial relationships among the different classes of descriptors", where the random forest determined that nine clusters ($x = 9$) optimizes the classification performance of the entire set, where each classification tree is constructed with the information of three descriptors (i.e., $x^{1/2}$). The information extracted by this analysis recovers the performance of the nonlinear method (i.e., random forest) and it therefore exhaust the possibilities of better performances with the totality of our data.

Implementation of our statistical model and the permeation rules

As an example of how our analysis can be applied on additional compounds to predict their OM permation, we carry out a testing evaluation on ten new compounds not seen before at any stage of this study. Figure 7a, b details the structures of these molecules, and shows the classification results according to permeation, respectively. The structures and origin of these compounds can be bounded to some of the main 16 chemotypes outlined in Fig. 1a in the following way: five compounds of chemotype 14 (C2-C6), three of chemotype 1 (C0, C7,C8), one of chemotype 7 (C1), and one of chemotype 9 (C9). Using the most consistent descriptors across several random splits of

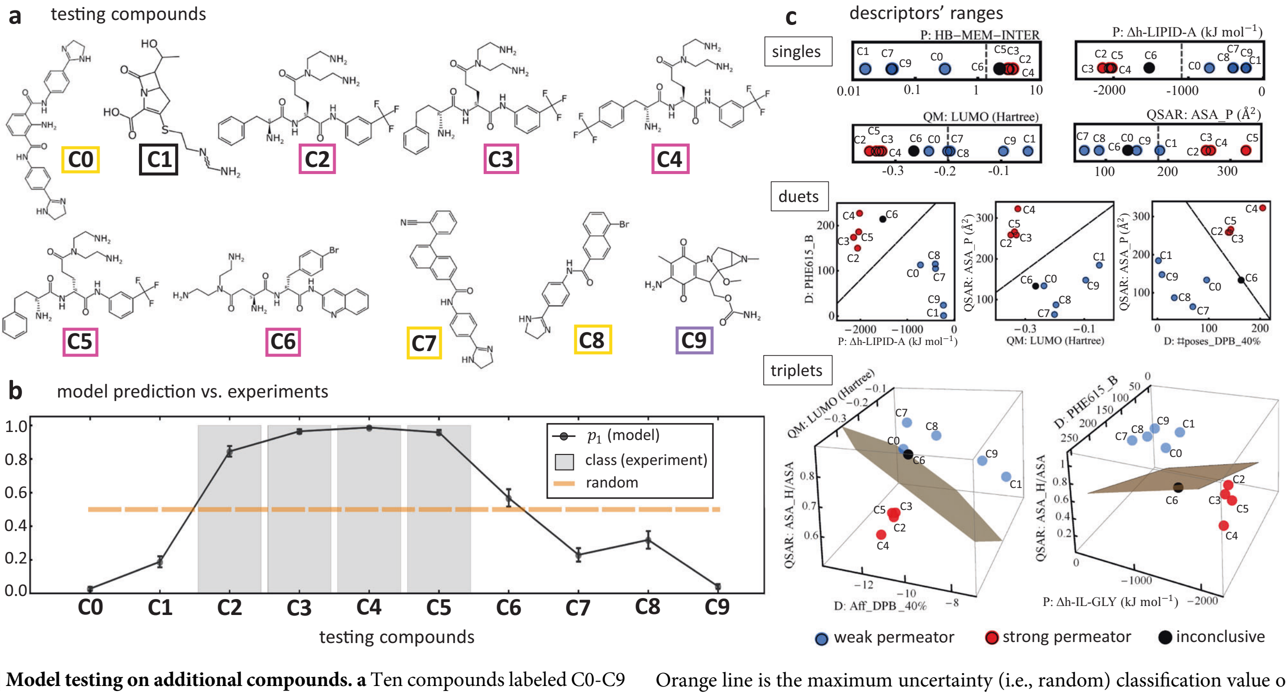


Fig. 7 | Model testing on additional compounds. a Ten compounds labeled C0-C9 structurally classified using the color code defined Fig. 1a. **b** Model prediction associated with the ten compounds (solid black line) against the target class (gray bars) assigned from the IC_{50} ratio measured experimentally in *Pseudomonas aeruginosa*. The prediction quantifies the probability that a given testing compound is a strong permeator, p_1. Error bars are the standard deviation of 100 model runs. Orange line is the maximum uncertainty (i.e., random) classification value of 0.5. The value of p_1 of compound C6 lies very close to this high uncertainty value. **c** Ranges of high-ranked descriptors as singles (top), duets (center), and triplets (bottom) for the testing compounds and classification given by the model. Each panel shows how these compounds' properties compare to the classification boundary of the training set (dark line or plane).

the training data, the algorithm calculates the probabilities that each compound is either a strong (class 1) or a weak (class 0) OM permeator i.e., p_1 and p_0, respectively. The model calculates consistent prediction probabilities for nine out of the ten compounds. Accordingly, it predicts that four of them are strong permeators (C2-C5) with an average probability of $p_1 = 0.94$, and five of them are classified as weak permeators (C0,C1,C7-C9) with an average probability of $p_0 = 0.84$. Experimental results of the IC_{50} ratios on these compounds in *P. aeruginosa* validates the prediction results for these nine testing compounds. This is illustrated in Fig. 7b where we plot the values of p_1 for each compound (black dots) against the target class found experimentally (gray bars). The probabilities calculated for the remaining compound (C6) are $p_1 = 0.56$ and $p_0 = 0.44$. These values are very close to those of a random classification (orange line), akin to the values found for compounds of the set B (see Fig. 5b). In addition, a structural analysis of this compound indicates that it is akin to subgroup SB167, where the trends between the descriptors and the molecule's permeation class are not conclusive. The subgroup SB167, which is characterized by having compounds containing an amide derived from 3-aminoquinoline (Supplementary Fig. S14 in the Supplementary Note 3), gathers a large fraction of compounds of the set B. Therefore, we determine that the classification of compound C6 is inconclusive. Our experiments indicate that the compound C6 is a weak permeator.

