University of Tennessee, Knoxville
TRACE: Tennessee Research and Creative Exchange

Doctoral Dissertations | Graduate School

8-2020

Computational Approaches to Understanding the Structure, Dynamics, Functions, and Mechanisms of Various Bacterial Proteins

Connor Cooper
University of Tennessee, scoope33@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

Recommended Citation

Cooper, Connor, "Computational Approaches to Understanding the Structure, Dynamics, Functions, and Mechanisms of Various Bacterial Proteins. " PhD diss., University of Tennessee, 2020.
https://trace.tennessee.edu/utk_graddiss/6795

This Dissertation is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Connor Cooper entitled "Computational Approaches to Understanding the Structure, Dynamics, Functions, and Mechanisms of Various Bacterial Proteins." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Jerry Parks, Major Professor

We have read this dissertation and recommend its acceptance:

Gladys Alexandre, Jennifer Morrell-Falvey, Margaret Staton

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Computational Approaches to Understanding the Structure, Dynamics, Functions, and Mechanisms of Various Bacterial Proteins

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Connor Jay Cooper
August 2020

Copyright © 2020 by Connor Jay Cooper
All rights reserved.

DEDICATION

I dedicate this work to my grandmother, Betty. Without her as a mentor and role model I undoubtedly would not be the person I am today.

ACKNOWLEDGEMENTS

I would first like to thank my family for always being supportive in every aspect of my life. I would also like to acknowledge my high school chemistry teacher Mr. Koch at Raymond High School and undergraduate research mentor at the University of New England, Dr. John Stubbs. Mr. Koch instilled my interest in science and my exposure to computational chemistry from working with Dr. Stubbs drove me to pursue this degree. I would also like to thank my Ph.D. advisor Dr. Jerry Parks for giving me seemingly endless opportunities to address interesting scientific questions all while providing unwavering support and guidance along the way. I would like to express gratitude for the help and guidance I received from the members of my committee as well as from Dr. Jeremy Smith. In addition, I would like to thank all of our collaborators along with the people in the Genome Science & Technology program and in the Center for Molecular Biophysics that provided help and useful insights. Lastly, I would like to express appreciation for all of the friends who have been supportive of me during this journey.

ABSTRACT

The 3D structure of a protein can be fundamentally useful for understanding protein function. In the absence of an experimentally determined structure, the most common way to obtain protein structures is to use homology modeling, or the mapping of the target sequence onto a closely related homolog with an available structure. However, despite recent efforts in structural biology, the 3D structures of many proteins remain unknown. Recent advances in genomic and metagenomic sequencing coupled with coevolution analysis and protein structure prediction have allowed for highly accurate models of proteins that were previously considered intractable to model due to the lack of suitable templates. Structural models obtained from homology modeling, coevolution-based modeling, or crystallography can then be used with other computational tools such as small molecule docking or molecular dynamics (MD) simulations to help understand protein function, dynamics, and mechanism.

Here coevolution-based modeling was used to build a structural model of the HgcAB complex involved in mercury methylation (Chapter I). Based on the model it was proposed that conserved cysteines in HgcB are involved in shuttling mercury, methylmercury, or both. MD simulations and docking to a homology model of *E. coli* inosine monophosphate dehydrogenase (IMPDH) provided insights into how a single amino acid mutation could relieve inhibition by altering protein structure and dynamics (Chapter II). Coevolution-based structure prediction was also combined with docking, and experimental activity data to generate machine learning models that predict enzyme substrate scope for a series of bacterial nitrilases (Chapter III). Machine learning was also used to identify physicochemical properties that describe outer membrane permeability and efflux in *E. coli* and *P. aeruginosa* and new efflux pump inhibitors for the *E. coli* AcrAB-TolC efflux pump were identified using existing physicochemical guidelines in combination with small molecule docking to a homology model of AcrA (Chapter IV). Lastly, quantum mechanical/molecular mechanical simulations were used to study the mechanism of a key proton transfer step in Toho-1 beta-lactamase using experimentally determined structures of both the apo and cefotaxime-bound forms. These simulations revealed that substrate binding promotes catalysis by enhancing the favorability of this initial proton transfer step (Chapter V).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Protein structures can provide valuable insights into enzymatic function. The overall fold, domain architecture, and spatial arrangement of residues involved in substrate recognition and catalysis all provide useful clues. In the absence of an experimentally determined structure, modeling is an alternative approach. When suitable templates are available, homology modeling is often the method of choice. However, the accuracy of the modeled structure depends on various factors, including scoring functions and conformational sampling strategies.[1] Homology model accuracy also depends on the quality of the template as well as the sequence similarity to the query sequence. However, templates are not available for many proteins. In the absence of suitable templates, models can be generated by using coevolutionary information obtained from large multiple sequence alignments of homologs to the target protein. Here, pairs of amino acids that are found to coevolve in sequence space are expected to be in close proximity to one another in the folded protein. Using this information, contact restraints can be derived and used to generate accurate structural models that reach homology level accuracy.[2-6]

Structural models and experimentally determined structures generated by any or all of these approaches can then be further studied with other computational tools to address questions pertaining to the structure, dynamics, functions, and mechanisms of various bacterial proteins. For example, molecular dynamics (MD) simulations can be used to investigate conformational changes in a protein upon mutation. Meanwhile, docking of small molecule ligands can be used to identify substrates for an enzyme and quantum mechanical/molecular mechanical (QM/MM) simulations can be used to study enzyme mechanisms in detail.

HgcAB

The *hgcAB* gene pair has been identified anaerobic bacteria and archaea that methylate mercury (Hg). [7-10] The *hgcAB* gene pair occurs in only ~1.4% of sequenced microbial genomes and deletion of *hgcA*, *hgcB*, or both has been show to completely eliminate methylmercury (MeHg) production.[11] Despite the rare occurrence of *hgcA* and *hgcB* in sequenced genomes, microorganisms with these genes are found throughout the world in highly diverse anaerobic settings. However, bacteria that methylate Hg are no less susceptible to Hg toxicity than those that do not, suggesting that it is not a mechanism for Hg detoxification.[12]

Through protein sequence analysis, HgcA was predicted to consist of an N-terminal corrinoid (i.e., vitamin B_{12}) binding protein (CBD) homologous to the corrinoid iron-sulfur protein (CFeSP) [13-15] and a C-terminal transmembrane domain (TMD) that consists of five helices.[7] Similar to CFeSP, HgcA was also predicted to have a "cap helix" that interacts non-covalently with the corrinoid. This helix contains several highly conserved residues, including a strictly conserved Cys that is absent in the CFeSP sequence. This Cys residue has been suggested to bind to the corrinoid cofactor based on observations from homology modeling. Site-directed mutagenesis experiments revealed that Ala or Thr mutations of this Cys (Cys93 in *Desulfovibrio desulfuricans* ND132) abolished Hg methylation, while mutation to His retained some activity.[16] Unlike the CBD, the TMD of HgcA lacks detectable sequence homology to any available protein structures. This domain has been shown to be essential for Hg methlylation activity, as C-terminal truncation mutants that removed the TMD are unable to methylate Hg.[16]

Bioinformatics analysis has suggested that HgcB is a bacterial ferredoxin with two separate CxxCxxCxxxCP motifs. These motifs are known to bind [4Fe-4S] clusters. HgcB also contains a strictly conserved Cys (Cys73 in *D. desulfuricans* ND132), located near the second [4Fe-4S]-binding motif. Mutation of this residue to Ala has been shown to eliminate Hg methylation in vivo.[10] There are also two additional Cys residues (Cys94 and Cys95 in *D. desulfuricans* ND13) at the C-terminus. For both of these Cys residues, Ala mutations have been shown to abolish Hg methylation activity, but single Cys mutants were not affected, suggesting that at least one of these Cys residues is required. Sequence homologs of HgcB have variable sequence length and number of C-terminal Cys residues.

HgcA and HgcB are expressed at very low levels,[17,18] making it particularly challenging to isolate and purify sufficient quantities of these proteins from the native host. Heterologous overexpression is difficult as Hg methylating microorganisms are obligate anaerobes and exposure to oxygen stops MeHg production.[19] In addition, the uptake of corrinoids is tightly regulated[20] and the iron-sulfur clusters require the proper machinery for assembly.[21] The TMD also poses a challenge in structure determination, as transmembrane proteins are difficult to crystallize. Lastly, structure determination of transmembrane proteins such as the TMD remains challenging for X-ray crystallography, nuclear magnetic resonance, and cryo-EM.

Structural models of the HgcAB complex would provide valuable insight into the biochemical mechanism of Hg methylation. These structures can be obtained using coevolution analysis. This method requires a large multiple sequence alignment and the massive amount of sequence data available provide a large number of diverse protein sequences. Recently, it was shown that highly accurate structures can be obtained using coevolution analysis on sequences obtained from metagenomes.[22] Due to the rarity of these genes in sequence genomes and the large number of sequences required to generate accurate contacts from coevolution analysis, metagenomes can be crucial for generating accurate models of the HgcAB complex. In Chapter I this complex is confirmed to bind corrinoid and iron-sulfur clusters, as predicted previously and models of the HgcAB complex were generated using metagenome sequences for coevolution analysis (**Figure 1A**). In addition, the relevant cofactors are incorporated to generate a complete model of the complex that is used to provide mechanistic insights into how microorganisms methylate Hg.

IMPDH

Microbes are able to utilize a wide array of compounds as carbon and energy sources. By expanding the range of compounds, a particular organism can use, new environmental niches can be accessed, and microbes can be engineered to use new feedstocks. In nature, horizontal gene transfer (HGT) allows for transfer of catabolic pathways between bacterial strains. In the laboratory setting metabolic engineering can be used for transferring pathways.[23, 24] However, these engineered pathways often fail to function properly in the host, requiring optimization to minimize deleterious interactions.[25,26, 27]

Pathways for catabolism of lignin-derived aromatic compounds are found to be widespread in nature[28], and often undergo HGT.[29] Previous metabolic engineering efforts have attempted to generate strains of *Escherichia coli* that use lignin-derived compounds (i.e., 4-hydroxybenzoate, 4-HB) as sole energy sources.[30, 31] However, introduction of the relevant pathways was not enough to enable growth. To overcome this issue, directed evolution was used to select for strains with improved growth in order to identify causal mutations that improve function of this engineered pathway. An extension of this work optimized these pathways for phenylpropanoid catabolism in *E. coli* and point mutations in the host were able to readily alleviate limitations in pathway activity. In one of these pathways, 4-hydroxybenzaldehyde, a pathway intermediate, inhibited purine

nucleotide biosynthesis. Interestingly, this inhibition was relieved by single amino acid replacements in the enzyme inosine-5′-monophosphate dehydrogenase (IMPDH). In Chapter II homology models of IMPDH were generated for small molecule docking to predict the inhibitor binding site followed by MD simulations to understand the structural and dynamic changes between the wild-type and mutant enzyme that lead to inhibition relief (**Figure 1B**).

Nitrilase

Enzymes are often able to accept multiple molecules as substrates and knowledge about the repertoire of substrates a given enzyme prefers (also known as substrate scope) can provide information about biological pathways and insights for metabolic engineering. Sequenced based methods are effective at predicting information about the broad categories of molecules that may act as substrates for a given enzyme to provide valuable information about potential protein function (i.e. active site residues, gene ontology terms, conserved domains), but are unable to predict the substrate scope. BRENDA[32] is a manually curated database of ~84,000 enzymes that contains information about substrate specificity, but is limited to only experimentally verified systems. The ability to predict the substrate scope of an enzyme based on limited experimental data would be beneficial to studying enzyme function.

Several efforts have been used to predict substrate specificity. For example, the substrate specificity of an enoyl-acyl carrier protein reductase was predicted using small molecule docking of putative substrates to an available crystal structure.[33] As mentioned previously, structural modeling (i.e., homology or coevolution-based modeling) can be used in the absence of an available crystal structure. These models can then be used for computational docking of ligands.[34-36] However, docking studies often are unable to differentiate between ligands with similar scaffolds and docking also neglects to account for chemical reactivity, making it challenging to predict enzymatic activity.[37] Combining information from structural modeling, docking, and other sources such as physicochemical properties of the ligand and the enzyme active site is expected to help overcome some of these limitations observed from docking to structural models of proteins.

