HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2021 June 22.

Published in final edited form as:

J Chem Inf Model. 2020 June 22; 60(6): 2838–2847. doi:10.1021/acs.jcim.0c00352.

Machine Learning Algorithm Identifies an Antibiotic Vocabulary for Permeating Gram-Negative Bacteria

Rachael A. Mansbach[†], **Inga V. Leus**[‡,§], **Jitender Mehla**[‡,§], **Cesar A. Lopez**[†], **John K. Walker**[¶], **Valentin V. Rybenkov**[‡], **Nicolas W. Hengartner**[†], **Helen I. Zgurskaya**[‡], **S Gnanakaran**[†]

[†]Department of Theoretical Biology and Biophysics, Los Alamos National Lab, MS-K710, P.O. Box 1663, Los Alamos, NM 87545-0001

[‡]Department of Chemistry and Biochemistry, University of Oklahoma, 101 Stephenson Parkway, SLSRC, Rm 1000, Norman, OK, 73019-5251

[¶]Pharmacology and Physiological Science, School of Medicine, Saint Louis University, Schwitalla Hall, Room M362, St. Louis, MO 63104

Abstract

Drug discovery faces a crisis. The industry has used up the “obvious” space in which to find novel drugs for biomedical applications, and productivity is declining. One strategy to combat this is rational approaches to expand the search space without relying on chemical intuition, to avoid rediscovery of similar spaces. In this work, we present proof-of-concept of an approach to rationally identify a “chemical vocabulary” related to a specific drug activity of interest without employing known rules. We focus on the pressing concern of multidrug resistance in *Pseudomonas aeruginosa* by searching for submolecules that promote compound entry into this bacterium. By synergizing theory, computation, and experiment, we validate our approach, explain the molecular mechanism behind identified fragments promoting compound entry, and select candidate compounds from an external library that display good permeation ability.

Graphical Abstract

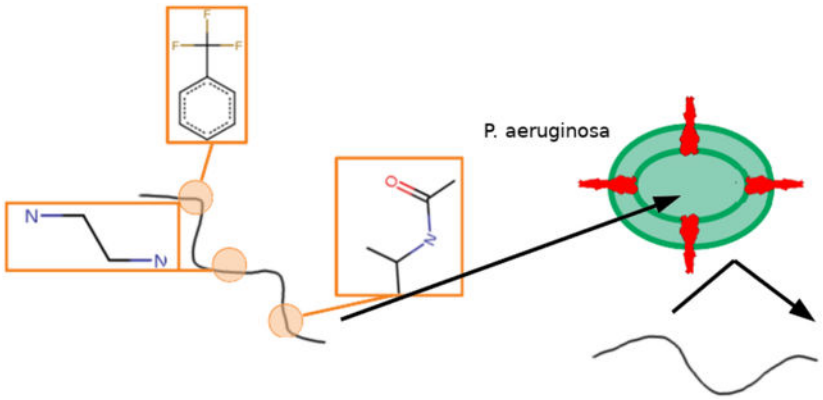
gnana@lanl.gov.

[§]These authors contributed equally to this work

[5]Data and Software Availability

The software and algorithms employed, generated during, and/or analysed during the current study are available from the corresponding author on reasonable request.

Supporting Information Available: We attach as supplementary files the training set of 595 compounds (file anc/training.xlsx), the topologies for the coarse-grained simulations (files in the anc/TOPOLOGIES folder), the full set of potential hits (file anc/potential_hits.xlsx), the set of 48 fragments returned by any run of the Hunting FOX algorithm (file anc/top_fragments.xlsx). We note that we do not report the structures of 46/595 training compounds and 5/463 potential hit compounds due to IP concerns. In addition, we have attached a Supporting Information document containing a detailed description of the software employed for multinomial logistic regression and the theory of multinomial logistic regression, a section on the theory of parallel-tempered well-tempered metadynamics, and a section on future improvements to the algorithm. This document also contains Table S1 and Figure S1, which further illustrate the process of fragment identification, Figure S2, which shows the physicochemical properties of our different datasets, Figure S3, which illustrates class assignment for the MIC ratio response variables, Figure S4, which illustrates the metadynamics setup, Figs S5 and S6, which are ROC curves of different classifiers, and Figures S7 and S8, which are enrichment curves of different classifiers.



1 Background

Gram-negative bacteria are notorious for their ability to evade antibiotic inhibition, partly because of the barrier presented by the highly-impermeable outer membrane (OM); that of the bacterium *Pseudomonas aeruginosa* presents one of the most impenetrable barriers[1–6]. Numerous high-throughput experimental studies have been performed to identify physicochemical properties of good antibiotics[3,7–11], but a lack of holistic understanding of the microscopic mechanisms and methods for improving certain underlying aspects such as drug permeability, particularly in *P. aeruginosa*, is still stimulating development of novel algorithms and studies[12–14]. The problem of designing new drugs to permeate Gram negative bacteria is a microcosm of the problems faced by the drug design industry in general, in which dwindling of the "obvious" chemical space in which to search for new drugs has led to a concomitant dwindling of the drug discovery pipeline[15–18]. It is therefore important to use both well-tested and new ideas in combination; in particular, we must seek ways to move past chemical intuition for drug design[19].

A well-established and primarily experimental technique in the field of drug discovery is "fragment-based drug design" (FBDD), in which libraries of small fragments are screened for activity against a target, followed by fragment growing, linking, or merging to optimize leads. FBDD demonstrates the advantage of a fragment-based approach through shrinking of the chemical space to be searched, as smaller fragments lead to a concomitantly lower number of possible atom combinations[20]. Much computational and experimental effort for FBDD has been focused on the definition of fragments for a fragment library[21]. Several software suites have been developed which seek to design a submolecular library through virtual fragmentation of a series of molecules[22,23]. Virtual similarity search of fragment libraries has allowed the design of GPCR ligands[24]. Training with fragments from large known fragment libraries was used to perform transfer learning with long short-term memory neural networks to generate new drug-like molecules[25]. It has been shown that linear models of molecular activity-a subset of quantitative structure-activity relation (QSAR) models-may also be predictive of the activity of submolecular fragments residing in those molecules[26–31].

Inspired in particular by the recent use of a "pseudolinear" approach to *manually* identify a set of 35 fragments for *in silico* design of heat shot protein 90 inhibitors[32], we develop an algorithm to *automatically* identify a set of relevant fragments for hybrid fragment-based design of molecules with the ability to permeate *P. aeruginosa*. Instead of, as previously done, relying on a linear or pseudolinear molecular model to generalize to the fragments contained in the molecule, we directly employ a fragment-based representation to train a linear model and use it to identify and validate predictive fragments. In specific, we define and exploit a chemical vocabulary–in spirit akin to the "n-grams" employed in natural language processing applications[33] from a set of known drugs with accompanying activity data. We call our algorithm "Hunting FOX" for "Hunting Fragments Of X" and focus on extracting fragments within compounds that can be incorporated into new hybrids to impart the desired experimental activity. Unlike the conventional FBDD approach, which relies on pre-defined fragments, this algorithm considers all possible unique fragments within a set of compounds, from a single bond length in radius about a central atom up to 10 bond lengths in radius. Although the focus of this work is computational, we note that from a medicinal chemistry perspective, such an approach potentially provides a pragmatic way to bring together diverse sub-molecular spaces in a rational ML-directed manner as building blocks[34] for novel hybrid drugs[35].

In this article, we apply Hunting FOX to automatically identify a chemical vocabulary relevant to compound permeation into the Gram-negative bacterium *P. aeruginosa* without any *a priori* chemical intuition. We validate the informational content of this chemical vocabulary through (i) *a posteriori* assessment and comparison with previous studies, and (ii) demonstrating that models trained with the fragment-based description are both *predictive* and *enriching*-i.e., they are able to narrow the search space for new drugs. In addition, using a biased coarse-grained molecular dynamics (CGMD) procedure to study a molecule rich in predictive fragments, we explore the molecular mechanism of permeation and the potential contributions of these fragments. Finally, we experimentally validate the identified vocabulary through showing that models trained on the fragment-based description are capable of identifying good permeators from an external library.

2 Experimental Methods

The overall workflow of the Hunting FOX algorithm is shown in Fig. 1 and consists of four major steps: (i) defining a representation for the compounds (Sec. 2.1); (ii) performing experimental measurements and data cleaning on a select subset of a curated dataset (Sec. 2.2) to set up the drug activity input to the algorithm; (iii) performing feature selection to identify a vocabulary of relevant submolecular fragments (Sec. 2.3); and (iv) fitting a predictive model based on the identified vocabulary (Sec. 2.4).