Figure 7c illustrates the values of some key descriptors of the testing compounds as singles (top panel), duets (central panel), and triplets (bottom panel), and how they compare with the permeation boundaries determined in the previous section with the compounds of the training batch. For the high-ranked descriptors (i.e., HB-MEM-INTER and enthalpy in the Lipid A subregion of the OM), most of the testing compounds are in very good agreement with the respective classification boundaries. This further supports our analysis and grants confidence in the applicability to other batches of compounds. Lower-ranked descriptors as singles, such as the LUMO level and polar surface area (ASA_P), have an excellent performance as a duet and also as triplets, highlighting an existing complementary relationship among the descriptors. Compound C6 shows an interesting behavior where, for some combination of descriptors, it sides with the strong permeators, while for other, it sides with the weak permeators. This inconclusive outcome with respect to descriptors sheds light on the limitations of an approach based only on these properties to classify some types of compounds. However, having characterized these specific subgroups by their distinguishable structural markers has prevented a possible error in its classification.

Discussion

Molecular diffusion across bacterial membranes is a very complex process which is largely dependent on the composition of the OM. In particular, Gram-negative organisms have capitalized on very sophisticated mechanisms to "screen" the passage of molecules from extracellular to intracellular regions. As a consequence, it is extremely difficult to develop new compounds to fight bacterial infections without a proper knowledge of the physical rules governing the overall translocation process. Our results show that small molecule permeation across the OM of *P. aeruginosa* can be predicted with high precision and accuracy based on the abilities of compounds to inhibit growth of cells with the native and hyperporinated OM. These predictions can be made for a library of structurally diverse compounds that likely use different mechanisms to penetrate the OM permeability barrier. However, the robustness of the model is increased by introducing descriptors of passive permeation across the OM model (permeation descriptors). This result provides further support to a recent finding that most antibiotics and nutrients accumulate inside cells by diffusion through the lipid bilayer of the OM.

We utilized the protocols of how the descriptors are calculated from our earlier work in ref. 37 and as before, inferred insights into the permeability barriers of *Pseudomonas aeruginosa* from the molecules' descriptor and the IC_{50} ratios. However, the current study has several crucial differences with ref. 37: (1) Our current dataset of compounds comprises 1260 molecules from at least 16 different structural chemotypes as listed in Fig. 1. Reference [37] used 290 compounds all belonging to a single structural class (Rempex), essentially peptidomimetic compounds. (2) Our analysis of correlations among descriptors and clustering is generalized to account for non-linear relationships going beyond what ref. 37 implemented, which is a traditional linear correlation analysis. (3) The target class in our current analysis is the OM permeation and inhibition capabilities that a given compound has given the values of its mechanistic descriptors. Reference 37 focused on efflux avoidance and did not consider OM permeation because Rempex compounds are polycationic and readily permeate the OM. (4) We deliver detailed lists of the ranges of the key single, pairs, and triplets sets of descriptors found to be predictive of OM permeation. Reference 37 lists single properties only and no explicit range of values for the descriptors is provided. (5) We demonstrate the applicability of the rules uncovered from our study into a new library of compounds, which illustrates the advantages as well as the limitations of our approach.

An accurate calculation of drug permeation is very challenging and requires extensive computing resources as the permeation is related to the exponent of the potential mean force[69]. It becomes impractical when the calculations are required for a large number of compounds such those considered here. Instead, we have generated a large number of molecular and mechanistic descriptors and used machine learning to identify the descriptors that are predictive of compound permeation. Interestingly, we find that descriptors indicative of interactions with different regions of the OM (HB, Δh) are among top ranked predictors for permeation. In fact, a favorable interaction with the membrane can lead to a positive chemical potential in virtue of high compound density, leading to an increase in translocation across the OM[70]. Not surprisingly, the presence of descriptors associated with both the Lipid-A and the core-2 of LPS indicates their importance during passive diffusion. With the exception of the highest ranked descriptor, hydrogen-bonding interactions at the membrane-water interface (HB-MEM-INTER), other highly ranked permeability descriptors can be substituted by QSAR descriptors, resulting in models with slightly lower accuracy. However, this can be beneficial when detailed calculations such MD simulations are not available. This is expected since inherent physicochemical properties of the compounds can be indicative of the chemical space they prefer. In particular, the importance of molecular connectivity (Randic index) and ASA properties for accurate prediction in our analysis indicates that surface exposure to the environment may be critical during the passage of the compounds across the OM. From the mechanistic perspective, the exposure of hydrophobic surfaces (ASA_H) can indeed enhance the interaction with the hydrophobic regions of the membrane[71] and is apparently in our calculations a more relevant descriptor for the prediction of permeation than the water-octanol partition coefficients (LogP). We want to highlight however, that QSAR descriptors by themselves are unable to replicate the level of accuracy obtained by using permeation descriptors, highlighting the importance of descriptors generated using all-atom MD simulations of compounds with a realistic Gram-negative OM model.

In reference to docking descriptors, our calculations indicate that distance to/contacts with ASN125 and THR130 are relevant. These two residues found in the substrate-binding pockets of RND transporters and have been shown to play a role in efflux[37,72,73]. In MexB, both THR130 and ASN135 are in the cave region of the Distal Binding Pocket and thus are in the path of substrate translocation. MD simulations and mutational analyses suggest that interaction between substrates and these residues contribute to substrate specificity and efflux efficiency[37,72]. Previous studies further suggested that the compound properties needed for permeation across the outer membrane and for recognition by efflux pumps complement each other[74]. In other words, compounds permeating the outer membrane are better substrates of efflux, because pumps are exposed to higher concentrations of these compounds in the periplasm.