Machine learning (ML) has recently been applied to a variety of problems in fields such as quantum mechanics, physical chemistry, and biophysics. For example, channelrhodopsin was

engineered with higher light sensitivity using directed evolution in combination with information from protein sequence and contact maps generated from crystal structures.[38] ML has also shown to be a promising way to predict substrate specificity.[39, 40] Recently, a decision tree-based classifier that incorporated sequence information and physicochemical properties of substrate donor and acceptor molecules with experimental activity data was able to accurately predict enzymatic activity (~90%).[41]An extension of this approach would be to directly incorporate 3D structural information.

Nitrilases are a family of enzymes that hydrolyze nitrile compounds to their corresponding carboxylic acids and ammonia. This enzyme family has a broad scope and is found in a wide range of organisms. These enzymes are involved in a variety of biological processes in addition to degradation of toxic nitrile compounds.[42] In plant-microbe interactions these enzymes are of interest for improving food crop production, as they are thought to be involved in hormone synthesis, nutrient assimilation, detoxification, and modulation of plant development and physiology.[43] In addition, nitriles are of interest in drug design.[44, 45] In general, nitrilases fall into three categories of substrate specificities: aliphatic, arylaceto-, and aromatic nitrilases.[43, 46] However, existing sequence-based annotations are limited in their ability to classify nitrilases.

Functional screening of microbial metagenomes has led to the identification of a diverse collection of nitrilases and reactivity toward specific substrates was found to be strongly correlated with phylogenetic relationships.[44] To evaluate a large number of putative nitrilases, a high-throughput method is essential. Various fluorogenic or chromogenic activity assays are available to do so.[45, 47, 48] Recently, a chromogenic method was developed to screen nitrilases produced in crude cell extracts which alleviates purification steps and facilitates screening.[49]

Shifts in substrate specificity were observed within specific subfamilies, suggesting that subtle changes in sequence can noticeably alter their substrate scope.[44] Chapter III describes an integrated and modular approach that combines experimental activity assays with coevolution-based protein structure prediction, small molecule docking, and calculation of physicochemical properties of a series of 12 bacterial nitrilases and a set of 20 nitriles (**Figure 1C**). This information is then used to train various machine learning classifiers to predict enzyme substrate scope.

Antibiotic outer membrane permeability and efflux

Bacteria are developing resistance to antibiotics at an alarming rate through an arsenal of resistance mechanisms and understanding these various resistance mechanisms is important for developing strategies to overcome multi-drug resistance and restore antibiotic effectiveness. The presence of multidrug resistant strains in clinics often leave clinicians with no therapeutic options and the discovery of new antibiotics that are active against these pathogens is a major challenge.[50]

Compared to Gram-positive bacteria, Gram-negative bacteria are more resistant to antibiotics. Gram-negative cell envelopes are comprised of the outer membrane (OM) that provides protection from toxic molecules and enzymatic attacks. The OM is an asymmetric bilayer of lipopolysaccharides and phospholipids along with both substrate-specific channels and non-selective porins.[51, 52] The inner membrane is a phospholipid bilayer that contains multidrug efflux pumps that protect intracellular functions by expelling small, toxic molecules from the cell.[53] The low-permeability of the OM coupled with active efflux in the inner membrane provides two synergistic barriers that are responsible for antibiotic resistance in Gram-negative bacteria. In particular, efflux is considered a major bottleneck in addressing multidrug resistance[54] and the discovery of new antibiotics is hindered by the lack of practical rules to maximize OM permeability and minimize active efflux.[55, 56]

Separation of these two barriers allows for different sets of rules to be established that define OM permeation and efflux.[57] To investigate efflux in the absence of the OM barrier, a hyperporination approach can be used that facilitates control of OM permeability in Gram-negative cells through inducible expression of a chromosomally encoded open pore.[58] The deletion of efflux pumps allows for OM permeability to be characterized without concern for the effect of efflux. Different classes of antibiotics can also be used to further investigate the individual and synergistic contributions of these two barriers. β-lactams (BLs) and fluoroquinolones (FQs) have been extensively developed and remain the major antibiotics administered in clinics. FQs target DNA replication by inhibiting DNA topoisomerases. Thus, to be effective, these antibiotics must penetrate both the outer and inner membranes and evade efflux pumps. In contrast, BLs act in the periplasm. The different targets for BLs and FQs allow for further investigation into the barriers limiting antibiotic activity in Gram-negative bacteria.

Pseudomonas aeruginosa and *E. coli* are two Gram-negative pathogens that differ significantly in their permeability barriers, despite having similar OM lipid compositions.[51, 59, 60] These two species differ in the composition and structure of their major general porins. *E. coli* has~200,000 copies per cell of OmpF/C porins, which have a molecular mass cutoff of ~600 Da, allowing for a significant portion of antibiotics to permeate.[61] In contrast, *P. aeruginosa* only utilizes substrate-specific porins to take up small compounds.[62] *P. aeruginosa* has shown susceptibility to FQs and some BLs, suggesting alternate routes of OM permeation. Fortunately, hyperporination negates the differences in OM permeability in these two species. [57, 58]

AcrAB-TolC is the main efflux pump in *E. coli*, which, which consists of three main components, AcrB, AcrA and TolC that assemble in 3:6:3 stoichiometry to span the entire cell envelope.[63] AcrB is a homotrimer with an integral membrane domain, a periplasmic porter domain that binds and extrudes substrates, and a docking domain that interacts with AcrA.[64, 65] AcrA consists of α-hairpin, lipoyl, β-barrel, and membrane-proximal domains. TolC is a trimeric protein that consists of a β-barrel domain embedded in the OM and a periplasmic α-helical coiled-coil domain.[66]

An effective efflux pump inhibitor must first bypass the OM. Recently, random forest ML was used to help determine that small-molecule compounds containing amine functional groups were most likely to accumulate in *E. coli,* and incorporation of a primary amine into the Gram-positive antibiotic deoxynybomycin resulted in a new antibiotic with broad-spectrum activity against multidrug-resistant Gram-negative bacteria.[67] In addition to containing an amine, antibiotics that were polar, amphiphilic, relatively rigid, and had low globularity were found to be more likely to permeate. In Chapter IV (**Figure 1D**) random forest classification is used to identify molecular properties of antibiotics that are associated with their activities, measured as minimum inhibitory concentrations in *P. aeruginosa* and *E. coli* strains with controlled permeability barriers. Physicochemical property guidelines are then used to identify novel efflux pump inhibitors.

β-lactamase

Enzymatic inactivation is often the preferred mechanism of resistance, as enzymes can catalyze chemical transformations that inactivate an entire class of antibiotics. The most extensively studied enzymatic inactivation mechanism is the inactivation of BL antibiotics by β-lactamase enzymes.

Since their introduction into the clinic, BLs have revolutionized medicine.[68, 69] However, the development of antibiotic resistance is inevitable. Despite the wide variety of BL antibiotics available today, resistant strains pose a threat to public health.

Sequence homology is used to divide β-lactamases into four classes (A-D).[70] Typical class A β-lactamases include extended-spectrum β- lactamase (ESBL) cefotaxime-resistant (CTX) M-type enzymes. CTX-M-type CTX-M ESBLs can inactivate monobactam antibiotics and all generations of cephalosporins.[70-72] Toho-1 is a class A CTX-M-type ESBL β-lactamase composed of two highly conserved domains.[73] All class A β-lactamases, including Toho-1, utilize a serine nucleophile to cleave the β-lactam bond. Several detailed mechanisms have been proposed for this reaction.[74-77] One way to differentiate between these mechanisms is to unambiguously identifying key protonation states and hydrogen-bonding interactions of the catalytically important residues and the substrate. Neutron crystallography is ideally suited to experimentally determine protonation states as it allows for hydrogen atom positions to be determined.

QM/MM calculations can provide key mechanistic insights that are complementary to crystallographic experiments by allowing for detailed inspection of short-lived intermediates and transition states and the quantification of reaction energetics of enzymatic reactions. Configurational sampling is required to provide information about free energies and can be achieved using umbrella sampling by running a series of restrained simulations along the reaction coordinate, in this case a hydrogen atom transfer. However, the computational cost of QM/MM umbrella sampling with density functional theory is high as it requires calculations at the quantum mechanical level at each timestep. Instead, semiempirical QM methods are often used to perform these calculations. Previous QM/MM studies of class A β-lactamases have focused on the acylation[78-81]and deacylation steps[82, 83], and have helped establish likely mechanisms for BL inactivation. However, the effect the substrate has on this mechanism has not yet been investigated. In Chapter V X-ray and neutron crystallography is combined with QM/MM simulations to address this question using protonation states confirmed crystallographically (**Figure 1E**).

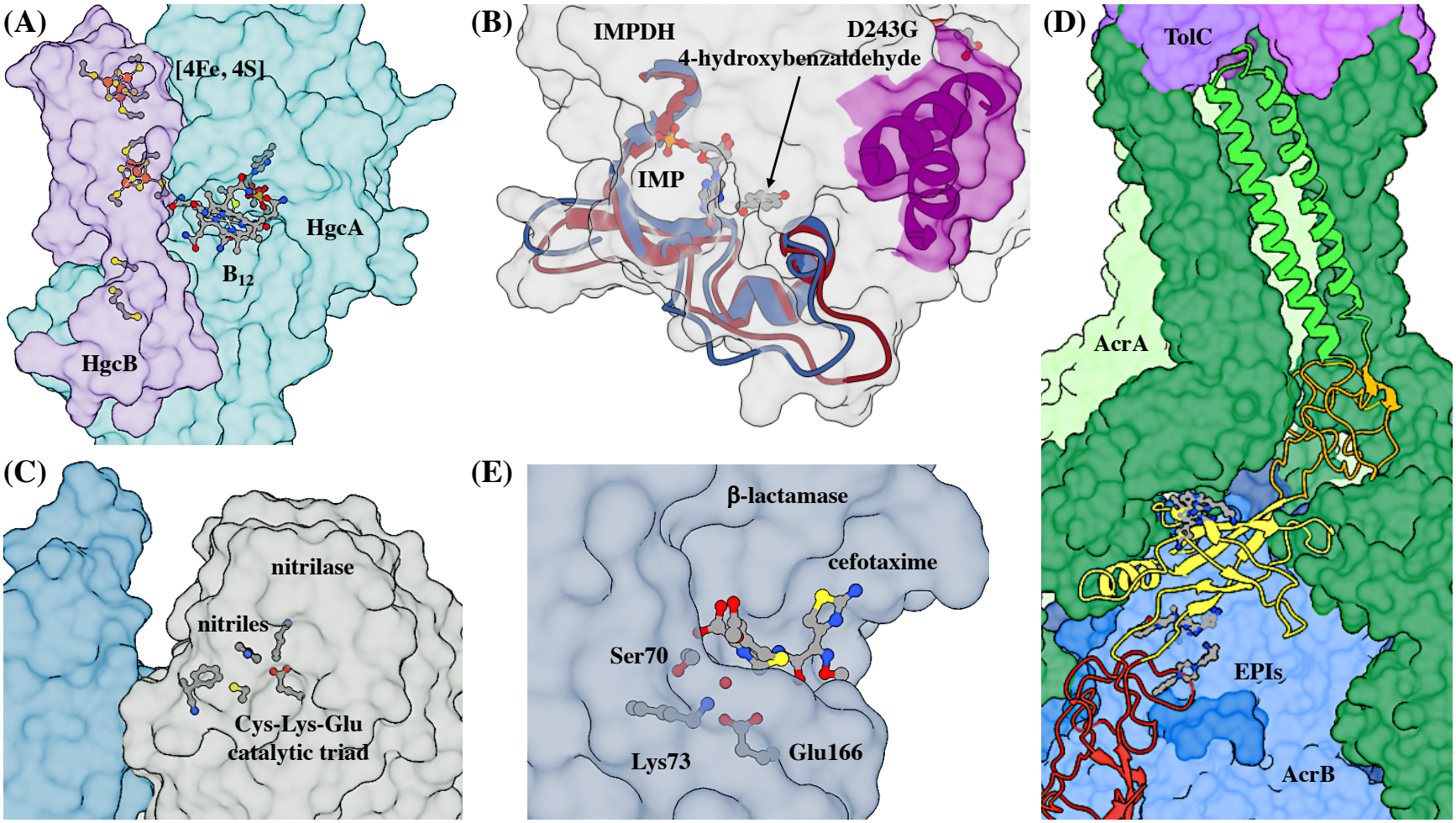


Figure 1. Schematic overview of chapters. (**A**) In Chapter I coevolution-based structural modeling is used to build a model of the HgcAB complex involved in mercury methylation. (**B**) In Chapter II homology modeling, docking, and molecular dynamics (MD) simulations are used to investigate how a single amino acid mutation relieves inhibition of *E. coli* inosine monophosphate dehydrogenase (IMPDH) by 4-hydroxybenzaldehyde (**C**) Chapter III describes the use of coevolution-based structure prediction, docking, and machine learning to predict the enzyme substrate scope of a series of bacterial nitrilases. (**D**) In Chapter IV machine learning is also used to identify physicochemical properties that define antibiotic permeability and efflux in *E. coli* and *P. aeruginosa*. Physicochemical property filters are then used to discover novel efflux pump inhibitors (EPIs) for the *E. coli* AcrAB-TolC efflux pump by docking to AcrA. (**E**) Chapter V describes the use of MD and quantum mechanical/molecular mechanical (QM/MM) free energy simulations of apo and cefotaxime-bound β-lactamase to investigate how substrate binding affects catalysis.