2.1 Representation of compounds

To define a representation for each compound from which we may extract a chemical vocabulary, we begin with the two-dimensional representation of a molecule as a set of atoms and bonds connecting the atoms. Using a sliding window and considering every atom in the molecule (see Fig. S1 for an example), we identify all fragments consisting of that

atom plus the atoms that lie within k bonds of it for all $1 \leq k \leq 10$ (see Fig. 2). In total, there are 22,139 different fragments comprising the training set of 595 molecules. We represent each molecule M as a $N_f = 22,139$-length vector of frequencies, $\overrightarrow{\ell}(M) = [f(x_1, M), \ldots, f(x_N, M)]$, where every entry is the number of times a particular fragment appears (may be 0), normalized by the number of atoms in molecule M, $L(M)$, such that,

$$f(x_i, M) = n(x_i, M)/L(M), \tag{1}$$

in which $n(x_i, M)$ represents the number of times fragment x_i appears in molecule M. Intuitively, containing larger numbers of relevant fragments should be correlated with increased activity; however, we wanted a metric that was independent of molecular size, thus our division by $L(M)$. Although this is a somewhat naive representation akin to one-hot encoding in traditional ML[36], we employ it as the simplest way to test whether information about permeation ability is contained within the fragments, although greater sophistication could be achieved through the use of more complex representations[37,38].

2.2 Compound activity data

The two major contributers to the impermeability of Gram-negative bacteria in general and *P. aeruginosa* in specific are the OM and the efflux pumps that actively remove molecules from the periplasm and cytoplasm[2,40]. To separate the effects of the efflux pumps from the effects of the OM, we have recently created different mutant strains of Gram-negative bacteria[41]. In this study, we focused on the effects of the OM alone by using two strategically designed mutant strains lacking the effects of efflux. In the first strain, compounds are impeded by the OM barrier, while in the second strain, they are not. Specifically, we studied mutants of the *P. aeruginosa* PAO1 strain. The "PΔ6" mutant is a variant of *P. aeruginosa* in which the genes encoding for the six best characterized efflux pumps have been deleted, which essentially removes the contribution of active efflux in antibacterial activities of antibiotics. It has no other effects; indeed, we have recently shown that there is no significant membrane disorganization introduced by deletions[8,14]. The "Pore" mutant is a variant-not studied in the work-modified to contain large (~2.4 nm in diameter) pores that allow nondiscriminate entry of drugs, which essentially removes the effects of the impermeable outer membrane with no other effects on cell physiology. The "PΔ6-Pore" mutant is a variant combining both previous modifications. In this study, we focus on the PΔ6 and PΔ6-Pore mutants, which both lack efflux pumps. For the drug property input to the algorithm, we experimentally measured the MICs of over 500 compounds exhibiting antibacterial activities in at least one out of the two different mutant strains of *P. aeruginosa* PAO1 (see Sec. 2.2.1 for a complete description of the curated dataset). We then computed the ratio of compound MIC values in the PΔ6-Pore mutant of *P. aeruginosa* PAO1 to their MIC values in the PΔ6 mutant of *P. aeruginosa* PAO1 $\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right)$. When this ratio goes to one, the drug's efficacy is the same whether or not the outer membrane is intact, because the main difference between the strains is whether or not the outer membrane has been hyperporinated. We define a drug possessing this property as being a "good permeator" or saying that it "permeates well." More specifically, we define a

set of five compound classes based on the ratio $\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}$, where $\mu_{P\Delta6\text{-Pore}}$ is the MIC of the compound in the *P*Δ6-Pore mutant and $\mu_{P\Delta6}$ is the MIC of the compound in the *P*Δ6 mutant. If $\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}} < 0.2$, $\text{Class}\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right) = 0$ ("non-permeators"); if $0.2 \leq \frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}} < 0.4$, $\text{Class}\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right) = 1$; if $0.4 \leq \frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}} < 0.6$, $\text{Class}\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right) = 2$; if $0.6 \leq \frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}} < 0.8$, $\text{Class}\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right) = 3$; and if $\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}} > 0.8$, $\text{Class}\left(\frac{\mu_{P\Delta6-\text{Pore}}}{\mu_{P\Delta6}}\right) = 4$ ("good permeators"). The class breakdown is as follows: ≈ 48% of MIC ratios fall into class 0, 10% into class 1, 9% into class 2, 10% into class 4, and 22% into class 4.

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. Minimum inhibitory concentration (MIC) determination was carried out using the 2-fold broth dilution method as described previously[8]. Two independent experiments were carried out, giving precision =/-two fold. The expression of the Pore was induced at OD_{600} ~0.3–0.4 by addition of 0.1 mM IPTG.

In practice, due to experimental limitations, the data must be cleaned. We set any ratios wherein the denominator was too high to measure to zero, assuming that such molecules will not contain substantial desirable fragments. Due to the 2-fold error in the MIC measurements, it was possible to compute ratios greater than one, which were set to one to avoid inclusion of erroneous information. Finally, certain measurements were given as ranges of MIC values and, in each of those cases, we chose the mean of the range.

2.2.1 Datasets—Because one of the benefits of ML algorithms such as Hunting FOX is their ability to narrow the search space for experiment, we initially curated a rather large database of 31,524 molecules of interest, which were organized with the CDD Vault from Collaborative Drug Discovery (Burlingame, CA. www.collaborativedrug.com)[39]. The bulk of compounds are from a commercially available library with diverse chemical structures (Chem-Bridge 30,000 Diversity Library) and others are known antibiotics and compound series generated by medicinal chemistry campaigns to optimize efflux pump inhibitors (EPIs)[42,43]. Of a representative set with known or predicted antibacterial activities, we selected a subset that could be reasonably screened in two strategically designed mutant strains of *P. aeruginosa* for growth inhibitory activities. From initial experimental screening, we further subsampled to identify 595 molecules (cf. Supplementary File anc/training.xlsx), for which MICs in both the *P*Δ6 and *P*Δ6-Pore mutants were measurable. In Fig. S2, we show several representative physicochemical characteristics (molecular weight, pKa, and logD), to emphasize that the chosen 595 are a reasonable microcosm of the full library in a broad sense. We do note that the test group is enriched with compounds that are known for their efflux pump inhibitory activities. Although we do not study the effects of efflux pumps, these EPIs display growth inhibitory activity as well even in mutants without efflux pumps. In addition, in order to reach the pumps they inhibit, it is necessary for EPIs, like more traditional antibiotics, to permeate the OM.

We employed the 595 as training/validation molecules for the Hunting FOX algorithm. In specific, on each iteration of the algorithm, we split the 595 molecules into five disjoint random groups with the same class distribution as the overall 595, and we used each of these five groups as the reserved validation set for a model trained on the remaining 4/5 (5-fold cross validation). Additionally, we used the remaining 30,929 on which we have no data concerning their ability to permeate *P. aeruginosa* for external experimental testing to show that Hunting FOX is capable of narrowing the search space of a large, chemically diverse library and identifying good molecules for experimental testing.

2.3 Feature selection

The feature selection portion of the Hunting FOX algorithm consists of two steps: (i) permissive LASSO regularization to eliminate non-predictive variables and (ii) hierarchical cleansing to eliminate remaining redundancies. First, we split the data into five disjoint subsets and fit a sparse multinomial classifier with the LASSO penalty[44] five times, each time reserving one of the five subsets for testing (5-fold cross-validation). We choose to employ multinomial classification due to the natural spread of the data (cf. Fig. S3), and we employ the LASSO penalty as a simple, well-verified way of discarding non-predictive features in the data. Step (i) results in five (overlapping) sets of relevant fragments with associated coefficients from the multinomial fit (see Sec. S1.1) for software details).

From the set of fragments that result in nonzero coefficients in the model, we retain only fragments that have positive coefficients for prediction of class 4 occupancy-that is, fragments whose existence imply a compound will have a high permeating ability-or negative coefficients for prediction of class 0 occupancy-that is, fragments whose existence will decrease the probability of being in the least-permeating class and therefore may be predictive of non-zero permeation. Due to class 0's higher heterogeneity, no fragments were found with positive coefficients for prediction of class 0 occupancy, so we are not able to present a set of fragments to avoid in addition to a set of fragments to include. We do not retain fragments with coefficients solely pertaining to the middle three classes because the interpretation of such fragments is less clear. We further sparsify the retained fragments by finding a subset that are not hierarchically related (step ii).

Our procedure in step ii is as follows: first, consider only fragments with non-zero coefficients in at least one classifier. Then, find the subset $\{\mathscr{H}\}$ that are hierarchically related within all classifiers, meaning that for any given fragment $x_i \in \{\mathscr{H}\}$ appearing in one classifier, either that same fragment or a fragment that it contains or that it is contained by appears in every other classifier. Next, consider fragments in order of increasing coefficient magnitude, where we consider the maximum coefficient magnitude of all the coefficients in all classifiers in which the fragment appears. Each time such a fragment is part of a hierarchical relationship with another fragment, remove it from consideration. Continue doing so until all that remains are fragments that are not hierarchically related to one another. This whole procedure preferentially retains fragments that are ranked as important to at least one classifier and also ensures that all retained fragments contain information present in all classifiers.