From a data analysis perspective, our main challenge is to circumnavigate the multidimensional set of descriptors in a way that rationally lessens the computational cost of running a classification algorithm on every single combination of them. The proposed analysis is designed to map, navigate, and reduce this extensive parameter set highlighting the theoretical/computational quantities that better correlate with our experiments. Since the reduction is done at the level of the cluster, it is key to use an optimal clustering technique that accounts for the nonlinearities found in the data. Hierarchical clustering is adequate because it uses correlations as its similarity measure[63] and when applied to a ranked data set (i.e., ranked correlations), it accounts for nonlinear monotonic relationships among the different features [Fig. 2]. In addition, we use a random forest classifier[65] that also identifies and benefits from nonlinear trends found in complex datasets[75–77]. Using a cluster-centered framework alleviates part of the computational cost of testing myriad combinations of parameters and grants a wider perspective on the overall properties that are linked to improvements in their permeation properties. As shown in Supplementary Fig. S8, although there are performance differences when using different members of a cluster, there are highly correlated descriptors with a performance that are comparable to each other, and the best descriptor is the one that is broadly represented by a general property of the compound (e.g., size) rather than a very specific one (e.g., Wiener index). Even after the implementation of these techniques, the number of possible combinations of descriptors is still very large. A sampling technique was therefore implemented to scan the clusters, rank them according to their predictive capacity, and in parallel, test the different combinations of descriptors of each sample. The effectiveness of this technique at finding a good parameter space for permeation is demonstrated by the more rigorous reverse analysis shown in Supplementary Fig. S7, where alternative combinations of descriptors within the same cluster were identified, and their score compares to that of the combination found during the sampling process. This effectiveness would have not been possible if the clustering technique implemented ignored the nonlinear relationships, since it yields a smaller number of clusters (29 clusters) restricting the parameter space and the descriptors exploration.

Our analysis identified nine key clusters containing the relevant descriptors that maximize the model's prediction performance, which in turn, allowed us to classify the compounds according to the consistency in their predictability. This classification identified three sets of compounds [Fig. 5]. The largest of them (set G accounting for 83.6% of the compounds) is the most consistent yielding correct predictions in every calculation performed. This gives us confidence in the robustness of the properties captured and in the statistical techniques employed. In reference to the sets R and B, a structural examination using a complete Tanimoto similarity analysis reveals that 10 subgroups contain most of these compounds pointing to a structural connection with their permeation predictability. We find that for some of these subgroups there are subgroup-specific descriptors able to correctly classify the compounds and bypass the prediction difficulty. However, for other subgroups, even the best-ranked descriptors appear to be unable to separate the permeation classes. Though this is a limitation of this approach based on descriptors, this is only found on 4 subgroups of the structural chemotype 14 (Rempex), which is 1 among the 16 structural chemotypes considered in this study. Focusing on the majority set (i.e., set G), we find that it is characterized by projecting strong and weak permeators in well-segregated parameter regions [Supplementary Fig. S12] allowing us to extract simple empirical rules associated with the descriptor space akin to OM permeation. Using the descriptors of the nine key clusters, we established one-, two- and three-body (i.e., descriptors) rules that better describe the patterns found in all of the active compounds. For example, the one-body analysis highlights the role of the permeation descriptors, especially the hydrogen bonds and the enthalpy computed in several regions of the OM [Supplementary Tables S5 and S6]. The patterns found reveal that weak permeators are characterized by having very limited hydrogen bond stabilization with the OM, as well as, having a very weak enthalpy of association [Fig. 6]. The two- and three-body analysis [Supplementary Figs. S13 and S14] revealed a complementary role of the compound's polar and hydrophobic surface areas that enhance the number of compounds correctly classified [Supplementary Tables S7 and S8]. Docking descriptors, particularly those describing properties of the deep pocket of MexB, are found to be highly correlated with the OM permeation. In fact, there are many examples of three-descriptors sets yielding correct classification scores that are comprised by one permeation, one docking, and one QSAR descriptor [Supplementary Table S7]. This highlights a complex relationship among these types of descriptors that captures well-rounded properties of both, weak and strong permeation, and hence, it facilitates their correct identification. An application of these uncovered rules on a new batch of compounds demonstrate their predictability power and opens the door to similar data-driven studies in other Gram-negative pathogens. This analysis complements similar efforts at determining the key properties that distinguishes strong and weak permeators[78].

In summary, our work combines experimental, computational, and statistical protocols in order to identify the critical properties that optimally predicts the passage of molecules across the bacterial OM and inhibit growth of Gram-Negative *P. aeruginosa*. The successful approach was able to reduce the spectrum of relevant mechanistic properties in a set of chemically diverse compounds with known antibacterial activity into simple but non-trivial empirical rules for the prediction of strong or weak permeators. We hope that the found relationships can guide additional experimental efforts and accelerate the rational design of new classes of molecules for combating antibiotic resistant strains. Our approach can be expanded for targeting the permeability of molecules to different biological membranes, regardless of their composition or distribution.

Methods

Experimental methods and chemical syntheses

The experimental set-up has been reported before[37,70]. Briefly, *P. aeruginosa* cells were grown in Luria Bertani Broth (LB) (10 g tryptone, 5 g yeast extract, 5 g NaCl per liter, pH 7.0) at 37 °C with shaking. Inhibitory concentration (IC_{50}) determination was carried out using the 2-fold broth dilution method. Two independent experiments were carried out. The expression of the Pore was induced at $OD_{600} = 0.3$–0.4 by addition of 0.1 mM IPTG. Chemical structures of the assembled library of 1260 compounds and the measured IC_{50} values are available upon request.

Computational setup and protocols for computing molecular descriptors

QSAR, QM, and MD calculations. For each compound we considered the protonation/charge state most populated at physiological pH. We used the ChemAxon's Marvin suite of programs[79] to obtain standard 1-2-3D descriptors used in QSAR studies (e.g., numbers of heavy atoms, rotatable bonds, H-bond donors/acceptors, van der Waals volume and surface, etc. see ref. 37). The geometry of the major microspecies has been used to perform QM calculations with the Gaussian16 package[80] as previously described[81]. Employing a polarizable continuum model to mimic the effect of water solvent we optimized the ground-state structure and performed full vibrational analysis, obtaining real frequencies in all cases. On the optimized geometry, we performed single-point energy calculations in vacuum to generate the atomic partial charges fitting the molecular electrostatic potential. Under the constraint of reproducing the electric dipole moment of the molecule, we used the Merz-Kollman scheme[82]. Atomic partial charges were generated through the two-step restrained electrostatic potential method[83] implemented in the AnteChamber package[84]. With this program we derived general Amber force field (GAFF) parameters[85]. QM descriptors associated with the ground-state optimized structure include static polarizabilities, frontier molecular orbital energies, permanent dipole moment, and rotational constants. For each compound, we performed 1-μs-long all-atom MD simulation in explicit water solution (0.1 M KCl) using the Amber18 package as described before[81]. From MD simulations, we obtained structural and dynamic features of the compounds investigated by means of the CPPTRAJ program[86]. The number and population of structural clusters were determined using a hierarchical agglomerative algorithm[87].