CHAPTER I

COEVOLUTION-BASED STRUCTURE PREDICTION OF THE HGCAB COMPLEX INVOLVED IN MERCURY METHYLATION

Text and figures are taken from the following:

Cooper, C.J., Ovchinnikov, S., Zheng, K., Rush, K.W., Podar, M., Pavlopoulos, G., Kyrpides, N.C., Johs, A., Ragsdale, S.W., and Parks, J.M. Structure determination of the HgcAB complex using metagenome sequence data: insights into the mechanism of mercury methylation. *Commun. Biol.* 2020. DOI: 10.1038/s42003-020-1047-5. In press.

S.O., G.P. and N.C.K performed the metagenome searches; C.J.C., S.O. and J.M.P. performed the structural modeling; K.Z., K.W.R. and S.W.R. performed the cloning, expression, purification and spectroscopy; C.J.C., A.J., B.J.S and J.M.P. performed the mechanistic analysis; M.P. performed the phylogenetic analysis; C.J.C., M.P. and J.M.P. prepared the manuscript with input from all other authors.

Abstract

Bacteria and archaea possessing the *hgcAB* gene pair methylate inorganic mercury (Hg) to form highly toxic methylmercury. HgcA consists of a corrinoid binding domain and a transmembrane domain, and HgcB is a dicluster ferredoxin. However, their detailed structure and function have not been thoroughly characterized. We modeled the HgcAB complex by combining metagenome sequence data mining, coevolution analysis, and Rosetta structure calculations. In addition, we overexpressed HgcA and HgcB in *Escherichia coli* and confirmed spectroscopically that they bind cobalamin and [4Fe-4S] clusters, respectively, and incorporated these cofactors into the structural model. Surprisingly, the two domains of HgcA do not interact with each other, but HgcB forms extensive contacts with both domains. The model suggests that conserved cysteines in HgcB are involved in shuttling Hg^{II}, methylmercury, or both. These findings refine our understanding of the mechanism of Hg methylation and expand the known repertoire of corrinoid methyltransferases in nature.

Introduction

Anaerobic bacteria and archaea possessing the *hgcAB* gene pair methylate inorganic mercury (Hg) to form methylmercury (CH_3Hg^+),[7-10] a potent neurotoxin. Deletion of *hgcA*, *hgcB*, or both completely abolishes the ability of microorganisms to make methylmercury. These genes are distributed somewhat sporadically among various Proteobacteria (*Deltaproteobacteria*), Firmicutes, and Euryarchaeota. They are also found in some Chloroflexi (*Dehalococcoides*), Chrysiogenetes, Nitrospirae, and others.

The *hgcAB* gene pair is relatively rare, occurring in only ~1.4% of sequenced microbial genomes.[11] Nevertheless, microorganisms harboring these genes are distributed worldwide in highly diverse anaerobic settings including soils, sediments, periphyton, rice paddies, invertebrate digestive tracts, and various extreme environments. It is not known why microorganisms methylate Hg, but this process is generally not thought to be a Hg detoxification mechanism because microorganisms harboring *hgcAB* genes are apparently no less susceptible to Hg toxicity than those lacking them.[12]

Protein sequence analysis revealed that HgcA (a subset of the CO dehydrogenase/acetyl-CoA synthase delta subunit family, PF03599) is a corrinoid (i.e., vitamin B_{12}-dependent) protein consisting of an N-terminal corrinoid binding domain (CBD) and a C-terminal transmembrane domain (TMD) with five TM helices.[7] The CBD of HgcA bears homology to the C-terminal domain of the large subunit of the corrinoid iron-sulfur protein (CFeSP) from the Wood-Ljungdahl pathway in acetogenic bacteria.[13-15]

HgcA was predicted to include a "cap helix" in its CBD similar to that in CFeSP.[13] The cap helix in CFeSP interacts noncovalently with the α face of the corrinoid cofactor. In HgcA, the putative cap helix region includes several highly conserved residues, one of which is a strictly conserved Cys residue (Cys93 in *Desulfovibrio desulfuricans* ND132), that is not present at the corresponding position in the sequence of CFeSP. On the basis of its position in a homology model of the CBD, this Cys residue was predicted to bind the corrinoid cofactor in a cobalt-thiolate, or "Cys-on" configuration.[7] Findings from in vivo site-directed mutagenesis experiments are consistent with Cys-on cofactor binding.[16] Mutation of Cys93 to Ala or Thr resulted in a complete loss of Hg methylation activity, but a His mutant, which can presumably still coordinate with Co, retained

partial activity. In addition, substitution of several amino acids in the cap helix region with a helix-breaking Pro residue drastically reduced or completely abolished activity. A quantum chemical study showed that Cys-on coordination promotes the exchange of one organometallic (Co–C) bond for another (Hg–C).[84] Recently, the first example of Cys-on coordination in a protein was observed for the bacterial vitamin B_{12} transporter BtuM co-crystallized with cobalamin.[85]

The TMD of HgcA has no detectable sequence homology (i.e., BLAST E-value < 10) to any structurally characterized protein. C-terminal truncation mutants of HgcA in which the TMD was deleted by introducing a stop codon after the nucleotides encoding either amino acid 166 or 187 were both unable to methylate Hg, indicating that this domain is essential for activity.[16]

HgcB is a 10.2 kDa bacterial ferredoxin (Pfam entries PF13237 and PF00037) that includes two CxxCxxCxxxCP motifs, which are known to bind [4Fe-4S] clusters. In addition, HgcB includes another strictly conserved Cys (Cys73 in *D. desulfuricans* ND132), located ~12 residues downstream of the second [4Fe-4S]-binding motif, and up to four additional Cys residues at its C-terminus. Two cysteines are present at the C-terminus of ND132 (Cys94 and Cys95). Homologs of HgcB have variable sequence length, in particular in the tail region near the C-terminus. Mutation of Cys73 to Ala completely abolished Hg methylation in vivo.[10] Mutation of either C-terminal cysteine (Cys94 or Cys95) individually to Ala did not affect Hg methylation activity, but mutation of both residues simultaneously to Ala led to a 95% reduction in activity compared to the wild-type. Thus, at least one Cys is required at the C-terminus for maximal Hg methylation activity.

In a proteomics study of *Geobacter sulfurreducens* PCA, another confirmed Hg-methylating bacterium, HgcA and HgcB were not detected due to low protein abundance.[17] In a subsequent study of *D. desulfuricans* ND132, HgcA was detected in low abundance but HgcB was again not detected.[18] Thus, isolation and purification of sufficient quantities of protein from a native host are expected to be challenging. Heterologous overexpression of HgcA and HgcB is complicated by a number of factors. For example, many Hg-methylating organisms are obligate anaerobes. Based on the proposed Hg methylation cycle, maintaining a low redox potential is essential for the function of HgcA and HgcB. It has been demonstrated that exposure to oxygen inhibits MeHg

formation in cell lysates of *D. desulfuricans* ND132.[19] In addition, incorporation of the corrinoid cofactor and [4Fe-4S] clusters is nontrivial in heterologous hosts such as *Escherichia coli* because the uptake of corrinoids is tightly regulated[20] and overexpression of recombinant proteins increases the demand on the machinery required to assemble iron-sulfur clusters.[21] Lastly, although tremendous progress has been made in recent years, structure determination of transmembrane proteins with X-ray crystallography, nuclear magnetic resonance, or cryo-electron microscopy remains a challenge.

In the absence of an experimentally determined structure, structural modeling is a viable means for obtaining mechanistic insight into protein function. Homology modeling is generally the method of choice, provided that suitable template structures are available. When templates are lacking, however, models can be generated by leveraging coevolution information inferred from a multiple sequence alignment. Pairs of amino acids that coevolve are likely to be in close spatial proximity in the folded protein. Thus, by imposing contact restraints derived from coevolution analysis with ab initio protein modeling, accurate structural models can be obtained.[2-6]

Coevolution analysis requires as input a multiple sequence alignment with a large number of sequences. The massive amount of data available in public repositories such as the UniRef100 database[86] and the DOE Joint Genome Institute (JGI) metagenome database[87] provide a rich source of diverse protein sequences. Recently, it was shown that the combination of metagenome sequences, coevolution analysis and Rosetta protein structure calculations can produce highly accurate structures.[22] For a multiple sequence alignment, when the effective number of sequences divided by the square root of the sequence length L is greater than 64 (where the effective number of sequences is defined as 1 over the number of sequences within 80% identity), then homology model-level accuracy or better can be obtained.

Structural models of HgcA and HgcB would provide valuable insight into the biochemical mechanism of Hg methylation. Here we express HgcA and HgcB individually in *E. coli* and show by UV-visible spectroscopy that they indeed bind corrinoid and iron-sulfur cofactors, as predicted from previous bioinformatics analyses. We then combine metagenome-based protein structure calculations to generate models of the individual domains of HgcA and of HgcB. We then show

how these domains assemble to form the HgcAB complex and incorporate a vitamin B_{12} corrinoid cofactor and two [4Fe-4S] clusters into the model. In addition, we analyze more than 4,300 genomic and metagenomic sequences of HgcA to show that the evolution of this enzyme family has been marked by extensive horizontal gene transfer. A large diversity of HgcA is present in organisms that have not yet been cultured.

Methods

For experimental details see:

Cooper, C.J., Ovchinnikov, S., Zheng, K., Rush, K.W., Podar, M., Pavlopoulos, G., Kyrpides, N.C., Johs, A., Ragsdale, S.W., and Parks, J.M. Structure determination of the HgcAB complex using metagenome sequence data: insights into the mechanism of mercury methylation. *Commun. Biol.* 2020. DOI: 10.1038/s42003-020-1047-5.

MSA generation and coevolution analysis

The sequences of HgcA and HgcB from *D. desulfuricans* ND132[88] (UniProt IDs: F0JBF0 and F0JBF1, respectively) were selected for 3D structural modeling. In microbial genomes, *hgcB* is nearly always located immediately downstream of *hgcA*, which facilitated generation of the paired multiple sequence alignment. Initial alignments were generated by searching the UniProt20 database (2015_06) with *hhblits*[89] from HH-Suite[90] and then filtering the results with *hhfilter* to remove sequences with >90% identity and columns with more than 50% gaps. A hidden Markov model (HMM) was then generated from the alignment with *hmmbuild* from HMMER version 3.1b1 with default parameters, and *hmmsearch* was used to search a combined database consisting of JGI metagenomes (IMG/M)[87] and the UniRef100 database.[86] Filtering was performed to generate the final paired alignment. GREMLIN[91, 92] was used to perform the coevolution analysis and predict intra- and interdomain contacts. A single GREMLIN calculation was performed on the paired multiple sequence alignment. The GREMLIN output provides predicted contacts that are ranked based on the strength of the coevolution signal between residue pairs. These raw contacts were then normalized and reweighted according to a previously described model that estimates the contact prediction accuracy from the normalized GREMLIN scores, the number of sequences in the MSA, and the length of the query sequence.[6]

HHΔ calculation

hhsearch from HH-Suite was used to search the PDB70 database of hidden Markov models (HMMs) for homologous proteins with known structures using the HgcAB query HMM as input. For the resulting list of potential templates, HHΔ was calculated to determine if the multiple sequence alignment was closer to the query protein than a given structural homolog.[91]

Ab initio modeling

The approach used to generate the model has been described previously.[22] Briefly, individual domains were folded with the standard Rosetta ab initio structure prediction method using restraints derived from the coevolution analysis. For each domain, we generated 10,000 models with sigmoidal restraints, 10,000 models with sigmoidal restraints and bounded restraints (with bounded restraints applied only during the centroid stage), and 4,000 *map_align* models with sigmoidal and bounded restraints. The program *map_align*[22] identifies structural homologs by aligning contact maps predicted from coevolution analysis with contacts in experimentally determined structures, in this case a subset of the Protein Data Bank with a maximum of 30% mutual sequence identity.[93]

The first nine residues of HgcA were excluded from the model because they are not highly conserved. The last ten residues of HgcB were not included in initial modeling but were added after the complex was assembled. Models were ranked by the sum of their Rosetta energy[94] and restraint score (scaled by a factor of 3). A diverse set of 30 top-scoring models selected on the basis of their pairwise TMscore[95] was then used as input for iterative hybridization.[1] The RosettaScripts interface[96] was used for both the *map_align* models and for iterative hybridization.