2.4 Non-sparse regression model

Having identified a set of likely active fragments, we train a new set of non-sparse classifiers, using only the non-hierarchically-related subset, that may be used to identify molecules from an external library. We employ the same train/test split as before to avoid cross-contamination and fit five multinomial classifiers with a ridge penalty. We run this set of five classifiers on an external testing set; any molecule that is predicted to have an MIC ratio $\frac{\mu PA6 - \text{Pore}}{\mu PA6} > 0.8$ by all five classifiers is returned as a hit by this run of the Hunting FOX algorithm. We employ five different classifiers to leverage the power of ensemble methods[45].

2.5 Coarse-grained molecular dynamics

To study the possible molecular contributions of the fragments, we performed biased coarse-grained molecular dynamics simulations employing the MARTINI force field to calculate the two-dimensional potentials of mean force of five molecules crossing a membrane mimic. The molecules we studied were amoxicillin, difloxacin, sarafloxacin, and two molecules from our curated database we refer to as OU-315 and OU-314 (cf. Fig. 4a–b). The parametrization of all molecules studied follows the general MARTINI philosophy, which consists of reproducing the partition coefficient between an organic solvent and water. In our case, the parameters of all drugs were tuned in order to match predicted octanol-water LogP values. (We have attached the itp files as part of the Supporting Material in the folder anc/TOPOLOGIES). In order to retain internal dynamics at the coarse-grained (CG) level, we have employed atomistic generated data using the General Amber force field (GAFF) with charges obtained using the RESP approach[46]. Such simulations were later used to incorporate the necessary bonded terms into the CG geometry as previously performed for similar molecules[47].

The simulated system consisted of a charge-neutralized outer membrane patch solvated with the polarizable water version of the MARTINI force field. We represented the OM using inhouse developed parameters[48], based on the GROMOS 53a6 GLYC. The outer leaflet is composed of a homogenous mixture of pure LPS derived from the *P. aeruginosa* PAO1 BandA strain, which was neutralized using CA++ cations. The inner leaflet is composed purely of DPPE lipids. Although it is possible in general for other lipids to be present in small amounts, this particular composition allows rapid equilibration and has previously been shown to be a good mimic of the *P. aeruginosa* outer membrane[49]. The system was simulated under constant ionic concentration of 150mM Na-Cl. The final membrane consisted of 12 LPS molecules and 36 DPPE lipids in the inner leaflet. Each compound was placed separately in the water-membrane interface and equilibrated at 310 K under semi-isotropic pressure coupling using a velocity-rescaling thermostat[50] and a Berendsen barostat[51] respectively for 1 μs before production simulations.

All simulations were carried out with GROMACS 5.1.2[52], compiled with the PLUMED v 2.5 software[53] in order to compute the membrane translocation free energy. Simulations used a 25 fs time-step. Particle Mesh Ewald electrostatics were used with a Coulomb cut-off of 1.1 nm and dielectric constant adjusted to $\epsilon = 2.5$ in order to maintain consistency with

the MARTINI polarizable water model[54]. A cut-off of 1.1 nm was used for calculating Lennard Jones interactions, and the Van der Waals potential was shifted to zero at the cut-off. The OM leaflets were coupled to a constant thermal bath maintained at 310 K using a velocity-rescaling thermostat[50] with a relaxation time of 1.0 ps. Compounds and solvents were coupled separately, due to the needs of the biasing algorithm (see below).

The free energy for membrane translocation was computed using parallel well-tempered metadynamics (PtWtMET)[55,56] (see Sec. S3 for details), requiring a set of four coupled different temperatures. For each drug, four simulations running at 310, 410, 610 and 1010 K were necessary in order to properly converge and overcome the energetic barriers. Production runs were adjusted to one of the coupling temperatures using a NVT ensemble. In order to prevent membrane disruption at higher temperatures, the phosphates of both the Lipid-A region and the DPPE inner leaflet lipids were position-restrained along the δ vector after equilibration.

In our free energy calculations, we chose to bias two specific reaction coordinates: (i) the center of mass (COM) of the drug with respect to the COM of the membrane and (ii) an angle defining the relative orientation of the drug with respect to the membrane (see Fig. S4). Specifically, in the case of amoxicillin, we bias the angle formed between the carbonyl group, the hydroxyl group in the aromatic ring and the COM of the membrane. For both fluoroquinolones, we bias the angle formed between the carbonyl, the distal nitrogen in the piperazine group and the COM of the membrane. Lastly, for OU-315 and OU-314, we bias the angle formed by the nitrogen in the leucoline ring, the central nitrogen in the distal di-amine groups and the COM of the membrane. In all cases, sigma values of 0.1 nm and 0.35 rads were applied to the COM-COM distance and the angle respectively. We find that a bias factor of k = 15 kJ/mol was enough to converge the simulations. We find that a combination of a CG model with the PtWtMET approach allows good convergence at ranges between 12–15 μs of MARTINI time scales.

3 Results and Discussion

In designing the Hunting FOX algorithm and the overall approach described in this article, our objectives were (i) to develop a vocabulary of submolecular fragments, similar to n-grams in natural language processing[33] that predict drug efficacy in permeating the outer membrane, and (ii) to validate our approach theoretically and experimentally. In the following section, we demonstrate the identification of promising submolecular fragments that specifically track the drug activity. We show a biochemical rationalization for why such fragments would be identified and then perform coarse-grained metadynamics to assess permeation of two compounds containing a relatively high number of important fragments. Finally, we theoretically and experimentally validate that classifiers trained on the relevant fragments possess significant predictive power and the ability to identify hits from an external library of drug compounds not contained in the original data set.

3.1 Hunting FOX discovers active submolecular fragments

We run the Hunting FOX algorithm twenty-eight times starting from a different random seed each time, which corresponds to a different disjoint split for the 5-fold testing data, and each

time identify a set of important molecular fragments (our “chemical vocabulary”), which we employ to train a second set of non-sparse classifiers. We believe this number of repetitions is sufficient to observe the statistics of the algorithm’s behavior. In Fig. 3, we report the number of fragments out of forty-eight that appear between one and twenty-eight times, while in Supporting File anc/top_fragments.xlsx we report all identified fragments and their frequency of appearance. We note a decent level of robustness in the algorithm: although no fragments are reported by every run, one out of forty-eight does appear in twenty-six runs, and twenty-nine fragments appear in more than one run. Although the lack of more fragments appearing repeatedly indicates some dependence on the training data, considering every fragment separately allows for a clear interpretation of the importance of each fragment individually. It might be possible to ameliorate this dependence by employing a representation that considers fragment similarity explicitly (cf. Sec. S4 in the Supporting Information); however such work is outside the scope of the current study.

There are several notable chemical features of the fragments that provide *a posteriori* empirical support for our procedure. First, twenty-six of forty-eight of all fragments and six of nine fragments appearing more than five times contain part of a benzene ring or one or more whole benzene rings. Such fragments have recently been demonstrated to improve permeability[12], and this result also highlights that the predictive algorithm is able to discriminate 3-dimensional features that are not directly employed in our code. In fact, the presence of rigid benzene rings dramatically improves both rigidity and globularity, important for membrane translocation[12]. Second, a large number of the fragments identified contain primary and secondary amine groups. Such groups are expected to strongly interact with both the 2-Keto-3-deoxy-octonate (KDO) and lipid-A regions of the lipopolysaccharide (LPS), functioning as specific anchors for highly anionic membranes (e.g. bacterial OM)[57]. Indeed, there is external evidence that including such groups does improve intracellular accumulation[12] and specificity for Gram-negative bacteria[58]. Third, we note the appearance of a trifluoromethyl fragment (Fig. 3, molecule at $a_{\mathrm{freq}} = 9$) in nine out of twenty-eight runs. Nowadays, fluorine containing compounds are synthesized in pharmaceutical research on a routine basis and about 10 percent of all marketed drugs contain a fluorine atom. The major rationale is that the presence of fluorine atoms in biologically active molecules can enhance their lipophilicity and thus their uptake and transport. In particular, the trifluoromethyl group (−CF3) confers increased stability and lipophilicity in addition to its high electronegativity. This enrichment with trifluoromethyl-containing fragments is somewhat to be expected, as our initial dataset contained a not-insubstantial number of compounds with this functional group. This provides an additional check of our algorithmic approach, and also points to a potential avenue for extension. It is remarkable that we were able to capture such behaviors without leveraging chemical expertise in the initial steps.

3.2 Molecular Dynamics simulations provide mechanistic rationale of fragment identification

We choose two representative molecules (OU-315, with an experimental MIC ratio of $\frac{\mu P\Delta 6 - \mathrm{Pore}}{\mu P\Delta 6}$ ~0.25 – 0.5, and OU-314, with an experimental MIC ratio of $\frac{\mu P\Delta 6 - \mathrm{Pore}}{\mu P\Delta 6}$ ~1.0) from the training set that summarize the different identified relevant fragments and probe

their mechanism of permeation in a relatively high-resolution manner to investigate possible molecular contributions of the fragments selected by the Hunting FOX algorithm in more than five iterations. In Fig. 4a–b we show the candidate molecules as well as highlighted fragments from the set of nine that were identified by the Hunting FOX algorithm. By using a coarse-grained two-dimensional metadynamics calculation in a previously-described membrane model[59], we compute a potential of mean force (PMF) that suggests mechanistic details by which the fragments enhance the permeation of the selected candidate.