P. aeruginosa OM set up for MD. The permeation descriptors were calculated from an outer membrane (OM) computational MD model of the Gram-negative bacteria *Pseudomonas aeruginosa*. Briefly, the OM consists of an inner leaflet composed of 1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine (DPPE) and an outer leaflet composed of a truncated LPS structure. The membrane is fully solvated using the TIP3P water[88] model and anionic charges in the LPS molecules are counter balanced with Ca2+ cations. A schematic representation of the model is provided in Fig. 1a and more details about its parameterization can be found in the original work[89]. The initial coordinates OM model are found in http://dqfnet.ufpe.br/biomat/software.html. The model has been parameterized in line with the GLYCAM force field[90] and parameters are adapted to run in the GROMACS[91] molecular dynamics engine.

Compounds were represented using the Amber force field. First, we optimized the ground-state structure of each compound employing a polarizable continuum model[92] as to mimic the effect of water solvent particularly to avoid formation of strong intramolecular H-bonds. This geometry was confirmed performing a full vibrational analyses, obtaining real frequencies in all cases. On the optimized geometry, we then performed single-point energy calculations in vacuum to generate the atomic partial charges fitting the molecular electrostatic potential. Under the constraint of reproducing the electric dipole moment of the molecule, we used the Merz-Kollman scheme[82] to construct a grid of points around the molecule. Atomic partial charges were then generated through the two-step restrained electrostatic potential method[83] implemented in the AnteChamber package[84]. Using this program, we derived general Amber force field (GAFF) parameter[85], which were transformed into GROMACS input files using the antechamber python parser interface (ACPYPE) tool[93]. In order to screen the molecular descriptors corresponding to the permeation along the OM membrane, each drug was placed into seven different molecular environments corresponding to specific regions along the normal of the OM [Fig. 1a]. These regions were explicitly selected in order to cover the influence of both the inner (DPPE) and outer leaflet (LPS) of the OM. Thus, seven independent simulations per drug were necessary in order to recapitulate the influence of the OM into the permeation process. The whole procedure was automated via a series of bash scripts, which iteratively connected the pulling code and energy minimization in GROMACS[91].

All simulations were run with the GROMACS 5.4.1 molecular dynamics engine[91] with a time step of 2 fs. The LINCS algorithm[94] was applied to constrain all bond lengths with a relative geometric tolerance of 10^{-4}. In line with its original parameterization, short-range interactions (vdW and Coulomb) were calculated using a cut-off scheme of 0.9 nm, which were evaluated based on a pair-list recalculated every five time steps. Long-range interactions were handled using a reaction field[95] correction with a permittivity dielectric constant of 66. After initial set-up, each system was energy minimized using 3000 steps of conjugated gradient, followed by a thermal equilibration of 1 ns. A harmonic potential of 1000 kJ mol^{-2}, along the Z vector connecting the center of mass (COM) of the drug and the OM of the membrane was applied in order to maintain the relative position of the drug with respect to each of the seven defined regions of the membrane [Fig. 1a]. During equilibration, bilayers were coupled to 1.0 bar using a Berendsen barostat[96] through a semi-isotropic approach with relaxation time of 1.0 ps. Afterwards, production runs were coupled using a Parrinello barostat[97] algorithm and a constant temperature of 310 K was maintained by weak coupling of the solvent and solute separately to a velocity-rescaling[98] scheme with a relaxation time of 1.0 ps. Production simulations were run for 20 ns and trajectories were saved each 20 ps. A total of 8841 (176 μs) trajectories were analyzed using in-home developed bash scripts, which were directly interconnected to the in-built GROMACS tools. Thus, for each simulation the following molecular descriptors were evaluated [Fig. 1a]: Number of hydrogen bonds between the drug with its first solvation shell (HB-WATER), number of hydrogen bonds between the drug and the surrounding OM environment (HB), lateral mean squared displacement of the Drug (Δxy), Total enthalpic component of interaction between drug and surrounding environment (Δh),and total cumulative entropy of the drug (Δs). All these analysis were carried with the in-built analysis tool set provided in GROMACS.

Ensemble docking to MexB. Docking descriptors were calculated using the default settings of the AutoDock Vina package[64] except for the exhaustiveness parameter which was set to 1024 (default of 8). Protein and ligand input files were prepared with AutoDock Tools[99]. Flexibility of docking partners was considered indirectly by using the ensemble of conformations. In particular, for each compound we used 10 different cluster representatives extracted from MD simulations in explicit water solution, while for MexB, we considered 6 conformations, including available X-ray crystal structures (PDB Ids 2V50, 3W9I, and 3W9J)[100,101] and MD snapshots extracted from MD simulations[46]. For each docking run, we retained the top 10 docking poses. Following ref. 102 we performed two sets of guided docking runs into the two major binding pockets of MexB: the access pocket of the access monomer (AP) and the deep binding pocket of the binding monomer (DP). In each case, the docking search was performed within a cubic volume of 40 × 40 × 40 Å^3 centered in the center of mass of the pocket. The interaction between compounds and MexB was quantified by means of a statistical analysis of all poses, yielding about 60 descriptors. These descriptors include average binding affinities (according to the docking scoring function) as well as the total number of contacts with single residues lining the two pockets (see Supplementary Table S1 in the supplementary).