Modeling of [4Fe-4S] clusters

Consistent with the expected Cys coordination patterns from other dicluster ferredoxins, such as that from *Clostridium acidurici* (PDB entry 2FDN),[97] preliminary de novo models of HgcB with coevolution restraints suggested that one [4Fe-4S] cluster is bound to Cys20, Cys23, Cys26, and Cys60 and another is bound to Cys50, Cys53, Cys56, and Cys30. Thus, after a preliminary model of the HgcAB complex was generated, additional restraints were included in subsequent hybrid modeling to enforce geometries consistent with cluster binding. The C-terminal tail of HgcB was also introduced at this step. All Cys restraints were generated on the basis of the 0.94 Å resolution

crystal structure of ferredoxin from *C. acidurici* (PDB entry 2FDN) and were the average values for the corresponding residues in each cluster. Harmonic distance restraints of 6.4 +/- 0.5 Å were applied to all pairs of Sγ atoms among the four cysteines coordinated to each [4Fe-4S] cluster. Harmonic angle restraints were applied to Cα-Cβ-Sγ angles in each Cys residue as follows: Cys20 and Cys50, 114.6 +/- 1 deg; Cys23 and Cys53, 116.9 +/- 1 deg; Cys26 and Cys56, 112.9 +/- 1 deg; Cys30 and Cys60, 108.9 +/- 1 deg. Circular harmonic restraints were applied to the C-Cα-Cβ-Sγ dihedrals in each cysteine as follows: Cys20 and Cys50, 56.1 +/- 2.3 deg; Cys23 and Cys53, -52.7 +/- 2.3 deg; Cys26 and Cys56, -71.6 +/- 2.3 deg; Cys30 and Cys60, 58.4 +/- 2.3 deg. Explicit [4Fe-4S] clusters were placed into the final model by aligning the Sγ atoms of cluster-binding cysteines of the model with those in 2FDN.

Modeling of the corrinoid cofactor

The specific corrinoid cofactor used by HgcA differs from organism to organism. For example, the corrinoid used by most species of *Geobacter* is 5-hydroxybenzimidazolyl cobamide. However, the cofactor used by ND132 is not known, so B_{12} was used. The cofactor was first placed in the binding pocket by superposing the CBD onto an X-ray structure of CFeSP. Polar residues in the CBD of CFeSP that interact with the B_{12} cofactor are conserved in HgcA. Thus, the following harmonic distance restraints were applied to facilitate cofactor binding in the HgcAB model: Thr60 (Oγ1)–B_{12} (N3B), 2.9 +/- 0.1 Å; Thr66 (Oγ1)–B_{12} (O4), 2.7 +/- 0.2 Å; Val91 (N)–B_{12} (O4), 3.0 +/- 0.05 Å; Ala153 (N)–B_{12} (O6R), 3.1 +/- 0.2 Å. Cys93 in HgcA was modeled as a chemically modified residue consisting of a coordinating bond between Sγ and the Co center in vitamin B_{12} with a harmonic distance restraint of 2.5 +/- 0.1 Å and a Cβ-Sγ-Co harmonic angle restraint of 108 +/- 5 degrees. We then generated 1,500 models with the Rosetta Relax application.[98] The model with the lowest Rosetta score was selected as the final model. The Dali web server[99] was used to identify structures in the PDB with folds that are similar to those of the HgcA and HgcB models. Figures were generated with PyMOL version 2.2.0.[100]

Phylogenetic analysis

HgcA sequences identified in UniRef100 and IMG/M included 296 sequences from genomes of isolated bacteria and archaea and from taxonomically assigned uncultured organisms (assembled genomes from single cells or metagenomes), as well as ~4,200 sequences (after filtering to a 90%

identity cutoff) identified in bulk metagenomes. The sequences were aligned with Muscle (v. 3.8.425)[101] in Geneious (version 10)[102] and the alignment trimmed to eliminate highly variable positions (<30% overall similarity). A phylogenetic tree was constructed using FastTree (v. 2.1.12)[103] and visualized in iTOL.[104]

Results

We cloned and expressed full-length HgcA from *D. desulfuricans* ND132 heterologously in *E. coli* as an N-terminal His-tagged construct (His-HgcA) (**Figure 2A** in **Appendix I**).[105] Similarly, HgcB was produced separately as a maltose-binding protein fusion construct (MBP-HgcB).

Electronic spectra of HgcA and HgcB

After purifying each protein, we obtained UV-visible spectra to confirm cofactor binding. The characteristic UV-visible peaks of dicyanocobalamin are 367, 540 and 580 nm.[106, 107] We obtained a spectrum from KCN and heat-treated His-HgcA (95 °C for 20 min) and compared it to that of 20 μM dicyanocobalamin dissolved in the same phosphate buffer (**Figure 2B**). Both spectra show the characteristic peaks of dicyanocobalamin, demonstrating that HgcA indeed binds cobalamin. Sodium dithionite (1 mM) was added to 12.5 μM HgcB (25 μM [4Fe-4S] cluster), quenching the absorbance in the 300-500 nm region, as is characteristic of reduced [4Fe-4S] cluster proteins (**Figure 2C**).

Lack of suitable templates for homology modeling

Structural models of HgcA published to date are limited to the core of the CBD.[7, 108] To determine whether including coevolutionary information is likely to provide more information for structural modeling of HgcA and HgcB than homology modeling, we searched a nonredundant subset of structures in the Protein Data Bank and calculated HHΔ for potential templates. HHΔ values less than 0.5 for a query and template sequence are generally considered to be good candidates for template-based modeling, whereas those with values greater than 0.5 are not. The lowest HHΔ value for the paired alignment of HgcA and HgcB is 0.77 (**Table 1** in **Appendix I**), with the top hit corresponding to an X-ray structure of the corrinoid iron-sulfur protein CFeSP (PDB entry 4DJD).[14] However, only the core of the CBD is covered by the template. No structures in the PDB were identified by *hhsearch* that could serve as templates for the TMD of HgcA. The lowest HHΔ

value for a template that covers HgcB is 0.92 for the Fe hydrogenase from *D. desulfuricans* (PDB entry 1HFE).[109]

Multiple sequence alignments and contact map predictions

To obtain a sufficient number of sequences for coevolution analysis, we searched a large master database comprising JGI metagenomes and the UniRef100 database for sequence homologs of HgcA and HgcB. Initial searches identified 7,505 and 19,317 putative HgcA and HgcB sequences, respectively. We then exploited co-occurrence and adjacency to generate a paired alignment of HgcA and HgcB. After pairing of HgcA and HgcB sequences based on whether two hits were from the same metagenomic contig, we obtained 3,025 sequences. We used 90% identity filtering to remove redundant sequences (2,432), but later reweighted by 80% identity to obtain the effective number of sequences (1,783). From the paired alignment, the estimated contact prediction accuracy is $N_f = seq/\sqrt{len} = 87.1$ for the 419 amino acids in HgcA and HgcB remaining after trimming regions at the N- and C-termini that are not well constrained by predicted contacts. This N_f value indicates that HgcA and HgcB are excellent candidates for structural modeling guided by coevolution-based contact restraints.

Structural modeling

Intra- and interdomain residue-residue contacts were predicted by performing a coevolution analysis of the HgcAB paired alignment. Surprisingly, the contact map includes very few predicted contacts between the two domains of HgcA (**Figure 3**). Gly33 is predicted to interact with Val186, and Leu32 is predicted to interact with Tyr189. In addition, Val173 and Thr174 are both predicted to interact with Glu179, but these residues are located near the boundary between the two domains. However, there is clear evidence for several contacts between the CBD of HgcA and HgcB.

CBD of HgcA

Rosetta modeling guided by coevolution analysis revealed that the core of the CBD of HgcA adopts a Rossmann fold with five β sheets, four major α-helices and two short helical regions (**Figure 4**). An additional α-helix is present near the N-terminus. A search of the Protein Data Bank with the Dali web server revealed several proteins with structural similarity to the CBD model (**Table 2**). As expected, the protein with the greatest structural similarity to the CBD of HgcA is CFeSP (PDB entry 2YCL, Z-score = 14.2). The sequence identity between the CBD of

HgcA (residues 15-166) and CFeSP (residues 291-445) is only 27%, but the binding pocket that accommodates the nucleotide tail of the cofactor is similar in the two proteins.[7] Besides the four conserved hydrogen bonds that were used as distance restraints (see Methods), the B_{12} cofactor forms hydrogen bonds with several other residues in the model (**Figure 4** and **Table 2**).

TMD of HgcA

The TMD consists of five TM helices, with helix 4 forming a central stalk that is mostly surrounded by helices 1, 2, 3 and 5 (**Figure 5**). Helices 1 and 2 are the longest, both consisting of 31 residues. Helix 5 includes 29 residues and helix 4 includes 24. Helix 3 is the shortest, comprising 21 residues. Based on the coevolution analysis, all adjacent pairs of helices in the model are predicted to be in contact with each other except for helices 1 and 5 (**Figure 5B**). A search of the Protein Data Bank with the Dali web server identified structural similarity between the TMD of HgcA and several membrane proteins (**Table 4).** Interestingly, the top hit is an X-ray structure of the homodimeric Mg^{2+} transporter MgtE from *Thermus thermophilus* (PDB entry 2YVX, Z-score = 6.8).[110]

HgcB

HgcB consists of an N-terminal core domain with a typical [4Fe-4S] ferredoxin fold[111] followed by an α-helical extension and a disordered tail at its C-terminus (**Figure 6**). The core domain of HgcB (residues 12-68) displays the same two-fold pseudosymmetry as the bacterial ferredoxin from *Clostridium acidurici*[97] and other ferredoxins. In addition, it is structurally similar to numerous proteins including heterodisulfide reductase, tungsten formylmethanofuran dehydrogenase subunit FwdA, photosystem I subunit PsaC, and adenylylsulfate reductase (**Table 5**). A similar α-helical extension is present in some ferredoxins, such as that from *Thauera aromatica*.[112] However, the additional disordered tail at its C-terminus appears to be unique to HgcB.

Cysteine residues 20, 23, 26 and 60 bind cluster A and residues 50, 53, 56, and 30 bind cluster B. The strictly conserved Cys73 in HgcB is located at the beginning of the α-helical extension and is located ~13 Å from the nearest Fe atom in cluster B in the model (**Figure 6B**). The number of cysteines and the total number of residues in the disordered tail vary among HgcB orthologs. Of

the 2432 sequences in the paired alignment, 1943 have at least one additional Cys located downstream of Cys73 and 1317 have two or more C-terminal cysteines. The majority of these sequences were obtained from metagenomes, so it is likely that some are truncated at their termini. Thus, these counts represent a lower bound for the number of cysteines located at or near the C-terminal tail of HgcB.