In addition, we choose three previously studied molecules with a range of MIC ratios for comparison: amoxicillin (MIC ratio $\frac{\mu_{P\Delta 6-\text{Pore}}}{\mu_{P\Delta 6}}$ ~0.015625), difloxacin, and sarafloxacin (MIC ratios $\frac{\mu_{P\Delta 6-\text{Pore}}}{\mu_{P\Delta 6}}$ ~0.25)[14]. Of these molecules, difloxacin contains a single one of the top nine fragments, whereas amoxicillin and sarafloxacin contain none. Amoxicillin exemplifies a very polar drug, while both difloxacin and sarafloxacin are fluoroquinolones containing aromatic rings in combination with halogen groups.

A one-dimensional projection of the permeation free energy of the studied compounds is provided in Fig. 4e. Comparison of OU-315 and OU-314 to the other three compounds illustrates two noteworthy features: i) a favorable energy basin in both the core region of the LPS and in the phosphate groups in the 1,2-Dipalmitoyl-sn-glycero-3-phosphoethanolamine (DPPE) lipids of the inner leaflet, and ii) the absence of an energy barrier within the hydrophobic region of the membrane: indeed, the PMF is approximately flat all the way from the outer leaflet to the inner leaflet. The first feature leads to the conclusion that the two compounds containing many favorable fragments are attracted by the anionic regions of the membrane, providing a direct advantage in terms of partitioning within the LPS leaflet. The second feature, the lack of a second energy well within the hydrophobic core of the membrane compared to the three compounds not containing many favorable fragments, provides a potential advantage in terms of permeation by reducing the number of barriers the drug must navigate. We note in addition the lack of these features in three non-permeating control molecules containing few or none of the chemical vocabulary.

The traced approximate lowest free energy paths (red lines) of OU-315 and OU-314 on two-dimensional PMFs including orientation as the second variable (Fig. 4c–d), demonstrate weak zigzag patterns, which suggests that these compounds bypass the aliphatic region via an orientational rearrangement, somewhat resembling the flip-flop mechanism of lipids, and that this mechanism occurs, to within thermal fluctuations, at constant free energy. A plausible explanation for this observed property in the PMF is supported by visual inspection of the simulation of OU-315 (Fig. 4f), which shows water molecules forming a solvation shell within the proximity of the drug, which is expected to overall reduce the penalty of translocating charged groups across the aliphatic region.

3.3 Theoretical and experimental verification of Hunting FOX predictions

We now validate our algorithm theoretically and experimentally, through showing that it is capable of producing well-performing classifiers that can identify novel molecules with desired properties from a library of molecules with unknown properties using only its

identified predictive fragments. We train a set of non-sparse classifiers (cf. Sec. 2.4) over 28 iterations of the Hunting FOX algorithm and employ them to make predictions on an external library.

3.3.1 Classifier performance—We assessed the performance of the non-sparse classifiers through the twin metrics of receiver operating characteristic (ROC) curves and enrichment curves on the accompanying test set. The ROC curve is an illustration of the true positive rate, *TPR*, versus the false positive rate, *FPR*. For a classifier whose performance is no better than random, the plot should lie along the line of equivalence, $TPR = FPR$. For a perfect classifier, the *TPR* would immediately rise to one. We define "enrichment," $\mathscr{E}$, of a class at a level m as,

$$\mathscr{E}(m) \equiv \frac{\frac{1}{m}\sum_{j=1}^{m} y_j^* - \frac{1}{N}\sum_{j=1}^{N} y_j^*}{\frac{1}{N}\sum_{j=1}^{N} y_j^*} \tag{2}$$

where $y_j^* \in [0, 1]$ is the list of true occupancies for the samples in a class reordered in order of decreasing probability as predicted by the classifier, N is the total number of samples, and $\frac{1}{N}\sum_{j=1}^{N} y_j^*$ is the total fraction of samples that belong to the class. The enrichment at a level m measures the difference between the percentage of hits identified by the classifier from the percentage of hits that would be found by random chance.

In Tables 1–2 and Figs S5 and S7, we report the average performance of five nonsparse classifiers per run trained in 28 iterations of Hunting FOX, where each set of five is trained on a different random stratified disjoint train/test split (see Secs. 2.2.1 and 2.3). We linearly interpolate the separate curves to assess performance at the same points and report the average and standard deviation of each set of 5 classifiers. To assess the predictivity of the best fragments, in Tables 1–2 and Figs S6 and S8, we also report the performance of classifiers trained on the same train/test split but employing only the top nine fragments identified by all 28 runs (cf. Fig. 3).

All classifiers perform well, which provides validation for our hypothesis that information is contained simply in the fragment composition of the molecules. Notably, employing only the top nine fragments results in a performance improvement in every metric other than enrichment of class 2, which demonstrates identical performance within error bars. Thus, we demonstrate that these are indeed the fragments to focus on for synthesizing new hybrids with greater ability to permeate *P. aeruginosa*.

In terms of specifics, all classifiers perform significantly better than random on all performance metrics, except for AUC for Class 1, which is quite poor. However, since we remove all fragments not pertaining directly to classes 0 and 4 during the feature selection step (cf. Sec. 2.3), it is encouraging that we are also able to train predictive classifiers for classes 2 and 3. The best performance in terms of ROCs is class 0 (non-permeating), with an AUC score of almost 90%, probably due to imbalance in the training data (almost 50% of training examples in class 0), but aside from the poor performance of class one, all AUC scores are at least 75% when trained on the top nine fragments and at least 73% when

trained on the top fragments returned by a single Hunting FOX run. The enrichment of class 4, which is of particular interest since it corresponds to the best permeators, is on average nearly 150% for classifiers trained only on the top nine fragments, while the maximum enrichment of any single iteration is about 270%; it is nearly 110% for classifiers trained on the fragments returned by one run, while the maximum of any single iteration is about 160%. In addition, the on-average monotonic decrease of the enrichment with m demonstrates that the probability rankings are sensible. Overall, the enrichment indicates that by taking the top ten percent most probable molecules as predicted by any given classifier, one should be able to on average find 2–2.5× as many hits as if one were to select compounds at random, which represents a significant shrinkage of the search space.

3.3.2 Experimental Validation—One of the potential strengths of an algorithm like Hunting FOX is its ability to narrow the search space for experiment by identifying and ranking possible compounds of interest. In order to both perform an experimental validation and demonstrate that this is indeed a strength of the algorithm, we employed a library of 30,929 molecules comprising efflux substrates, efflux inhibitors, outer membrane permeators and non-permeators and known antibiotics that were not part of the training or testing set for any of the regression models and used the non-sparse regression models of 28 repeated randomized iterations of the Hunting FOX algorithm to identify which of these molecules might be expected to display outer membrane permeation properties as measured by the MIC ratios $\frac{\mu_{P\Delta 6-\text{Pore}}}{\mu_{P\Delta 6}}$. In each iteration, we reported as potential "hits" those molecules that were predicted by all five non-sparse classifiers to have MIC ratios of $\frac{\mu_{P\Delta 6-\text{Pore}}}{\mu_{P\Delta 6}} > 0.8$ (class 4).

Nine compounds were identified as potential permeators by at least 50% of the repeated runs of the Hunting FOX algorithm (Fig. 5). (We report all 463 compounds identified by any run in the Supporting File anc/potential_hits.xlsx.) Among the top nine identified compounds, five possessed antibacterial activities and their permeation properties could be assessed by measuring MICs in PΔ6 and PΔ6-Pore strains. For all these compounds, the ratio of MICs $\frac{\mu_{P\Delta 6-\text{Pore}}}{\mu_{P\Delta 6}} \approx 1.0$, indicating that they are good permeators. Compounds OU-572 and OU-559 were unavailable, but belong to the same structural series as OU-457 and OU-466, sharing with them certain structural fragments, and are likely to have similar properties. The remaining compounds OU-1729 and OU-2015 are not expected to display measurable MICs and their permeation could not be assessed using growth inhibition assays. However, the fact that all the drugs that were assessable in this manner were hits demonstrates the use of the algorithm in narrowing experimental search space as well as more broadly providing experimental validation of our models and of our algorithm.