Statistical methods

Agglomerative clustering. This is an unsupervised statistical technique that uses correlations among random variables to form groups (or clusters) of highly correlated quantities, resulting in clusters that are highly dissimilar from one another. This is a bottom-up technique that starts with clusters formed by a single random variable. Then the correlations coefficients among all the pairs are computed and ranked. The pair with the lowest dissimilarity measure is merged together into a cluster of size two. The dissimilarity D_{ij} is defined as the square-root of one minus the square of the correlation coefficient between the pair i and j:

$$D_{ij} = \sqrt{1 - R_{ij}^2}, \tag{1}$$

where, R_{ij} is the correlation coefficient between variables i and j. Subsequently, all correlation coefficients are computed again treating the cluster of two as a single variable in which the resultant correlation between the pair and another variable is derived as the average of the correlation with each member of the cluster individually. Then, the dissimilarity measures among all groups are ranked and the pair with the lowest one is merged into a larger cluster. This process is repeated until only one cluster remains.

In our analysis we implement the ranked correlations coefficients consisting on replacing the value of the random variable (i.e., molecular descriptor of the compound) for the low-to-high rank of such value within the distribution. For example, for the molecular property of molecular weight, the lightest compound would have a rank of one, the second lightest a rank of two, and so one. We do the same procedure for all descriptors. Then, the ranked correlations are calculated by computing the Pearson correlation coefficient over the list of ranked values.

In order to determine the optimal number of clusters we use the fractional variance explained defined as the ration of the variance between groups (i.e., points residing in different clusters) to the total variance (i.e., all points):

$$\sigma_{fve} = \frac{\sigma_{between\_groups}}{\sigma_{total}} = \frac{\sum_{ij|c_i \neq c_j} D_{ij}^2}{\sum_{ij} D_{ij}^2} \tag{2}$$

This quantity increases as the number of clusters decreases and then stabilizes, which points to an appropriate number of clusters. At this point variance from within clusters is small enough hinting at a relative closeness among points within clusters and otherwise for points in different clusters. The L method is employed to identify the optimal number of clusters n_c. First we create the list of the fractional variance explained σ_{fve} vs the number of clusters n. For each candidate number n we find the best straight line fit of all points before and after n, and compute the weighted sum of the root mean square error (RMSE) associated to the fits. The value of n that minimizes RMSE corresponds to the point in which the variance stopped increasing as a function of the number of clusters. We consider this point as the optimal number of clusters.

Cluster reduction algorithm. We start by dividing the set of 600 compounds into a large group of 480 compounds and a small group of 120. The large group is used to train/validate on a subset of x descriptors a nonlinear classifier and quantify the importance of each descriptor of the subset. The small group is used to test the efficacy of the trained/validated model at predicting the respective target class. The value of x is determined by the number of clusters considered in the calculation, and the subset of descriptors is comprised by one descriptor per cluster randomly selected. We start with the complete set of 37 clusters (i.e., $x = 37$ initially). From this set, 200 subsets of x descriptors each, are randomly assembled. For each subset, a random forest (i.e., bagged ensemble) classifier[65] comprised by $N_e = 1001$ estimators (i.e., 1001 classification trees) that use $x^{1/2}$ descriptors for each estimator assigned randomly with equal weight is implemented using the scikit-learn package in python[66]. Using the properties of the compounds in the training portion, each estimator determines which permeation class is more appropriate for each compound of the testing portion. The dominant class i.e., the one that is assigned by the majority of estimators, becomes the class prediction of the compound in question. The classification algorithm is trained over the target class of 95% of compounds in the large group of compounds, and the remaining 5% is used for validation, which helps to control for overfitting. For each subset of x descriptors, we carry out 50 classification runs over random 95:5 training/validation splits (see dotted circle in Fig. 4a). Hence, considering all 200 subsets of x descriptors, there are 10,000 classification runs that are carried out for each value of x. These calculations allow us to measure the cluster score according to their performance (see details below in cluster score). This measure is then used to rank the clusters accordingly. Finally, the number of clusters is reduced by eliminating the lowest scoring one and the cycle is restarted for the reduced set of clusters (i.e., $x \rightarrow x - 1$). In addition to this process, for each of the classification runs, the fitted model is tested on the small group of 120 compounds (orange arrow in Fig. 4a), where we compute the standard confusion matrix and its associated evaluation metrics of accuracy, precision, recall, specificity, and F1. In this way we keep track of how better or worse the model performs for the different combinations of x descriptors, as well as, when the number of clusters decreases and find the local optimal combination (see definition below).

Cluster score. We compute a performance score for each cluster by aggregating the contributions of each of the x tested descriptors in the random set. For each of the 200 random sets, which are tested across 50 training/validation splits (i.e., 10^4 calculations), we compute the performance score of each descriptor and aggregate the results according to the descriptor's cluster membership. The aggregated score value for each cluster is what we call cluster score. The performance score associated with each descriptor is defined as the impurity (i.e., Shannon) entropy, which measures how good the information from the descriptor decreases the classification uncertainty. This is calculated with the built-in function "feature_importances_" provided in the sklearn package[66].

Optimal combination. We define the optimal combination as the group of descriptors that maximizes the model performance accuracy of the testing portion of the data. In Fig. 4b this value is shown as the orange dot with the highest y-axis value, which occurs at $x = 9$.

Evaluation metrics. The model evaluation metrics computed in this work are based on combinations of the output from the traditional confusion matrix, which compares the truthfulness of the prediction (i.e., true or false), with the binary classification (class 1 or class 0) of the real data. Hence, each prediction outcome can be classified as either true positive (TP, or class 1 correctly identified), true negative (TN, or class 0 correctly identified), false positive (FP, or a real class 0 identified as class 1), or false negative (FN, or real class 1 identified as class 0). The accuracy (a_0) is the ratio of correct predictions to all predictions, i.e., the fraction of correct predictions:

$$a_0 = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

In addition, the measure of precision is targeted to minimize false positives, and it is also known as the positive predictive value (PPV). It quantifies the fraction of positive predictions that are real:

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

An equivalent measure to precision, but targeted to quantify the fraction of negative predictions that are real, is known as the negative predictive value (NPV):

$$NPV = \frac{TN}{TN + FN}. \tag{5}$$

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated and analyzed during this study are included in this published article (and its supplementary information files). Our library of compounds used in this study, including chemotype, SMILES, descriptor values and IC_{50}'s, is provided in the Supplementary Data 1. The definitions of all descriptors are provided in the Supplementary Data 2.