Assembly and analysis of the HgcAB complex

Using the top predicted interdomain contacts to guide docking of the individual domains together (**Figure 7A**), we generated a model of the HgcAB complex. Based on the ratio of the number of contacts in the model to those expected from the coevolution analysis given the number of sequences in the paired alignment and the GREMLIN score,[6] the estimated accuracy of the model, R_c, is 0.87. R_c values for native proteins range from 0.7 to 1.2. Thus, in general the HgcAB structural model fits the predicted contact set well (**Figure 7B**).

Interfacial residues

In the assembled complex (**Figure 8**), residues in the CBD of HgcA interact with the core of HgcB via several polar contacts: Gly96 (O)–Arg58 (NE), Gly132 (N)–Asn59 (OD1), Thr131 (OG1)–Asn59 (OD1), Arg136 (NH1)–Pro61 (O), Gly132 (O)–Ser25 (OG), Glu168 (OE2)–Lys2 (NZ), and Val (N)–Pro31 (O) (**Figure 9** and **Table 6**). Polar contacts between residues in the TMD of HgcA and HgcB include: Asn245 (O)–Arg5 (NH1), Arg250 (NH2)–Arg5 (O), Arg250 (N)–Asp8 (OD2), and Tyr303 (O)–Asp8 (N). The α-helical extension of HgcB interacts with TM helices 4 and 5 in HgcA, which protrude above the expected position of the membrane head groups. All contacts between the C-terminal extension and the TM helices of HgcA are nonpolar.

The distance between the closest Fe atom in cluster B and Co in the assembled model is 14.9 Å. The strictly conserved Cys73 in HgcB is located at the beginning of the C-terminal extension and is oriented away from the corrinoid in the CBD **(Figure 8**). The C-terminal cysteines in HgcB (Cys94 and Cys95) are located at the end of a long, disordered tail, which is likely to be highly flexible. Both Cys73 and the B_{12} cofactor are accessible by Cys94 and Cys95, suggesting a possible role of the cysteine pair in the transfer of Hg^{2+}, $[CH_3Hg]^+$, or both.

Oligomerization state

Several pieces of evidence suggested that HgcAB could function as a dimer of heterodimers, i.e., $(HgcAB)_2$: (i) Helices 1 and 5 in the TMD are not predicted to contact each other (**Figure 5**), which suggests that the TMD may not form a tight, cylindrical bundle but may instead be more open or splayed out and may interact with another protein. (ii) The closest structural homolog to the TMD model identified by the Dali server is a homodimeric Mg^{2+} transporter (PDB entry 2YVX).[110] (iii) There appears to be self-complementarity in the shape of the HgcAB subunit, particularly in the TMD. (iv) Three functionally important residues in HgcB, Cys73, Cys94 and Cys95 are all oriented away from the B_{12} cofactor in the HgcAB model (**Figure 8**), but these residues in one HgcB protomer would be oriented toward the corrinoid in the opposite HgcA protomer in a dimer of heterodimers model. (v) Some of the predicted contacts, particularly in the TMD, are relatively long in the model and could potentially be interpreted as inter-oligomeric contacts. We therefore explored this possibility by performing symmetric docking[113] of two copies of HgcAB using ambiguous restraints. However, we found that the inter-oligomeric contacts were all longer and therefore less favorable than those in the original HgcAB model (**Figure 7**). Thus, the present coevolution analysis appears to support a 1:1 rather than a 2:2 oligomerization state.

Phylogenetic analysis

In addition to providing input for coevolution analysis, the deep multiple sequence alignment obtained in this work enables an unprecedented phylogenetic analysis of HgcA diversity in nature. It has been shown previously that the phylogeny of HgcAB is not congruent with that of Bacteria and Archaea species, suggesting the genes have been horizontally transferred across the different microbial lineages.[11] The more than tenfold expansion of the number of available sequences based on more recent metagenomes and additional cultured organisms provides much deeper insight into the diversity of Bacteria and Archaea that we predict to be able to methylate mercury, in a variety of environments. Although HgcA sequences from methanogens appear to remain confined to a single major clade, the genes from important methylating bacteria such as Deltaproteobacteria and Firmicutes are distributed across three or four distinct clades, suggesting multiple horizontal gene transfer events followed by independent diversification (**Figure 11**). The various groups of HgcA also include sequences from a variety of cultured bacterial phyla (including Chloroflexi, Nitrospirae, Spirochaetes, Bacteroidetes) but also phyla with few or no cultured representatives

(e.g., Raymondbacteria, Saganbacteria, Lentispherae). Several archaeal phyla with no cultured representative also appear to include potential methylators, such as Heimdallarchaeota and Theionarchaea. Interestingly, distinct sequence clades composed of dozens of metagenomic sequences cannot be assigned to any specific microbial taxa, suggesting we still have much to learn about the diversity of bacteria and archaea that can methylate mercury.

Discussion

We have combined coevolution-based contact prediction and Rosetta modeling to generate a model of the HgcAB complex, which is responsible for Hg methylation in anaerobic microorganisms. This system is challenging to model because HgcA includes a transmembrane domain with no detectable sequence homology to any structurally characterized protein, and the complex consists of a unique heterodimeric structure in which the two domains of HgcA do not interact with each other but are instead bridged by interactions with HgcB. In addition, both proteins bind complex metal cofactors, which we have confirmed experimentally through heterologous expression and UV-visible spectroscopic characterization. These cofactors, vitamin B_{12} and two [4Fe-4S] clusters were incorporated into the model, which is consistent with available data from in vivo site-directed mutagenesis experiments targeting highly conserved residues in both HgcA and HgcB.[16]

Some of the predicted residue-residue contacts in the fully assembled model are longer than expected (**Figure 7**), suggesting that structural rearrangements (i.e., domain motions) may occur during catalysis.[114] The closest Fe atom from [4Fe-4S] cluster B is ~15 Å from the Co center in the B_{12} cofactor. However, it is likely that the CBD can move slightly closer to enable efficient electron transfer. Corrinoid-dependent enzymes with Rossmann domains often bind to $(\beta/\alpha)_8$ triosephosphate isomerase (TIM) barrel proteins to perform tightly controlled radical chemistry.[115] In addition, the CBD of the closest known homolog of HgcA, the corrinoid/iron-sulfur protein (CFeSP), is known to undergo large-scale conformational rearrangements, as revealed by X-ray co-crystal structures with its methyltransferase, a TIM barrel protein.[14] In the HgcAB model, the CBD is oriented toward the expected location of the membrane surface (**Figure 8**). Such a conformation would preclude the approach and binding of a relatively large TIM barrel protein,

suggesting that movement of the CBD would be required to accommodate a TIM barrel protein as a methyl donor.

The C-terminal tail of HgcB from *D. desulfuricans* ND132 includes a pair of cysteine residues (Cys94 and Cys95). Pairs of cysteines are commonly observed in proteins and enzymes involved in metal trafficking and detoxification, such as the proteins and enzymes encoded by the *mer* operon in Hg-resistant bacteria.[116] For example, the mercuric reductase (MerA), which catalyzes the reduction of Hg^{II} to Hg^{0}, includes two Cys residues at its C-terminus that acquire Hg^{II} and then transfer it to another pair of Cys residues in the active site. Whereas a double mutant of MerA in which both C-terminal Cys residues were substituted with Ala retained less than 0.1% of wild-type activity, a single Ala mutant maintained the same activity as the wild-type enzyme when an exogenous small-molecule thiol was present.[117] These findings suggest that when one of the Cys residues in the pair is replaced with Ala, a small molecule thiolate can substitute for the missing Cys to satisfy the valence of Hg^{II}. However, loss of both Cys residues completely eliminates the tether that binds and properly positions Hg^{II}, resulting in a major reduction in activity.

Formation of MeHg by HgcAB has been previously proposed to proceed through a multi-step reaction involving (i) reduction of the corrinoid cofactor to form a Co^{I} species, (ii) methylation of the Co^{I} center to form a CH_3-Co^{III} species, and (iii) methyl transfer to a Hg^{II} substrate to form $[CH_3Hg^{II}]^{+}$ **(Figure 12A)**.[7] The reduction step is presumed to be carried out by HgcB. The reduction potentials of the [4Fe-4S] clusters in HgcB and the corrinoid bound to HgcA have not been reported. However, parallels to CFeSP, in which a single [4Fe-4S] cluster serves a reductive activation role,[118, 119] would put the $Co^{II/I}$ couple below -500 mV versus SHE. Loss of the axial Cys93 ligand is expected upon reduction to Co^{I} to give a four-coordinate complex, which is supported by DFT calculations.[84] Subsequent oxidative addition of the methyl group and coordination of Cys93 from HgcA by the reduced corrinoid form the proposed active species for mercury methylation. The Hg substrate that is then methylated by HgcA to produce methylmercury is not known but is assumed to be a Hg^{II} bis(thiolato) species.

Our model provides insight into how HgcAB orchestrates the transfer and transformation of Hg. Specifically, we propose that Cys94 and Cys95 from HgcB acquire Hg^{II} (from an unknown source)

and deliver it to the corrinoid cofactor for methylation (**Figure 12B**). The Hg methylation step has been proposed to proceed through either a methyl anion transfer or radical ligand exchange pathway.[7, 84] A relativistic DFT study found that the latter pathway is energetically more favorable when spin-orbit effects were taken into account.[120] Assuming that the reaction proceeds through radical ligand exchange, a crosslinked HgcB-Cys94/95(Sγ)–Co^{III}–(Sγ)Cys93-HgcA intermediate would be formed. Reduction of the Co center to Co^{I} would then release both thiolate ligands and allow the C-terminal tail to deliver $[CH_3Hg]^+$ to Cys73 from HgcB. Either of the C-terminal cysteines (Cys94/95) could facilitate delivery of the $[CH_3Hg]^+$ product, as only a single Cys thiolate is required to bind this species. An exogenous thiolate, possibly a cysteine residue on a protein, would then displace Cys73 to liberate $[CH_3Hg]^+$ from HgcB, completing the reaction cycle. We expect that this structural model of HgcAB will facilitate the development of hypotheses addressing more detailed structural and functional questions that can then be tested experimentally.

Appendix I

Tables

Table 1. HHsearch results for HgcAB

PDB ID	Coverage	Prob (%)	HHΔ	Description
4DJD_C	0.37	100	0.77	5-methyltetrahydrofolate corrinoid/Fe-S protein
2H9A_A	0.37	100	0.78	CO dehydrogenase/acetyl-CoA synthase, Fe-S protein
1HFE_L	0.22	99.5	0.92	Fe-only hydrogenase
3GYX_B	0.19	99.4	0.93	adenylylsulfate reductase
1DWL_A	0.13	99.3	0.93	ferredoxin I
1F2G_A	0.13	99.3	0.93	ferredoxin II
1JNR_B	0.19	99.3	0.93	adenylylsulfate reductase
1XER_A	0.07	99.3	0.93	ferredoxin
4ID8_A	0.14	99.3	0.93	putative ferredoxin
1IQZ_A	0.16	99.3	0.93	ferredoxin

Table 2. Top ten Dali results for the CBD of HgcA versus PDB25

PDB ID	Z-score	RMSD (Å)	N_res	%ID	Description
2YCL_A	14.2	2.7	442	27	CO dehydrogenase corrinoid/iron-sulfur protein
3D0K_B	5.9	3.1	293	9	putative poly(3-hydroxybutyrate) depolymerase LpqC
2DST_A	5.5	3.5	122	16	hypothetical protein TTHA1544
3B48_F	5.1	3.2	135	5	uncharacterized protein
3GDW_B	4.9	3.6	138	5	sigma-54 interaction domain protein
2XDQ_A	4.8	3.8	425	7	light-independent protochlorophyllide reductase
3LFH_B	4.7	3.4	144	11	phosphotransferase system
1CVR_A	4.6	3.1	433	7	gingipain R
6HSW_A	4.6	3.5	422	14	carbohydrate esterase family 15 domain protein
5ELM_B	4.4	3.5	236	16	Asp/Glu racemase family protein

Table 3. Interactions between the B12 cofactor and residues in the CBD of HgcA

B_{12} atom	CBD atom	Distance (Å)
N3B	Thr60 (OG1)	2.9[a]
N3B	Ala61 (N)	3.3
O4	Thr66 (OG1)	3.0[a]
O2	Thr66 (OG1)	3.2
N52	Gly88 (O)	2.8
O51	Asn90 (N)	2.7
O3	Asn90 (ND2)	2.8
O4	Val91 (N)	3.0[a]
O5	Trp92 (N)	2.9
O39	Lys97 (NZ)	2.9
O7R	Gln127 (O)	3.0
O6R	Ala153 (N)	3.2[a]
O8R	Ala153 (N)	3.1