3.4 Conclusions

In this article, we provide proof of concept of the Hunting FOX algorithm, which combines traditional machine learning approaches with a chemical vocabulary-based molecular description inspired by natural language n-grams and FBDD. Despite employing no *a priori* expert input, it identifies the chemical submolecules that impart compounds with desirable

properties and identifies existing drugs with those properties from an external library whose size makes it experimentally intractable, leading to a unique avenue for rational hybrid drug design and drug reuse. Specifically, we have employed our algorithm to identify a set of fragments expected to confer on drugs the ability to permeate the OM of *P. aeruginosa*, as well as nine compounds expected to be good OM permeators, of which thus far five have been directly experimentally validated. We have also used biased MD simulations to determine the mechanism of permeation of two molecules containing many of the top reported fragments, thus uncovering their molecular-level importance. Hunting FOX opens new portions of chemical space through new combinations of fragments, and also serves as useful tool to generate novel low-molecular weight fragments governing specific target or activity in fragment-based drug discovery. As demonstrated here, this work represents an important step forward for rational hybrid drug design, particularly for antibiotics against Gram-negative bacteria. This simple chemical vocabulary-based methodology generalizes easily to any response variable of interest: for example, one might classify drugs based on their ability to inhibit the growth of cancerous cells compared to that of normal cells and thus identify the reusable chemical vocabulary pertaining to preferential tumor-binding in a rational manner. In future, we plan to employ synthetic biology techniques to design unique hybrid antibiotics from a medicinal chemistry perspective based on the fragments identified herein.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by NIAID/NIH grant number R01AI136799. RAM acknowledges a Los Alamos Director's Postdoctoral Fellowship. This research used resources provided by the Los Alamos National Laboratory Institutional Computing Program, which is supported by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001. Triad National Security, LLC (Los Alamos, NM, USA) operator of the Los Alamos National Laboratory under Contract No. 89233218CNA000001 with the U.S. Department of Energy. We thank Olga Lomovskaya, Qpex Biopharma for providing the Rempex compounds. We thank Paolo Ruggerone and Giuliano Malloci for providing Amber topologies of selected compounds. We thank Liam Herndon for discussions. We thank Illia S. Affanasiev for technical assistance. We thank Dr. Keith Haynes and Dr. Napoleon D'Cunha for synthesis of certain compounds used in the analysis.

References

(1). Arzanlou M; Chai WC; Venter H Intrinsic, Adaptive and Acquired Antimicrobial Resistance in Gram-Negative Bacteria. Essays Biochem. 2017, 61, 49–59. [PubMed: 28258229]

(2). Zgurskaya HI; López CA; Gnanakaran S Permeability Barrier of Gram-Negative Cell Envelopes and Approaches To Bypass It. ACS Infect. Dis 2015, 1, 512–522. [PubMed: 26925460]

(3). Richter MF; Hergenrother PJ The Challenge of Converting Gram-Positive-Only Compounds Into Broad-Spectrum Antibiotics. Ann. N. Y. Acad. Sci 2018, 1435, 18–38. [PubMed: 29446459]

(4). Schuster S; Vavra M; Schweigger TM; Rossen JWA; Matsumura Y; Kern WV Contribution of AcrAB-TolC to Multidrug Resistance in an Escherichia Coli Sequence Type 131 Isolate. Int. J. Antimicrob. Agents 2017, 50, 477–481. [PubMed: 28689875]

(5). Zgurskaya HI; Rybenkov VV; Krishnamoorthy G; Leus IV Trans-Envelope Multidrug Efflux Pumps of Gram-Negative Bacteria and Their Synergism with the Outer Membrane Barrier. Res. Microbiol 2018, 169, 351–356. [PubMed: 29454787]

(6). Cama J; Henney AM; Winterhalter M Breaching the Barrier: Quantifying Antibiotic Permeability across Gram-Negative Bacterial Membranes. J. Mol. Biol 2019, In press.

(7). Brown DG; May-Dracka TL; Gagnon MM; Tommasi R Trends and Exceptions of Physical Properties on Antibacterial Activity for Gram-Positive and Gram-Negative Pathogens. J. Med. Chem 2014, 57, 10144–10161. [PubMed: 25402200]

(8). Krishnamoorthy G; Leus IV; Weeks JW; Wolloscheck D; Rybenkov VV; Zgurskaya HI Synergy between Active Efflux and Outer Membrane Diffusion Defines Rules of Antibiotic Permeation into Gram-Negative Bacteria. mBio 2017, 8, 01172–17.

(9). Graef F; Vukosavljevic B; Michel J-P; Wirth M; Ries O; De Rossi C; Windbergs M; Rosilio V; Ducho C; Gordon S; Lehr C-M The Bacterial Cell Envelope as Delimiter of Anti-Infective Bioavailability – An in Vitro Permeation Model of the Gram-Negative Bacterial Inner Membrane. J. Controlled Release 2016, 243, 214–224.

(10). Pawlowski AC; Johnson JW; Wright GD Evolving Medicinal Chemistry Strategies in Antibiotic Discovery. Curr. Opin. Biotechnol 2016, 42, 108–117. [PubMed: 27116217]

(11). Silver LL A Gestalt Approach to Gram-negative Entry. Bioorg. Med. Chem 2016, 24, 6379–6389. [PubMed: 27381365]

(12). Richter MF; Drown BS; Riley AP; Garcia A; Shirai T; Svec RL; Hergenrother PJ Predictive Compound Accumulation Rules Yield a Broad-Spectrum Antibiotic. Nature 2017, 545, 299–304. [PubMed: 28489819]

(13). Ivanenkov YA et al. Identification of Novel Antibacterials Using Machine Learning Techniques. Front. Pharmacol 2019, 10, 913. [PubMed: 31507413]

(14). Cooper SJ; Krishnamoorthy G; Wolloscheck D; Walker JK; Rybenkov VV; Parks JM; Zgurskaya HI Molecular Properties That Define the Activities of Antibiotics in Escherichia coli and Pseudomonas aeruginosa. ACS Infect. Dis 2018, 4, 1223–1234. [PubMed: 29756762]

(15). Griffen EJ; Dossetter AG; Leach AG; Montague S Can We Accelerate Medicinal Chemistry by Augmenting the Chemist with Big Data and Artificial Intelligence? Drug Discovery Today 2018, 23, 1373–1384. [PubMed: 29577971]

(16). Pammolli F; Magazzini L; Riccaboni M The Productivity Crisis in Pharmaceutical R&D. Nat. Rev. Drug Discovery 2011, 10, 428–438. [PubMed: 21629293]

(17). Sedo K; Kararli T GLOBAL REPORT – 2017 Global Drug Delivery & Formulation Report. Drug Dev. Delivery 2018, 1–4.

(18). Hay M; Thomas DW; Craighead JL; Economides C; Rosenthal J Clinical Development Success Rates for Investigational Drugs. Nat. Biotechnol 2014, 32, 40–51. [PubMed: 24406927]

(19). Lu W; Xiao R; Yang J; Zhang W Data Mining-Aided Materials Discovery and Optimization. J. Materiomics 2017, 3, 191–201.

(20). Erlanson DA; Fesik SW; Hubbard RE; Jahnke W; Jhoti H Twenty Years On: the Impact of Fragments on Drug Discovery. Nat. Rev. Drug Discovery 2016, 15, 605–619. [PubMed: 27417849]

(21). Osborne J; Panova S; Rapti M; Urushima T; Jhoti H Fragments: Where Are We Now? Biochem. Soc. Trans 2020, 48, 271–280. [PubMed: 31985743]

(22). Liu T; Naderi M; Alvin C; Mukhopadhyay S; Brylinski M Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. J. Chem. Inf. Model 2017, 57, 627–631. [PubMed: 28346786]

(23). Gunera J; Kolb P Fragment-Based Similarity Searching with Infinite Color Space. J. Comput. Chem 2015, 36, 1597–1608. [PubMed: 26119231]

(24). Li Y; Sun Y; Song Y; Dai D; Zhao Z; Zhang Q; Zhong W; Hu LA; Ma Y; Li X; Wang R A Fragment-Based Computational Method for Designing GPCR Ligands. J. Chem. Inf. Model 2019, acs.jcim.9b00699.

(25). Awale M; Sirockin F; Stiefl N; Reymond J-L Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. J. Chem. Inf. Model 2019, 59, 1347–1356. [PubMed: 30908913]

(26). Speck-Planche A; Dias Soeiro Cordeiro MN Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. ACS Comb. Sci 2017, 19, 501–512. [PubMed: 28437091]

(27). Prado-Prado FJ; García-Mera X; González-Díaz H Multi-target Spectral Moment QSAR Versus ANN for Antiparasitic Drugs Against Different Parasite Species. Bioorg. Med. Chem 2010, 18, 2225–2231. [PubMed: 20185316]

(28). Speck-Planche A; Cordeiro MNDS Chemoinformatics for Medicinal Chemistry: In Silico Model to Enable the Discovery of Potent and Safer Anti-Cocci Agents. Future Med. Chem 2014, 6, 2013–2028. [PubMed: 25531966]

(29). Speck-Planche A; Cordeiro MND Fragment-based in Silico Modeling of Multi-Target Inhibitors Against Breast Cancer-Related Proteins. Mol. Diversity 2017, 21, 511–523.

(30). Speck-Planche A; Cordeiro MND De Novo Computational Design of Compounds Virtually Displaying Potent Antibacterial Activity and Desirable in Vitro ADMET Profiles. Med. Chem. Res 2017, 26, 2345–2356.