Code availability

All algorithms utilized in this study are open-source and referenced in the manuscript and the supplementary information.

Received: 4 October 2023; Accepted: 27 March 2024;
Published online: 12 April 2024

References

1. World Health Organization. *Antibacterial agents in clinical development: An analysis of the antibacterial clinical development pipeline, including tuberculosis*. Tech. Rep., (World Health Organization, 2017). http://www.jstor.org/stable/resrep35853.1.
2. Bush, K. & Page, M. G. P. What we may expect from novel antibacterial agents in the pipeline with respect to resistance and pharmacodynamic principles. *J. Pharmacokinet. Pharmacodyn.* **44**, 113–132 (2017).
3. Li, X.-Z., Plésiat, P. & Nikaido, H. The challenge of efflux-mediated antibiotic resistance in gram-negative bacteria. *Clin. Microbiol. Rev.* **28**, 337–418 (2015).
4. Krishnamoorthy, G. et al. Synergy between active efflux and outer membrane diffusion defines rules of antibiotic permeation into gram-negative bacteria. *mBio* **8**, e01172–17 (2017).
5. Masi, M., Réfregiers, M., Pos, K. M. & Pagés, J.-M. Mechanisms of envelope permeability and antibiotic influx and efflux in gram-negative bacteria. *Nat. Microbiol.* **2**, 17001 (2017).

6. Fernández, L. & Hancock, R. E. W. Adaptive and mutational resistance: role of porins and efflux pumps in drug resistance. *Clin. Microbiol. Rev.* **25**, 661–681 (2012).

7. Zgurskaya, H. I., López, C. A. & Gnanakaran, S. Permeability barrier of gram-negative cell envelopes and approaches to bypass it. *ACS Infect. Dis.* **1**, 512–522 (2015).

8. Zgurskaya, H. I. & Rybenkov, V. V. Permeability barriers of gram-negative pathogens. *Ann. N. Y. Acad. Sci.* **1459**, 5–18 (2020).

9. Viale, P., Giannella, M., Tedeschi, S. & Lewis, R. Treatment of mdr-gram negative infections in the 21st century: a never ending threat for clinicians. *Curr. Opin. Pharmacol.* **24**, 30–37 (2015).

10. Antibiotic resistance threats in the united states, https://www.cdc.gov/drugresistance/pdf/threats-report/2019-ar-threats-report-508.pdf (2019).

11. Fraimow, H. S. & Tsigrelis, C. Antimicrobial resistance in the intensive care unit: mechanisms, epidemiology, and management of specific resistant pathogens. *Crit. Care Clin.* **27**, 163–205 (2011).

12. Theuretzbacher, U. Global antimicrobial resistance in gram-negative pathogens and clinical need. *Curr. Opin. Microbiol.* **39**, 106–112 (2017).

13. Who global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed (2017).

14. Silver, L. L. Challenges of antibacterial discovery. *Clin. Microbiol. Rev.* **24**, 71–109 (2011).

15. Pang, Z., Raudonis, R., Glick, B. R., Lin, T.-J. & Cheng, Z. Antibiotic resistance in pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. *Biotechnol. Adv.* **37**, 177–192 (2019).

16. Vesselinova, N., Alexandrov, B. S. & Wall, M. E. Dynamical model of drug accumulation in bacteria: Sensitivity analysis and experimentally testable predictions. *PloS One* **11**, e0165899 (2016).

17. Westfall, D. A. et al. Bifurcation kinetics of drug uptake by gram-negative bacteria. *PLOS One* **12**, 1–18 (2017).

18. Nichols, W. W. Modeling the kinetics of the permeation of antibacterial agents into growing bacteria and its interplay with efflux. *Antimicrobial Agents Chemother.* **61**, e02576–16 (2017).

19. Manrique, P. D. & Gnanakaran, S. Microscopic approach to intrinsic antibiotic resistance. *J. Phys. Chem. B* **125**, 3114–3118 (2021).

20. Manrique, P. D., López, C. A., Gnanakaran, S., Rybenkov, V. V. & Zgurskaya, H. I. New understanding of multidrug efflux and permeation in antibiotic resistance, persistence, and heteroresistance. *Ann. N. Y. Acad. Sci.* **1519**, 46–62 (2023).

21. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinforma. Chem. AABC* **8**, 37–47 (2015).

22. Durrant, J. D. & Mccammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9**, 71 (2011).

23. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* **59**, 4035–4061 (2016).

24. Aminpour, M., Montemagno, C. D. & Tuszynski, J. A. An overview of molecular modeling for drug discovery with specific illustrative examples of applications. *Molecules* **24**, 1693 (2019).

25. López, C. A., Zgurskaya, H. & Gnanakaran, S. Molecular characterization of the outer membrane of pseudomonas aeruginosa. *Biochim. Biophys. Acta Biomembranes* **1862**, 183151 (2020).

26. Parkin, J., Chavent, M. & Khalid, S. Molecular simulations of gram-negative bacterial membranes: A vignette of some recent successes. *Biophys. J.* **109**, 461–468 (2015).

27. Carpenter, T. S., Parkin, J. & Khalid, S. The free energy of small solute permeation through the escherichia coli outer membrane has a distinctly asymmetric profile. *J. Phys. Chem. Lett.* **7**, 3446–3451 (2016).

28. Kim, S. et al. Bilayer properties of lipid a from various gram-negative bacteria. *Biophys. J.* **111**, 1750–1760 (2016).

29. Hsu, P.-C., Jefferies, D. & Khalid, S. Molecular dynamics simulations predict the pathways via which pristine fullerenes penetrate bacterial membranes. *J. Phys. Chem. B* **120**, 11170–11179 (2016).

30. López, C. A., Travers, T., Pos, K. M., Zgurskaya, H. I. & Gnanakaran, S. Dynamics of intact mexab-oprm efflux pump: Focusing on the mexa-oprm interface. *Sci. Rep.* **7**, 16521 (2017).