[a] Interactions in CFeSP that was used as distance restraints for docking B_{12} into the HgcAB model

Table 4. Top ten Dali results for the TMD of HgcA versus PDB25

PDB ID	Z-score	RMSD (Å)	N_res	%ID	Description
2YVX_A	6.8	3.8	442	12	Mg^{2+} transporter MgtE
4TQ4_D	6.5	4.9	290	9	prenyltransferase
6IU4_A	5.6	3.7	225	10	iron transporter VIT1
5YCK_A	5.6	7.3	449	5	multidrug efflux transporter
6FV7_A	5.5	7.9	421	6	multidrug resistance transporter Aq_128
5EDL_A	5.4	4.5	197	6	S-component of ECF transporter, Putative HMP/thiamine permease protein YkoE
3FNB_A	5.4	4.9	374	6	acylaminoacyl peptidase, hydrolase
5WEO_A	5.3	5.5	989	5	glutamate receptor 2, voltage-dependent calcium channel
4IDN_B	5.3	4.7	423	3	atlastin-1, hydrolase
2XZE_A	5.1	2.9	141	7	stam-binding protein, hydrolase/transport

Table 5. Top ten Dali results for HgcB versus PDB25

PDB ID	Z-score	RMSD (Å)	N_res	%ID	Description
5ODC_A	8.0	6.2	653	27	heterodisulfide reductase
5T5M_F	7.9	3.4	342	31	tungsten formylmethanofuran dehydrogenase
5T5I_P	7.8	1.3	81	39	tungsten formylmethanofuran dehydrogenase
5OY0_c	7.8	2.9	81	27	photosystem I trimer
3GYX_B	7.7	4.3	166	33	adenylylsulfate reductase
5C4I_E	6.9	2.3	312	24	oxalate oxidoreductase subunit alpha
1SIZ_A	6.7	2.1	66	26	ferredoxin
3J16_B	6.4	6.9	608	24	ribosomal protein
1XER_A	6.2	1.4	103	35	ferredoxin
1IQZ_A	5.6	2.2	81	18	ferredoxin

Table 6. Polar interactions between HgcA and HgcB in the HgcAB model

	HgcA	HgcB
CBD-HgcB (core)	Gly96 (O)	Arg58 (NE)
	Thr131 (OG1)	Asn59 (OD1)
	Gly132 (O)	Ser25 (OG)
	Gly132 (N)	Asn59 (OD1)
	Arg136 (NH1)	Pro61 (O)
	Glu168 (OE2)	Lys2 (NZ)
	Val173 (N)	Pro31 (O)
TMD-HgcB (tail)	Asn245 (O)	Arg5 (NH1)
	Arg250 (NH2)	Arg5 (O)
	Arg250 (N)	Asp8 (OD2)
	Tyr303 (O)	Asp8 (N)

Figures

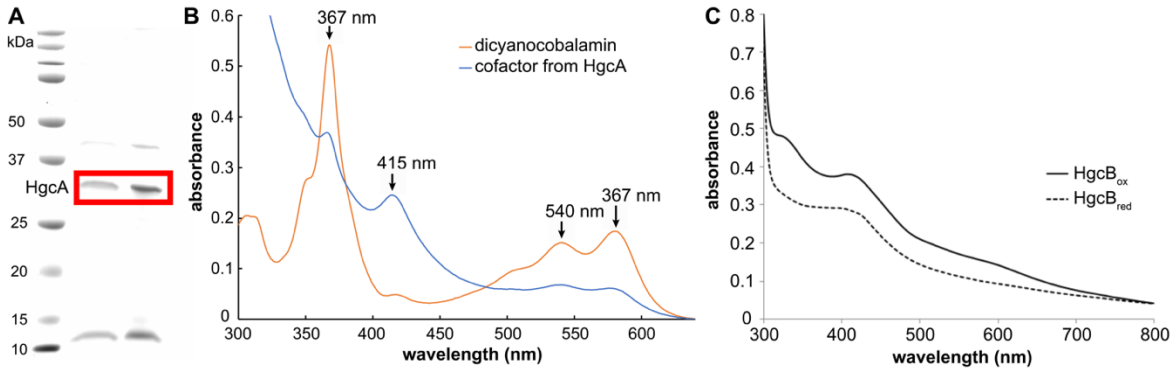


Figure 2. Purification and UV-visible spectroscopy of HgcA and HgcB. (**A**) SDS-PAGE gel of purified HgcA. The bands enclosed in the red rectangle are HgcA in elution buffer and after buffer exchange, respectively, as verified by western blot analysis using an antibody against the His-tag. (**B**) UV-visible spectrum of dicyanocobalamin (orange) and cofactor extracted from purified, His-tagged HgcA by heating to 95 °C with KCN (blue). HgcA was dissolved in phosphate buffer (50 mM K_2HPO_4, 100 mM NaCl, 10% glycerol, 2 mM BME, 10 mM imidazole, pH 7.4). (**C**) UV-visible spectrum of oxidized, as-isolated MBP-HgcB ($HgcB_{ox}$) and MBP-HgcB after reduction with sodium dithionite ($HgcB_{red}$). See ref 105 for experimental methods.

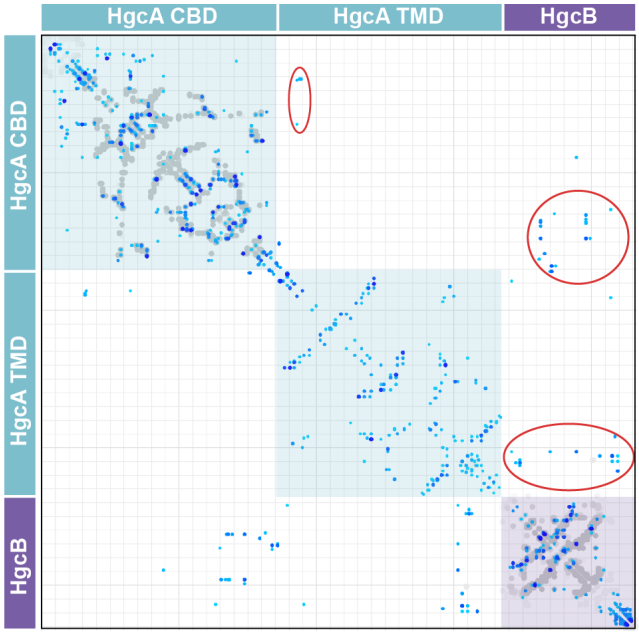


Figure 3. HgcAB contact map predicted from coevolution analysis of the paired multiple sequence alignment. Contacts are shown in shades of blue (darker blue = higher probability) and contacts from X-ray structures of homologs are shown in grey. Individual domains are labeled and interdomain contacts are circled in red.

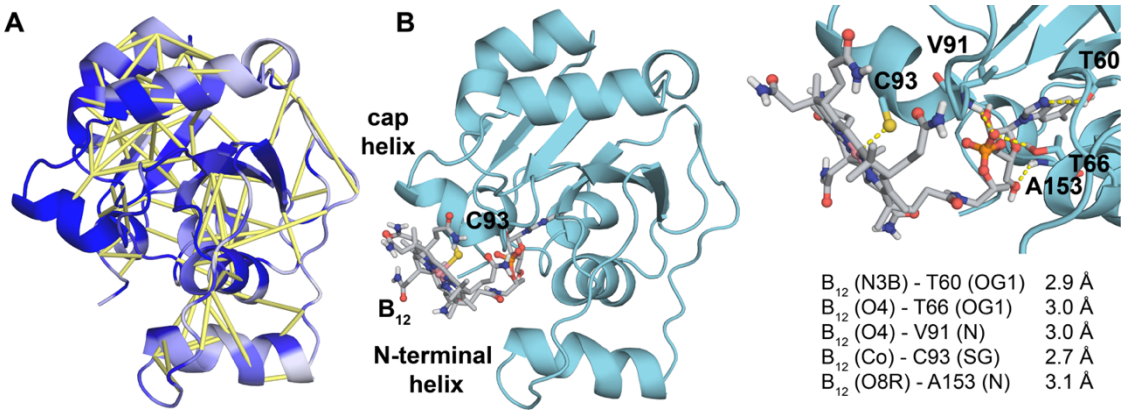


Figure 4. Model of the corrinoid binding domain of HgcA. (**A**) predicted contacts shown as yellow bars and residues colored according to sequence conservation (dark blue = highest) and (**B**) model including the B_{12} cofactor. Key residues and distances are shown.

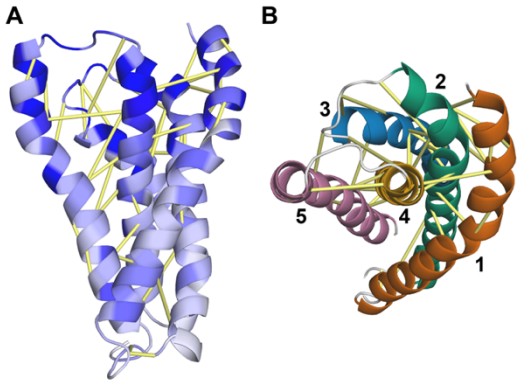


Figure 5. Model of the transmembrane domain of HgcA. (**A**) Predicted contacts are shown as yellow bars and residues are colored according to sequence conservation (dark blue = highest). (**B**) Model rotated ~90 degrees forward to show a 'top' view from the cytoplasmic side. Each helix is shown in a different color.

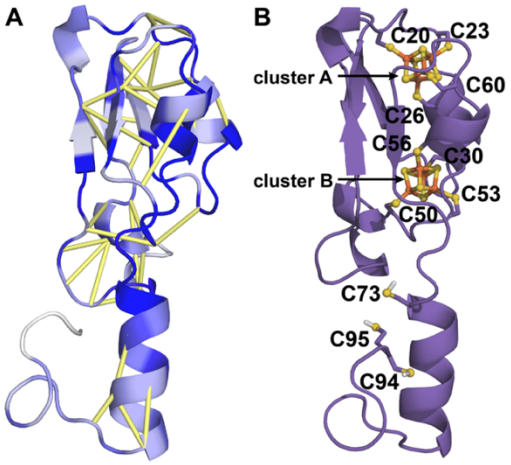


Figure 6. Model of HgcB. (**A**) Predicted contacts are shown as yellow bars and residues colored according to sequence conservation (dark blue = highest). (**B**) Location of conserved Cys residues and incorporation of [4Fe-4S] clusters.

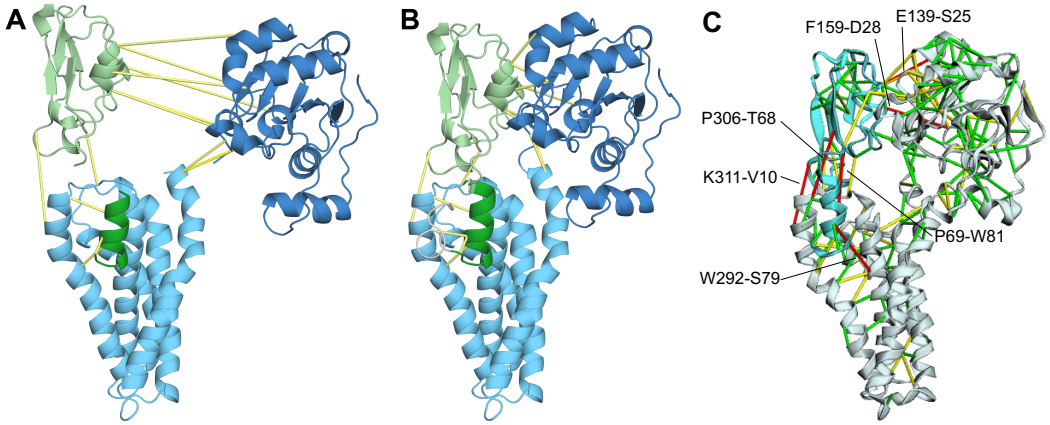


Figure 7. Assembly of the HgcAB model. (**A**) Top interdomain contacts (yellow) in the HgcAB complex predicted from the coevolution analysis. (**B**) Interdomain contacts in the assembled HgcAB complex after docking. *Colors*: Dark blue, CBD of HgcA; light blue, TMD of HgcA; light green, core of HgcB; dark green, C-terminal extension of HgcB. (**C**) Predicted contacts in the HgcAB model color coded by Cα-Cα distance. *Colors*: green (<5 Å), yellow (5-10 Å), red (>10 Å). Residues with distances >10 Å are labeled. All labels refer to the residue from HgcA followed by the residue from HgcB except P69-W81, for which both residues are from HgcB.