(31). Kleandrova VV; Ruso JM; Speck-Planche A; Dias Soeiro Cordeiro MN Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. ACS Comb. Sci 2016, 18, 490–498. [PubMed: 27280735]

(32). Speck-Planche A Combining Ensemble Learning with a Fragment-Based Topological Approach To Generate New Mol. Diversity in Drug Discovery: In Silico Design of Hsp90 Inhibitors. ACS Omega 2018, 3, 14704–14716. [PubMed: 30555986]

(33). Broder AZ; Glassman SC; Manasse MS; Zweig G Syntactic Clustering of the Web. Comput. Networks ISDN 1997, 29, 1157–1166.

(34). Goldberg FW; Kettle JG; Kogej T; Perry MW; Tomkinson NP Designing Novel Building Blocks is an Overlooked Strategy to Improve Compound Quality. Drug Discovery Today 2015, 20, 11–17. [PubMed: 25281855]

(35). Domalaon R; Idowu T; Zhanel GG; Schweizer F Antibiotic Hybrids: the Next Generation of Agents and Adjuvants against Gram-Negative Pathogens? Clin. Microbiol. Rev 2018, 31, 00077–17.

(36). Gori M Machine Learning: A Constraint-Based Approach; Elsevier: Cambridge, MA, 2018.

(37). Mikolov T; Chen K; Corrado G; Dean J Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 2013, 1–12.

(38). Rodríguez P; Bautista MA; Gonzàlez J; Escalera S Beyond One-Hot Encoding: Lower Dimensional Target Embedding. Image Vision Comput. 2018, 75, 21–31.

(39). Hohman M; Gregory K; Chibale K; Smith PJ; Ekins S; Bunin B Novel Web-Based Tools Combining Chemistry Informatics, Biology and Social Networks for Drug Discovery. Drug Discovery Today 2009, 14, 261–270. [PubMed: 19231313]

(40). Zgurskaya HI; Rybenkov VV Permeability Barriers of Gram-Negative Pathogens. Ann. N. Y. Acad. Sci 2019,

(41). Krishnamoorthy G; Wolloscheck D; Weeks JW; Croft C; Rybenkov VV; Zgurskaya HI Breaking the Permeability Barrier of Escherichia coli by Controlled Hyperporination of the Outer Membrane. Antimicrob. Agents Chemother 2016, 60, 7372–7381. [PubMed: 27697764]

(42). Haynes KM; Abdali N; Jhawar V; Zgurskaya HI; Parks JM; Green AT; Baudry J; Rybenkov VV; Smith JC; Walker JK Identification and Structure-Activity Relationships of Novel Compounds that Potentiate the Activities of Antibiotics in Escherichia coli. J. Med. Chem 2017, 60, 6205–6219. [PubMed: 28650638]

(43). Renau TE et al. Conformationally-Restricted Analogues of Efflux Pump Inhibitors that Potentiate the Activity of Levofloxacin in Pseudomonas Aeruginosa. Bioorg. Med. Chem. Lett 2003, 13, 2755–2758. [PubMed: 12873508]

(44). Tibshirani R Regression Shrinkage and Selection Via the Lasso. J. Royal Stat. Soc.: Series B (Methodological) 1996, 58, 267–288.

(45). Dietterich TG Ensemble Methods in Machine Learning; Springer, Berlin, Heidelberg, 2000; pp 1–15.

(46). Bayly CI; Cieplak P; Cornell W; Kollman PA A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. J. Phys. Chem 1993, 97, 10269–10280.

(47). Marrink SJ; Risselada HJ; Yefimov S; Tieleman DP; de Vries AH The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. J. Phys. Chem. B 2007, 111, 7812–24. [PubMed: 17569554]

(48). López CA; Zgurskaya H; Gnanakaran S Molecular Characterization of the Outer Membrane of Pseudomonas Aeruginosa. Biochim. Biophys. Acta, Biomembr 2020, 1862, 183151. [PubMed: 31846648]

(49). Lins RD; Straatsma TP Computer simulation of the rough lipopolysaccharide membrane of Pseudomonas aeruginosa. Biophys. J 2001, 81, 1037–1046. [PubMed: 11463645]

(50). Bussi G; Donadio D; Parrinello M Canonical Sampling through Velocity Rescaling. J. Chem. Phys 2007, 126, 014101. [PubMed: 17212484]

(51). Berendsen HJC; Postma JPM; van Gunsteren WF; DiNola A; Haak JR Molecular Dynamics with Coupling to an External Bath. J. Chem. Phys 1984, 81, 3684–3690.

(52). Páll S; Abraham MJ; Kutzner C; Hess B; Lindahl E Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS; Springer, Cham, 2015; pp 3–27.

(53). Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. Nat. Methods 2019, 16, 670–673. [PubMed: 31363226]

(54). Yesylevskyy SO; Schäfer LV; Sengupta D; Marrink SJ Polarizable Water Model for the Coarse-Grained MARTINI Force Field. PloS Comput. Biol 2010, 6, e1000810. [PubMed: 20548957]

(55). Abrams C; Bussi G Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration. Entropy 2013, 16, 163–199.

(56). Bussi G; Gervasio FL; Laio A; Parrinello M Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics. J. Am. Chem. Soc 2006, 128, 13435–13441. [PubMed: 17031956]

(57). Savage PB; Li C; Taotafa U; Ding B; Guan Q Antibacterial Properties of Cationic Steroid Asntibiotics. FEMS Microbiol. Lett 2002, 217, 1–7. [PubMed: 12445638]

(58). Moretti A; Weeks RM; Chikindas M; Uhrich KE Cationic Amphiphiles with Specificity against Gram-Positive and Gram-Negative Bacteria: Chemical Composition and Architecture Combat Bacterial Membranes. Langmuir 2019, 35, 5557–5567. [PubMed: 30888181]

(59). López CA; Zgurskaya H; Gnanakaran S Molecular Characterization of the Outer Membrane of Pseudomonas Aeruginosa. Biochim. Biophys. Acta, Biomembr 2019, 183151. [PubMed: 31846648]

(60). Humphrey W; Dalke A; Schulten K VMD: Visual Molecular Dynamics. J. Mol. Graphics 1996, 14, 33–38.

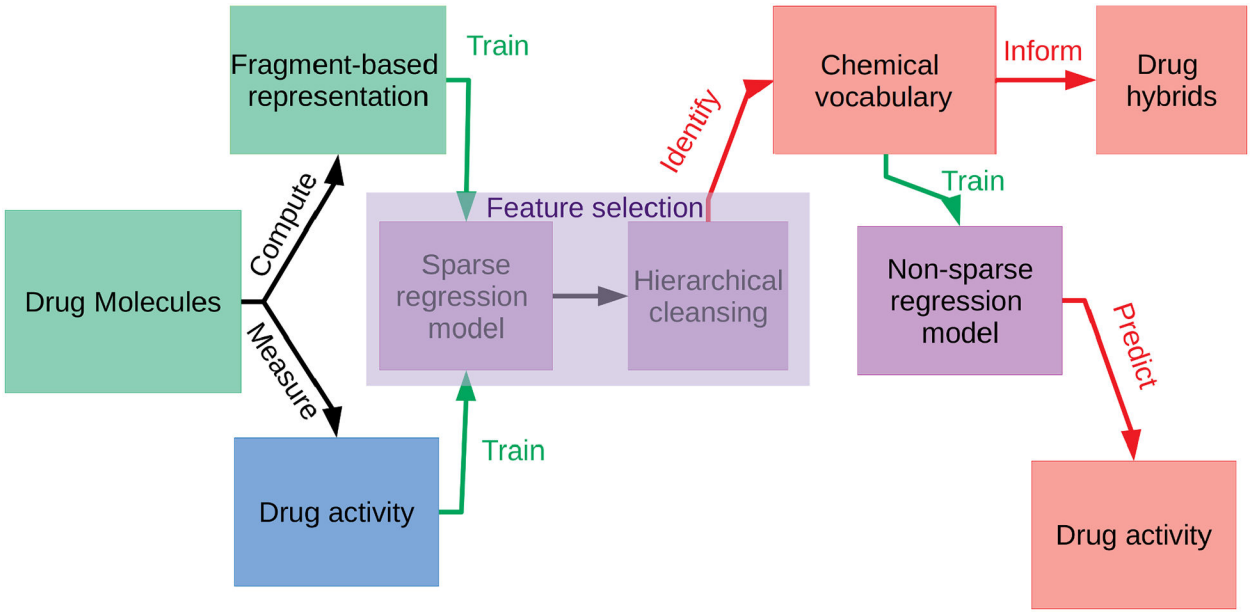


Figure 1:

Schematic of the basic Hunting FOX algorithm. We compute a fragment-based representation of drugs and use a combination of sparse regression and a hierarchical cleansing procedure to select a subset of relevant fragments that define a learned chemical vocabulary. We use these fragments to train a non-sparse regression model, from which we may predict drug class for novel molecules. In this study, the drug activities (blue) employed were MIC ratios of compounds in two different mutant strains of *P. aeruginosa* PAO1. The algorithm used these MIC ratios to classify a set of compounds based on their ability to permeate the outer membrane.