31. Bruzzese, A., Dalton, J. A. R. & Giraldo, J. Statistics for the analysis of molecular dynamics simulations: providing p values for agonist-dependent gpcr activation. *Sci. Rep.* **10**, 19942 (2020).

32. Gapsys, V. & de Groot, B. L. On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. *eLife* **9**, e57589 (2020).

33. Likić, V. A., Gooley, P. R., Speed, T. P. & Strehler, E. E. A statistical approach to the interpretation of molecular dynamics simulations of calmodulin equilibrium dynamics. *Protein Sci.* **14**, 2955–2963 (2005).

34. Cooke, B. & Schmidler, S. C. Statistical prediction and molecular dynamics simulation. *Biophys. J.* **95**, 4497–4511 (2008).

35. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in trna:protein complexes. *Proc. Natl Acad. Sci.* **106**, 6620–6625 (2009).

36. Manrique, P. D. et al. Network analysis uncovers the communication structure of sars-cov-2 spike protein identifying sites for immunogen design. *iScience* **26**, 105855 (2023).

37. Mehla, J. et al. Predictive rules of efflux inhibition and avoidance in pseudomonas aeruginosa. *mBio* **12**, e02785–20 (2021).

38. May, K. L. & Grabowicz, M. The bacterial outer membrane is an evolving antibiotic barrier. *Proc. Natl Acad. Sci.* **115**, 8852–8854 (2018).

39. Strateva, T. & Yordanov, D. Pseudomonas aeruginosa – a phenomenon of bacterial resistance. *J. Med. Microbiol.* **58**, 1133–1148 (2009).

40. Chatterjee, M. et al. Antibiotic resistance in pseudomonas aeruginosa and alternative therapeutic options. *Int. J. Med. Microbiol.* **306**, 48–58 (2016).

41. Breidenstein, E. B., de la Fuente-Núñez, C. & Hancock, R. E. Pseudomonas aeruginosa: all roads lead to resistance. *Trends Microbiol.* **19**, 419–426 (2011).

42. Schweizer, H. P. Efflux as a mechanism of resistance to antimicrobials in pseudomonas aeruginosa and related bacteria: unanswered questions. *Genet. Mol. Res. GMR* **2**, 48–62 (2003).

43. Nikaido, H. & Zgurskaya, H. I. Antibiotic efflux mechanisms. *Curr. Opin. Infect. Dis.* **12**, 529–536 (1999).

44. Alav, I. et al. Structure, assembly, and function of tripartite efflux and type 1 secretion systems in gram-negative bacteria. *Chem. Rev.* **121**, 5479–5596 (2021).

45. Vargiu, A. V. et al. Computer simulations of the activity of rnd efflux pumps. *Res. Microbiol.* **169**, 384–392 (2018).

46. Ramaswamy, V. K., Vargiu, A. V., Malloci, G., Dreier, J. & Ruggerone, P. Molecular determinants of the promiscuity of mexb and mexy multidrug transporters of *p*seudomonas aeruginosa. *Front. Microbiol.* **9**, 1144 (2018).

47. 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.* **1459**, 38–68 (2020).

48. Ornik-Cha, A. et al. Structural and functional analysis of the promiscuous acrb and adeb efflux pumps suggests different drug binding mechanisms. *Nat. Commun.* **12**, 6919 (2021).

49. Klenotic, P. A., Moseng, M. A., Morgan, C. E. & Yu, E. W. Structural and functional diversity of resistance–nodulation–cell division transporters. *Chem. Rev.* **121**, 5378–5416 (2021).