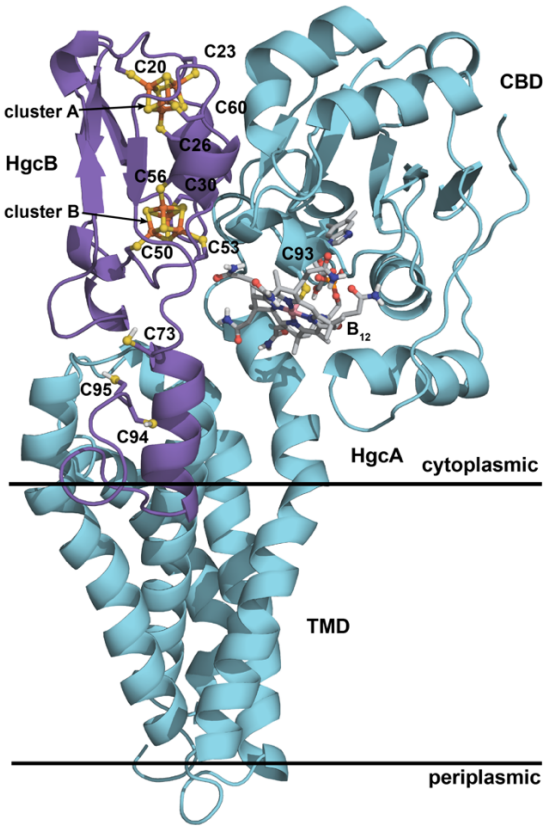


Figure 8. Model of the HgcAB complex.

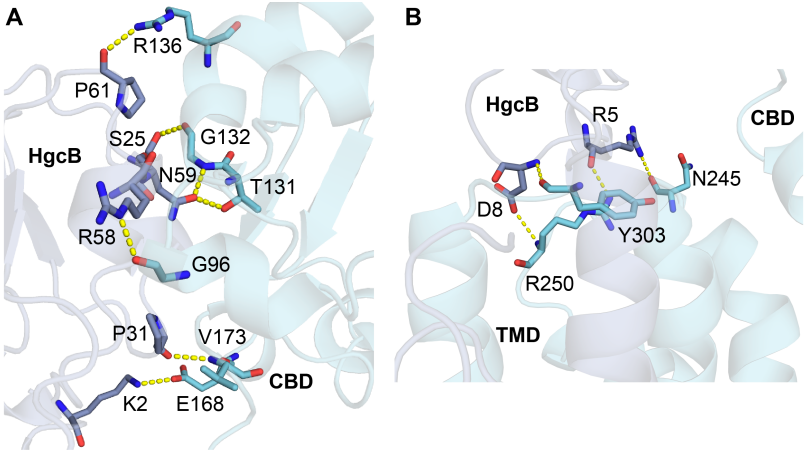


Figure 9. Polar contacts between (**A**) CBD and HgcB and (**B**) TMD and HgcB.

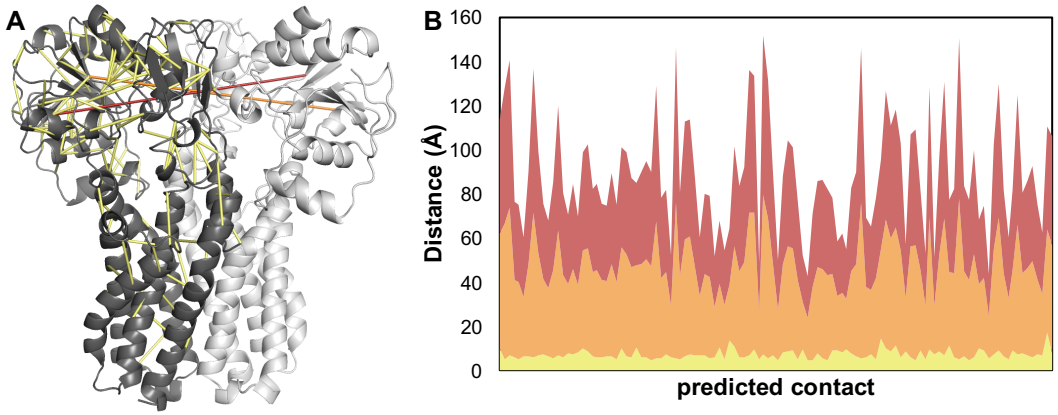


Figure 10. Oligimerization state of HgcAB. **(A)** Dimer-of-heterodimers model generated by applying ambiguous restraints during symmetric docking of two HgcAB heterodimers. One HgcAB heterodimer is shown in dark gray and the other is shown in light gray. Predicted residue-residue contacts within a single HgcAB dimer are shown as yellow lines. Only contacts with probability >0.99 are shown. Representative examples of contacts that could potentially be satisfied between two separate heterodimers are shown as orange and red lines. **(B)** Stacked bar chart of possible contacts. The contacts are ordered by predicted probability with the highest on the left. In all cases the contacts within a single HgcAB heterodimer were shorter and therefore more favorable than inter-heterodimeric restraints, suggesting that the coevolution analysis supports a 1:1 HgcAB model rather than a 2:2 $(HgcAB)_2$ model.

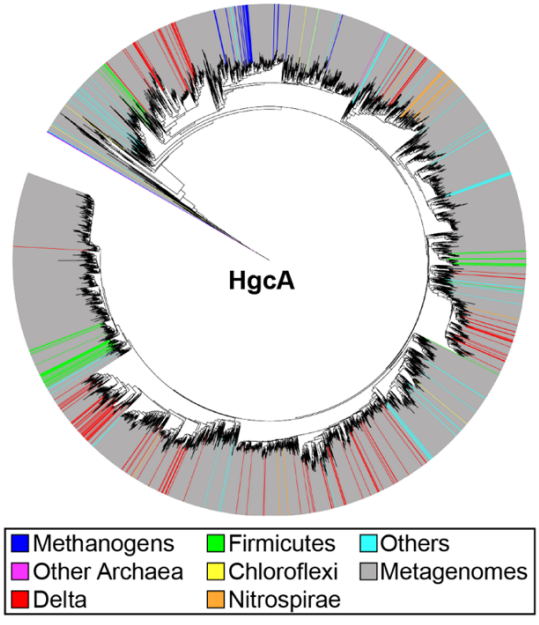


Figure 11. Phylogenetic tree of HgcA. See Methods for details.

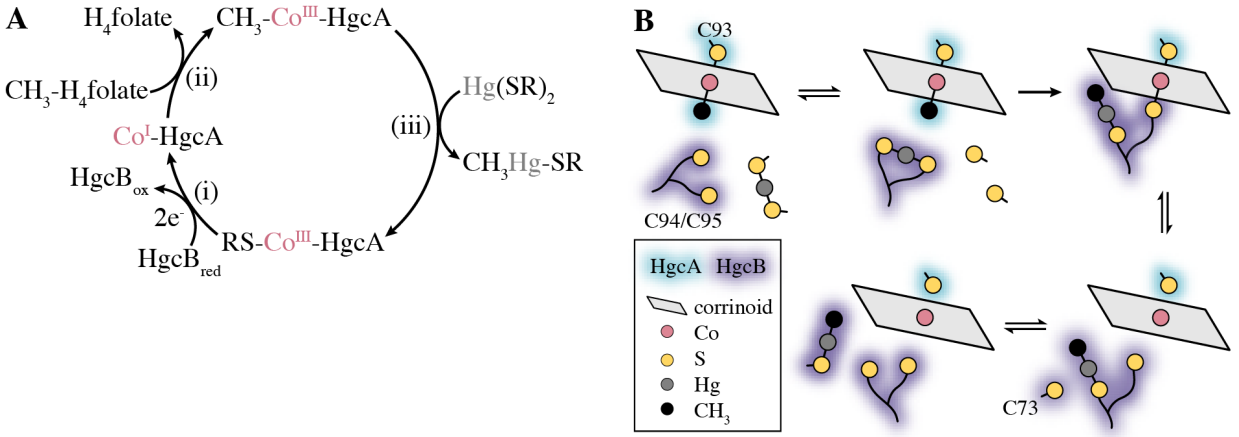


Figure 12. Mechanistic insights. (**A**) Proposed Hg methylation cycle from refs [7], [84] and [120]. (**B**) Proposed pathway for Hg^{II} acquisition, methylation, and $[CH_3Hg]^+$ release.

CHAPTER II

UNDERSTANDING THE DYNAMICS UNDERLYING HOW A SINGLE AMINO ACID MUTATION RELIEVES INHIBITION OF INOSINE MONOPHOSPHATE DEHYDROGENASE (IMPDH)

Text and figures are taken from the following:

Close, D.M.[#], Cooper, C.J.[#], Wang, X., Chirania, P., Gupta, M., Ossyra, J.R., Giannone, R.J., Engle, N., Tschaplinski, T.J., Smith, J.C., Hedstrom, L., Parks, J.M., and Michener, J.K. Horizontal transfer of a pathway for coumarate catabolism unexpectedly inhibits purine nucleotide biosynthesis. *Mol. Microbiol.* 2019, 112, 1784-1797.

(#) These authors contributed equally to this work.

D.M.C. and J.K.M performed the laboratory evolution experiments; L.H. performed the kinetic assays; C.J.C. and J.M.P. generated the homology models and performed the docking; J.R.O. set up MD simulations and M.G. parameterized the covalent intermediate; C.J.C, and M.G. analyzed the MD simulations and generated figures; J.K.M., J.M.P., D.M.C., and C.J.C. prepared the manuscript with input from all other authors.

Abstract

A microbe's ecological niche and biotechnological utility are determined by its specific set of co-evolved metabolic pathways. The acquisition of new pathways, through horizontal gene transfer or genetic engineering, can have unpredictable consequences. Here we show that two different pathways for coumarate catabolism failed to function when initially transferred into *Escherichia coli*. Using laboratory evolution, we elucidated the factors limiting activity of the newly acquired pathways and the modifications required to overcome these limitations. Both pathways required host mutations to enable effective growth with coumarate, but the necessary mutations differed. In one case, a pathway intermediate inhibited purine nucleotide biosynthesis, and this inhibition was relieved by single amino acid replacements in IMP dehydrogenase. A strain that natively contains this coumarate catabolism pathway, *Acinetobacter baumannii*, is resistant to inhibition by the relevant intermediate, suggesting that natural pathway transfers have faced and overcome similar challenges. Molecular dynamics simulation of the wild type and a representative single-residue mutant provide insight into the structural and dynamic changes that relieve inhibition. These results demonstrate how deleterious interactions can limit pathway transfer, that these interactions can be traced to specific molecular interactions between host and pathway, and how evolution or engineering can alleviate these limitations.

Introduction

Microbes can use a wide variety of compounds as carbon and energy sources. Expanding the breadth of compounds that a strain can catabolize can allow access to new environmental niches or enable engineered microbes to use new feedstocks. Correspondingly, catabolic pathways are frequently transferred between strains, either in nature through horizontal gene transfer (HGT) or in the laboratory through metabolic engineering.[23, 24] However, newly-acquired pathways often fail to function effectively in their new host.[25] In these cases, productive use of a new pathway may require post-transfer refinement to optimize expression and minimize deleterious interactions.[26, 27] The pathway activity immediately following transfer may be very different from the potential activity after optimization, complicating predictions about engineering or HGT.