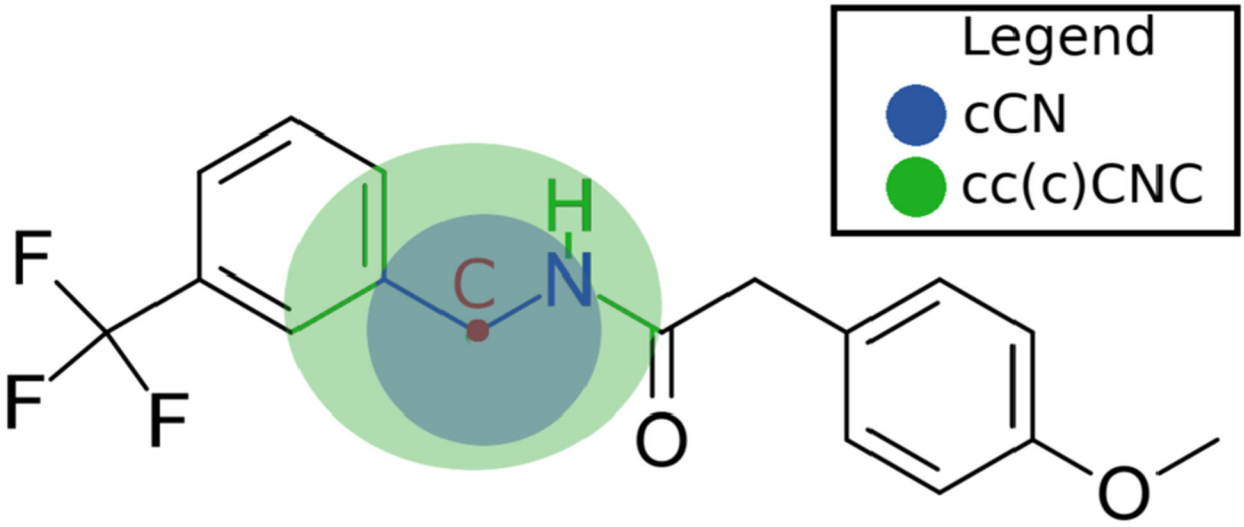


Figure 2:

Example of fragment definition by radius. The figure shows an arbitrarily chosen compound (OU-31237) with an arbitrarily chosen central atom in red. We show both the fragment of radius 1 bond about the central atom (in blue) and the fragment of radius 2 bonds about the central atom (in green, also includes the blue bonds and atoms). In the legend we indicate the SMILES string associated with each fragment. In Table S1 in the Supplemental Information we list all fragments of radius 1 and 2 contained in this molecule. Visualization of molecule chemical structure was rendered employing tools from the CDD Vault from Collaborative Drug Discovery (Burlingame, CA. www.collaborativedrug.com)[39]

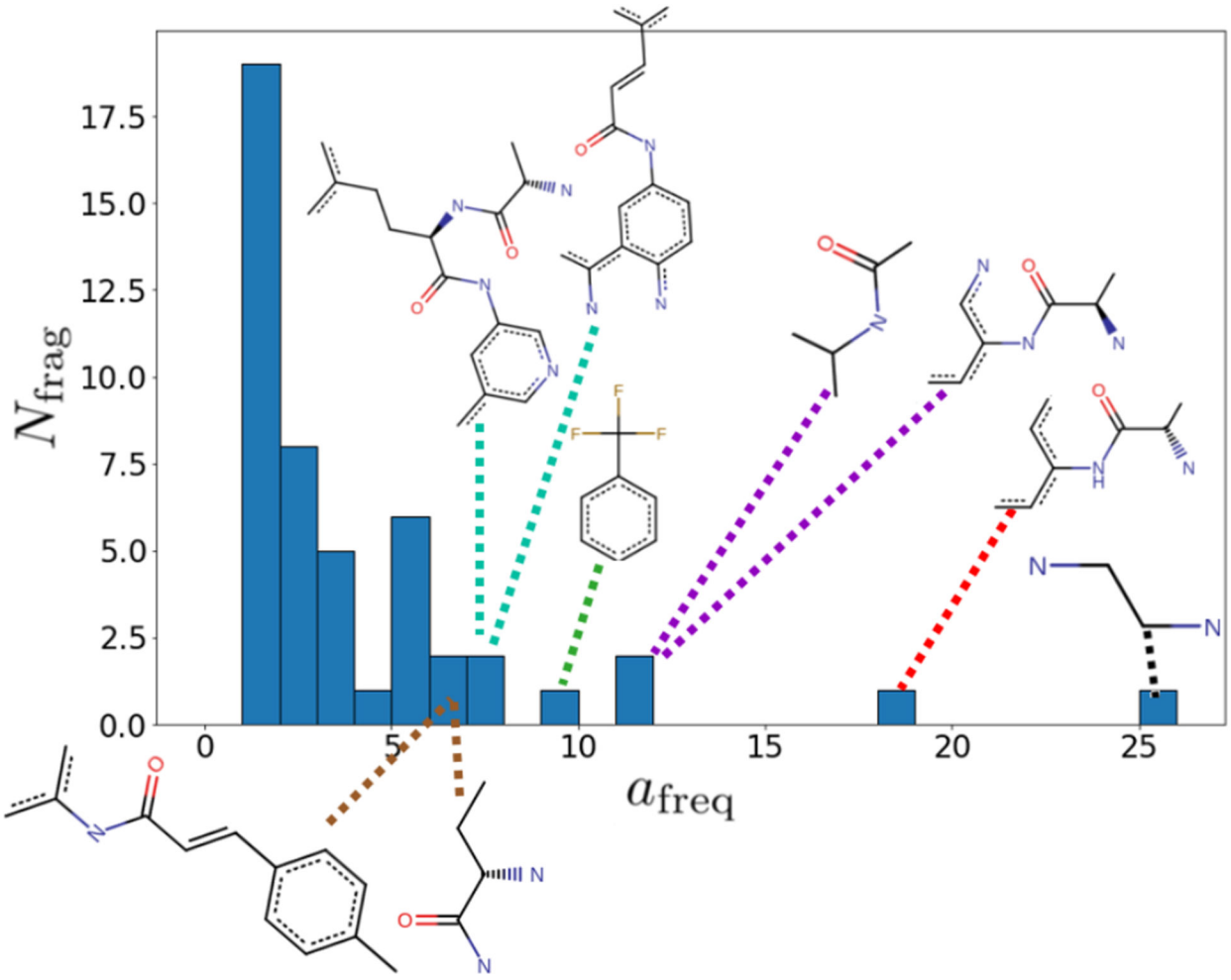


Figure 3:

Histogram showing the number of fragments N_{frag} identified by $a_{\mathrm{freq}}/28$ runs of Hunting FOX with randomly shuffled training/testing data. We render the 9 fragments that are reported by more than five runs and identify which bin they fall into. The fragments were rendered using Marvin 19.16.0, 2019, ChemAxon (http://www.chemaxon.com). We do not show hydrogen occupancy as it may change depending on how a fragment is connected to a molecule. Different colored dotted lines indicate fragments lying in different bins. Note that although we remove hierarchy in each iteration, we do not remove it between separate iterations, so there may be some hierarchically-related molecules in the final reported forty-eight.