50. Rybenkov, V. V. et al. The whole is bigger than the sum of its parts: Drug transport in the context of two membranes with active efflux. *Chem. Rev.* **121**, 5597–5631 (2021).
51. Richter, M. F. et al. Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **545**, 299–304 (2017).
52. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
53. Leus, I. V. et al. Property space mapping of pseudomonas aeruginosa permeability to small molecules. *Sci. Rep.* **12**, 8220 (2022).
54. Gaucher, B. & Dreier, J. Efflux-pump inhibitors and therapeutic uses thereof. https://patents.google.com/patent/WO2017042099A1/en (2016).
55. Renau, T. E. et al. Peptidomimetics of efflux pump inhibitors potentiate the activity of levofloxacin in pseudomonas aeruginosa. *Bioorg. Med. Chem. Lett.* **12**, 763–766 (2002).
56. Haynes, K. M. et al. Identification and structure-activity relationships of novel compounds that potentiate the activities of antibiotics in escherichia coli. *J. Med. Chem.* **60**, 6205–6219 (2017).
57. Cao, F. et al. Identification and structure-activity relationships for a series of n, n-disubstituted 2-aminobenzothiazoles as potent inhibitors of s. aureus. *Bioorg. Med. Chem. Lett.* **89**, 129301 (2023).
58. D'Cunha, N. et al. Mechanistic duality of bacterial efflux substrates and inhibitors: Example of simple substituted cinnamoyl and naphthyl amides. *ACS Infect. Dis.* **7**, 2650–2665 (2021).
59. Zhao, H. et al. Small molecule condensin inhibitors. *ACS Infect. Dis.* **4**, 1737–1745 (2018).
60. Green, A. T. et al. Discovery of multidrug efflux pump inhibitors with a novel chemical scaffold. *Biochim. Biophys. Acta Gen. Subj.* **1864**, 129546 (2020).
61. Abdali, N. et al. Reviving antibiotics: Efflux pump inhibitors that interact with acra, a membrane fusion protein of the acrab-tolc multidrug efflux pump. *ACS Infect. Dis.* **3**, 89–98 (2017).
62. Tanimoto, T. *An Elementary Mathematical Theory of Classification and Prediction* (International Business Machines Corporation, 1958).
63. Salvador, S. & Chan, P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *16th IEEE International Conference on Tools with Artificial Intelligence*, 576–584 (IEEE, 2004).
64. Trott, O. & Olson, A. J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
65. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
66. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).
68. Geddes, E. J. et al. Porin-independent accumulation in pseudomonas enables antibiotic discovery. *Nature* **624**, 145 – 153 (2023).
69. Menichetti, R., Kanekal, K. H. & Bereau, T. Drug-membrane permeability across chemical space. *ACS Cent. Sci.* **5**, 290–298 (2019).
70. Mansbach, R. A. et al. Machine learning algorithm identifies an antibiotic vocabulary for permeating gram-negative bacteria. *J. Chem. Inf. Model.* **60**, 2838–2847 (2020).
71. Yoshimura, F. & Nikaido, H. Permeability of pseudomonas aeruginosa outer membrane to hydrophilic solutes. *J. Bacteriol.* **152**, 636–642 (1982).
72. Gervasoni, S. et al. Molecular determinants of avoidance and inhibition of pseudomonas aeruginosa mexb efflux pump. *mBio* **14**, e01403–23 (2023).
73. Morgan, C. E. et al. Cryoelectron microscopy structures of adeb illuminate mechanisms of simultaneous binding and exporting of substrates. *mBio* **12**, e03690–20 (2021).
74. Cooper, S. J. et al. Molecular properties that define the activities of antibiotics in escherichia coli and pseudomonas aeruginosa. *ACS Infect. Dis.* **4 8**, 1223–1234 (2018).
75. Auret, L. & Aldrich, C. Interpretation of nonlinear relationships between process variables by use of random forests. *Miner. Eng.* **35**, 27–42 (2012).
76. Deloncle, A., Berk, R. A., D'Andrea, F. & Ghil, M. Weather regime prediction using statistical learning. *J. Atmos. Sci.* **64**, 1619 – 1635 (2007).
77. Touw, W. G. et al. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? *Brief. Bioinforma.* **14**, 315–326 (2013).
78. Ude, J. et al. Outer membrane permeability: Antimicrobials and diverse nutrients bypass porins in pseudomonas aeruginosa. *Proc. Natl Acad. Sci.* **118**, e2107644118 (2021).
79. ChemAxon. Marvin suite, https://chemaxon.com (2017).
80. Frisch, M. J. et al. Gaussian, inc. https://gaussian.com (2016).
81. Gervasoni, S. et al. Ab-db: Force-field parameters, md trajectories, qm-based data, and descriptors of antimicrobials. *Sci. Data* **9**, 148 (2022).
82. Singh, U. C. & Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **5**, 129–145 (1984).
83. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *J. Phys. Chem.* **97**, 10269–10280 (1993).
84. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **25 2**, 247–60 (2006).
85. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
86. Roe, D. R. & Cheatham, T. E. I. Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
87. Shao, J., Tanner, S. W., Thompson, N. & Cheatham, T. E. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *J. Chem. Theory Comput.* **3**, 2312–2334 (2007).
88. Chen, F. & Smith, P. E. Simulated surface tensions of common water models. *J. Chem. Phys.* **126 22**, 221101 (2007).
89. Kirschner, K. N., Lins, R. D., Maaß, A. & Soares, T. A. A glycam-based force field for simulations of lipopolysaccharide membranes: Parametrization and validation. *J. Chem. Theory Comput.* **8 11**, 4719–31 (2012).
90. Kirschner, K. N. et al. Glycam06: A generalizable biomolecular force field. carbohydrates. *J. Comput. Chem.* **29**, 622–655 (2008).
91. Páll, S., Abraham, M. J., Kutzner, C., Hess, B. & Lindahl, E. Tackling exascale software challenges in molecular dynamics simulations with gromacs. In *Solving Software Challenges for Exascale*, (eds. Markidis, S. & Laure, E.) 3–27 (Springer International Publishing, Cham, 2015).
92. Tomasi, J., Mennucci, B. & Cammi, R. Quantum mechanical continuum solvation models. *Chem. Rev.* **105**, 2999–3094 (2005).
93. Sousa da Silva, A. & Vranken, W. Acpype - antechamber python parser interface. *BMC Res. Notes* **5**, 1–8 (2012).
94. Hess, B. P-lincs: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* **4 1**, 116–22 (2008).
95. Tironi, I. G., Sperb, R. P., Smith, P. E. & van Gunsteren, W. F. A generalized reaction field method for molecular dynamics simulations. *J. Chem. Phys.* **102**, 5451–5459 (1995).

96. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

97. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

98. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).

99. Morris, G. M. et al. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

100. Sennhauser, G., Bukowska, M. A., Briand, C. & Grütter, M. G. Crystal structure of the multidrug exporter mexb from pseudomonas aeruginosa. *J. Mol. Biol.* **389**, 134–145 (2009).

101. Nakashima, R., Sakurai, K. & Yamasaki, Sea Structural basis for the inhibition of bacterial multidrug exporters. *Nature* **500**, 102–106 (2013).

102. Atzori, A. et al. Identification and characterization of carbapenem binding sites within the rnd-transporter acrb. *Biochim. Biophys. Acta Biomembranes* **1861 1**, 62–74 (2019).

Acknowledgements

The authors thank Dr. Olga Lomovskaya and Rempex Pharmaceuticals, and Jurg Dreier and Basilea Pharmaceutica for providing compounds and sharing their structures. This work was supported by NIAID/NIH grants number R01AI136799 and R01AI136795, and by U.S. Air Force Office of Scientific Research awards FA9550-20-1-0382 and FA9550-20-1-0383.

Author contributions

J.K. Walker, P. Ruggerone, V.V. Rybenkov, H.I. Zgurskaya and S. Gnanakaran conceptualized the study. I.V. Leus, J. Mehla and H.I. Zgurskaya carried out the experiments. J.K. Walker, R.K. Kinthada and H.I. Zgurskaya assembled the library of compounds. C.A. López, G. Malloci, S. Gervasoni and A.V. Vargiu calculated the molecular descriptors. P.D. Manrique and L. Herndon curated the data. P.D. Manrique and N.W. Hengartner designed and implemented the statistical model. P.D. Manrique, C.A. López, H.I. Zgurskaya and S. Gnanakaran analyzed the results and wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-024-01161-y.

Correspondence and requests for materials should be addressed to Pedro D. Manrique or S. Gnanakaran.

Peer review information *Communications Chemistry* thanks Fiona Kearns and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.