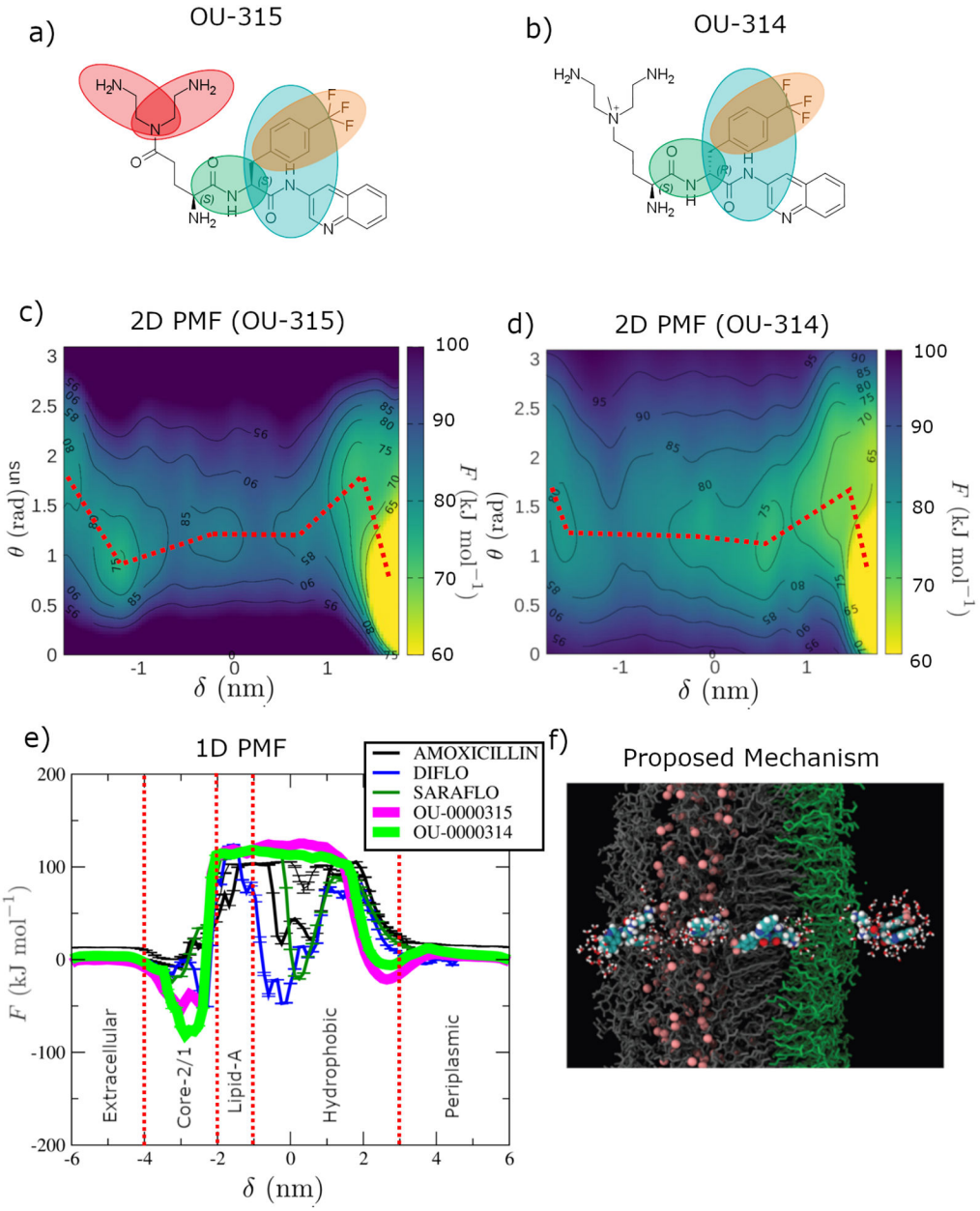


Figure 4:

Proposed mechanism of membrane translocation for compounds OU-315 and OU-314. a-b) Chemical structure of the target compounds, highlighting chemical fragments that were identified by more than five iterations of the Hunting FOX algorithm. Fragments with the same chemical identity are given the same color. c-d) 2D projection of the translocation free energy onto the biasing variables θ (orientation with respect to the membrane) and δ (perpendicular distance across the membrane) for c) OU-315 and d) OU-314. Red dashed lines indicate approximate path of lowest free energy. e) 1-D projection of the translocation onto δ potential of mean force (PMF) for Amoxicillin (black line), Difloxacin (blue line), Sarafloxacin (thin dark green line), OU-315 (thick magenta line), and OU-314 (thick bright green line). The PMF is projected based on the relative distance of the center of mass

(COM) of the drug with respect to the COM of the bacterial membrane. Different regions of the membrane are delimited by red dashed lines. f) Close-up view of the membrane translocation for OU-315. Note the presence of water molecules around the drug that enhance the permeation process. Image rendered with VMD[60].

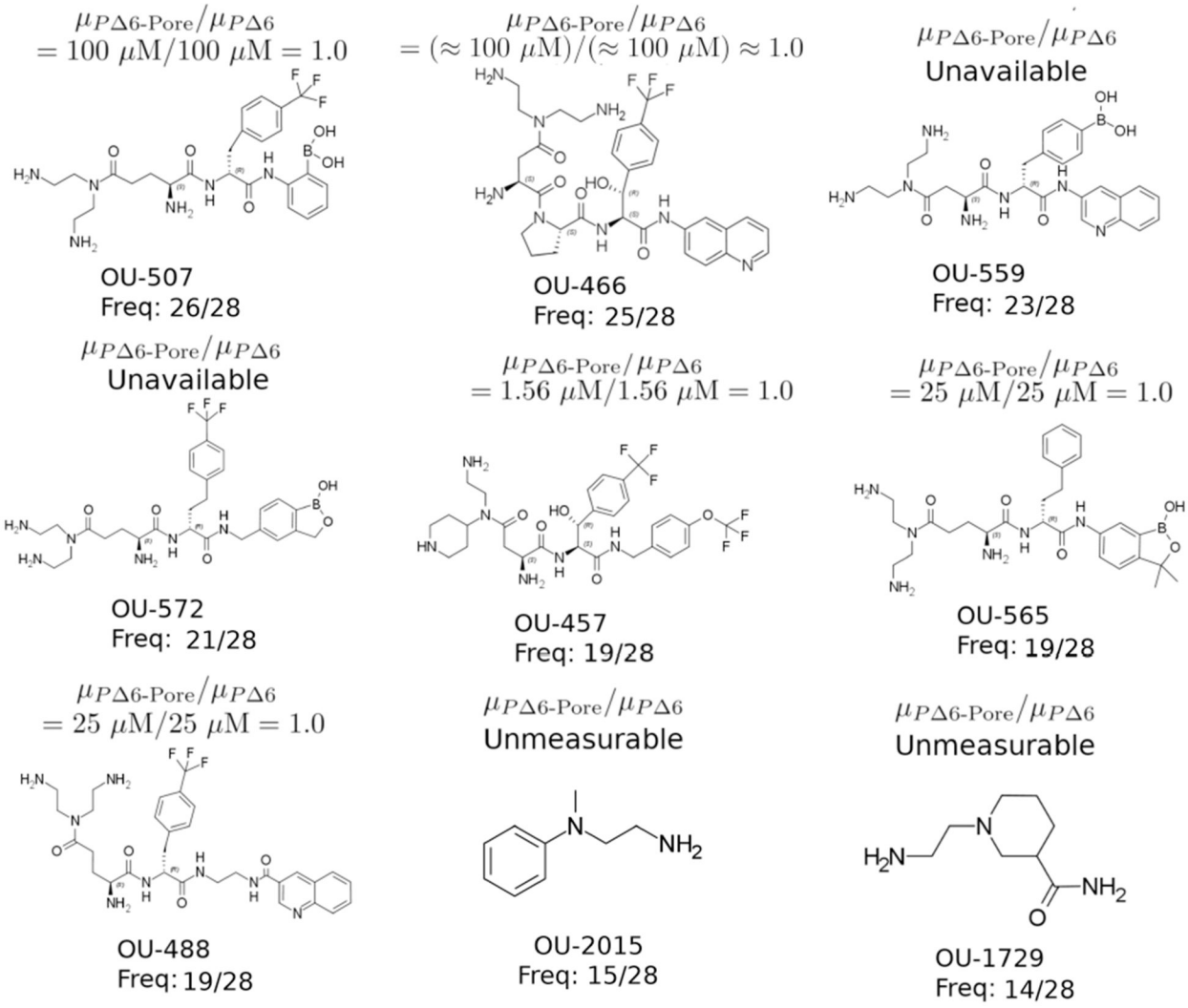


Figure 5:

Molecules selected as hits by at least 14/28 repeated runs of the Hunting Fox algorithm. We note the experimental measurements above the chemical structures of the molecules and the molecule name below. Visualizations of molecule chemical structure were rendered employing tools from the CDD Vault from Collaborative Drug Discovery (Burlingame, CA. www.collaborativedrug.com)[39].

Table 1:

Average area under the ROC curve for classifiers. We consider both the performance of classifiers trained on the fragments selected in a single run (AUC_{sing}) and the performance of classifiers trained on the top nine fragments returned in at least 5 runs of the algorithm (AUC_{top9}). We report the average performance as the mean of the 28 means of the 5 classifiers along with the standard error in the mean.

Class	$\langle AUC \rangle_{sing}$	$\langle AUC \rangle_{top9}$
0	0.842 ± 0.007	0.8819 ± 0.0005
1	0.51 ± 0.1	0.600 ± 0.003
2	0.764 ± 0.005	0.797 ± 0.001
3	0.804 ± 0.005	0.847 ± 0.001
4	0.735 ± 0.006	0.768 ± 0.001

Table 2:

Average percent enrichment at a ranking of 10% for classifiers. We consider both the performance of classifiers trained on the fragments selected in a single run ($\langle \epsilon(10\%) \rangle_{\text{sing}}$) and the performance of classifiers trained on the top nine fragments returned in at least 5 runs of the algorithm ($\langle \epsilon(10\%) \rangle_{\text{top9}}$). We report the average performance as the mean of the 28 means of the 5 classifiers along with the standard error in the mean.

Class	$\langle \epsilon(10\%) \rangle_{\text{sing}}$ (%)	$\langle \epsilon(10\%) \rangle_{\text{top9}}$ (%)
0	65 ± 1	82.0 ± 0.5
1	32 ± 8	154 ± 2
2	134 ± 4	133 ± 2
3	126 ± 3	173 ± 3
4	107 ± 3	141 ± 2