We have explored this issue using pathways for catabolism of lignin-derived aromatic compounds, since these pathways are widespread in nature,[28] are often transferred by HGT,[29] have biotechnological applications,[121] and involve challenging biochemistry.[122] We previously constructed strains of *E. coli* that grow with the model lignin-derived compounds protocatechuate (PCA) and 4-hydroxybenzoate (4-HB) as sole sources of carbon and energy using the 3,4-cleavage pathway for protocatechuate catabolism from *Pseudomonas putida* and a 4-hydroxybenzoate 3-monooxygenase from *P. putida* or *Paenibacillus* sp. JJ-1b.[30, 31] Introduction of the relevant catabolic pathways was not sufficient to enable rapid growth with either carbon source. We then used experimental evolution to select for strains with improved growth. By resequencing the evolved variants and reconstructing mutations in the parental strains, we identified causal mutations that improved function of the heterologous pathway.

In this work, we extended those pathways to allow growth with a model phenylpropanoid, coumarate. There are two known oxidative routes for coumarate catabolism, differing in their specific reaction chemistry and resulting intermediates (**Figure 13** in **Appendix II**). These pathways are exemplified by the *hca* pathway from *Acinetobacter baylyi* ADP1[123] and the *cou* pathway from *Rhodococcus jostii.*[124] Both pathways begin by conjugating the phenylpropanoid substrate to coenzyme A. The *hca* pathway then uses a retro-aldol reaction to produce an intermediate benzaldehyde derivative, while the *cou* pathway uses a hydrolytic retro-Claisen reaction to produce the benzoate derivative directly. Since these two phenylpropanoid pathways

use different biochemistry and intermediates, their interactions with the host may also differ substantially.[125] Identifying the likeliest pairing of host and pathway, either for engineering or HGT, will depend on understanding the specific challenges imposed by each potential pathway and the mechanisms to overcome these challenges available to the host.

Using a combination of engineering and evolution, we constructed and optimized both representative pathways for phenylpropanoid catabolism in *E. coli*. We show that pathway activity is initially limited due to pathway-specific molecular interactions that can readily be alleviated through point mutations to the host. Similar compensatory mechanisms are present in a strain that natively contains the appropriate pathway. Molecular dynamics simulations of the wild-type and mutant enzymes demonstrate how subtle modifications to the enzyme distant from the active site can relieve inhibition while preserving catalysis. Identifying and alleviating the specific molecular interactions between an engineered metabolic pathway and its heterologous host will aid our efforts to rapidly engineer metabolic capabilities.

Methods

For experimental details see:

Close, D.M., Cooper, C.J., Wang, X., Chirania, P., Gupta, M., Ossyra, J.R., Giannone, R.J., Engle, N., Tschaplinski, T.J., Smith, J.C., Hedstrom, L., Parks, J.M., and Michener, J.K. Horizontal transfer of a pathway for coumarate catabolism unexpectedly inhibits purine nucleotide biosynthesis. *Mol. Microbiol.* 2019, 112, 1784-1797.

Homology modeling

To generate models of IMPDH from *E. coli* (accession number P0ADG7) with the corresponding cofactors and substrates, HHpred[126] was used to search the Protein Data Bank for suitable structural templates. Five templates were chosen based on their similarity to the query sequence, inclusion of cofactors and substrates, or both (**Table 7** in **Appendix II**). The sequences were aligned with MAFFT (L-INS-i)[127] (**Figure 14**) and homology models were generated using RosettaCM [128] with fragment files obtained from the Robetta web server.[129, 130] The top-scoring model was used for docking.

The flap containing the catalytic dyad (R401 and Y402) is not resolved in most X-ray crystal structures of IMPDH but is present in the structure of the phosphate-bound "apoenzyme" from *Bacillus anthracis* (PDB entry 3TSB).[131] Thus, we used this structure as a template to generate homology models of *E. coli* IMPDH in the closed conformation and the top scoring model was selected to generate a model of the C305-XMP* covalent intermediate for molecular dynamics simulations. Sequences were aligned with MAFFT (L-INS-i) (**Figure 15**).

Ligand docking

Structure files in mol2 format for IMP (ZINC04228242), NAD^+ (ZINC08214766), and 4-hydroxybenzaldehyde (ZINC00156709) were obtained from http://zinc.docking.org.[132] RosettaLigand[133] was used to dock 4-hydroxybenzaldehyde into the active site of the IMPDH model following a previously described protocol.[34] Top binding poses were ranked on the basis of their '*interface_delta*' score in Rosetta energy units.

Molecular dynamics simulations

Initial coordinates for XMP were extracted from the crystal structure of *B. anthracis* IMPDH complexed with XMP (PDB entry 3TSD), and the covalently bound C305-XMP* complex was generated for chain A using the Molefacture plugin in VMD.[134] The force field toolkit (ffTK) plugin in VMD was used to generate CHARMM-compatible force field parameters for C305-XMP*. Gaussian09[135] was used to perform geometry optimizations, compute Hessian matrices, and calculate water interaction energies of the C305-XMP* fragment. NAMD 2.11 was used for charge, bond, angle, and dihedral optimization.[136]

The CHARMM36 force field[137] and TIP3P water model[138] were used to describe the protein and solvent, respectively. Each system (wild-type and mutant) was solvated in a periodic box of 168 Å × 168 Å × 104 Å and 0.15 M KCl ions were added using CHARMM-GUI,[139] resulting in a system of ~274,000 atoms. All-atom molecular dynamics (MD) simulations were performed using the OpenMM 7.0 package[140] with GPU acceleration using CUDA 7.2. Ten thousand steps of energy minimization were performed to eliminate clashes, followed by equilibration in the NPT ensemble at 310 K. Temperature was maintained using the Langevin thermostat with a damping coefficient of 1 ps^{-1}. To enable a 5-fs time step, which was used in all simulations, all bond lengths were constrained to their equilibrium distances and the masses of hydrogens were repartitioned to

the parent heavy atoms.[140] Five separate runs of 100 ns each were performed for both the wild type and mutant IMPDH and the last 50 ns of each run was used for analysis.

Results

Combining engineering and evolution enabled coumarate catabolism

We designed and synthesized two constructs containing genes for phenylpropanoid import and degradation, each of which converts coumarate into 4-hydroxybenzoate (**Figure 13**).[141] Each pathway was introduced into *E. coli* strains, JME38 and JME50, that had previously been engineered to grow with 4-HB using *pobA* and *praI*, respectively.[31] None of the engineered strains acquired the immediate ability to grow with coumarate as the sole source of carbon and energy.

To understand the factors preventing pathway function, we used experimental evolution to select for strains with the ability to catabolize coumarate. Three replicate cultures of each engineered strain were propagated in minimal medium containing 1 g/L coumarate (~6.1 mM). After 300 generations, individual mutants were isolated from each population and characterized for growth with protocatechuate (PCA), 4-HB, coumarate, and caffeate. Representative isolates were chosen for each replicate population for further characterization. All isolates could grow with PCA and coumarate, though growth with caffeate and 4-HB varied between replicates.[141]

Genome resequencing and reconstruction identified causal mutations

The genomes of the selected isolates were resequenced to identify new mutations. Several of the mutations have previously been described for their effects on catabolism of 4-HB, such as synonymous mutations to the gene encoding the 4-hydroxybenzoate monooxygenase *pob.A*.[31] Among the strains with the *hca* pathway, five of the six isolates had additional mutations to the native gene *guaB*, encoding inosine monophosphate (IMP) dehydrogenase (IMPDH), and to the intergenic region between *hcaB* and *hcaC* in the engineered pathway. The exception was JME96, which had a mutation to *rpoS*, encoding the RNA polymerase sigma factor σ^{38}, that is expected to be highly pleiotropic.[142]

In the strains with the *cou* pathway, the acquired mutations were less consistent across replicates, with several mutations to genes that are expected to be pleiotropic. However, parallel mutations were observed in JME106 and JME109, with mutations to both *couL* and *nadR*. The mutations to

couL, which encodes the CoA ligase, were coding mutations, L192R and S134Y. NadR is involved in both regulation and catalysis for NAD salvage.[143] One of the mutations to *nadR* led to a frameshift that precisely removed the C-terminal ribosylnicotinamide kinase (RNK) domain, which converts N-ribosylnicotinamide into β-nicotinamide mononucleotide during NAD salvage.[143] Similarly, the second *nadR* mutation also occurred in the RNK domain. The physiological consequences of these mutations are unclear.

To test the causality of the identified mutations, we reconstructed representative mutations in the engineered parental strains. We assumed that parallel mutations to a given gene produced similar effects, and therefore only tested one representative mutation (*e.g.* D243G in *guaB*). Two mutations, to *pobA* and *hcaABCK,* were necessary for growth with coumarate in JME64, while a third mutation to *guaB* significantly increased growth.[141] Similarly, mutations to *pobA*, *couLHTMNO*, and *nadR* were all required for growth with coumarate using the *cou* pathway in JME65.

Parallelism of mutations within replicates of a pathway, but divergence between pathways, strongly suggests that the mutations are specific to a particular pathway. To test this hypothesis, we replaced the *hca* pathway in JME131 with either the wild-type or evolved *cou* pathways. Neither strain was able to grow with coumarate as the sole source of carbon and energy.

Inhibitory crosstalk between engineered and native pathways limits function

A mutation to *guaB* was necessary for growth with coumarate using the *hca* pathway. IMPDH, encoded by *guaB*, converts inosine monophosphate (IMP) to xanthosine monophosphate (XMP) with the reduction of NAD^+ during guanine nucleotide biosynthesis.[144] Five independent amino acid replacements in IMPDH were identified: A48V, D243G, G330D, L364Q, and P482L. IMPDH uses different conformations to catalyze each step of the catalytic cycle: an open conformation for hydride transfer that produces a covalent intermediate with the catalytic C305 (E-XMP*) and a closed conformation for hydrolysis of the E-XMP* (**Figure 16**). We generated homology models of wild-type *E. coli* IMPDH in both the closed and open conformations. The mutations are distant from each other and from the active site, with no obvious effect on catalysis (**Figure 17**).

To understand the consequences of these mutations, we measured metabolite levels in the parent and engineered strains during growth with coumarate. Consistent with our genetic analysis above, we chose to focus on the D243G mutation. Compared to the D243G *guaB* mutant (JME131), the strain with wild type *guaB* (JME129) showed higher levels of inosine nucleotides. We hypothesized that growth with coumarate led to inhibition of IMPDH and accumulation of IMP, and that this inhibition was relieved in the *guaB* mutants. To determine whether inhibition of nucleotide biosynthesis limited growth with coumarate, we supplemented the growth medium with guanosine. Addition of guanosine increased growth with coumarate in a strain with the wild-type IMPDH, but not the mutant.

Mutations to *guaB* improved growth with the *hca* pathway but not with the *cou* pathway. The *hca* pathway produces an intermediate, 4-hydroxybenzaldehyde, that is not present in the *cou* pathway (**Figure 13**). To test whether this intermediate was responsible for the inhibition of IMPDH, we grew strains containing WT and mutant IMPDH in varying concentrations of 4-hydroxybenzaldehyde.[141] Both strains were inhibited by high concentrations of 4-hydroxybenzaldehyde, but the mutation to *guaB* decreased inhibition.

Next, we purified WT and mutant IMPDH and measured inhibition *in vitro* with 4-hydroxybenzaldehyde. This compound is a weak inhibitor of WT *Ec*IMPDH, with a $K_{i,app}$ of 320 ± 20 μM. Introduction of the D243G replacement had little effect on catalytic activity but increased the $K_{i,app}$ to 1250 ± 50 μM, indicating a substantial reduction of inhibition in the mutant. The *hca* pathway that we used came from *A. baylyi* ADP1, and we hypothesized that the native IMPDH of this strain would have faced similar selective pressures to minimize inhibition by 4-hydroxybenzaldehyde. As a surrogate, we tested the IMPDH of *A. baumannii*, since this strain contains a homologous *hca* pathway (83-94% amino acid identity) and IMPDH (91% amino acid identity). As predicted, the *A. baumannii* IMPDH has a $K_{i,app}$ of 720 ± 30 μM, substantially higher than that of wild-type *Ec*IMPDH.

Subtle changes in enzyme dynamics relieve inhibition while maintaining catalysis

To gain insight into possible mechanisms of inhibition by 4-hydroxybenzaldehyde, we measured enzyme activity at varying inhibitor concentrations. 4-hydroxybenzaldehyde is an uncompetitive

inhibitor with respect to IMP and a noncompetitive (mixed) inhibitor with respect to NAD^+ for both wild type and D243G *Ec*IMPDH.[141] Similar patterns of inhibition have been observed for compounds that bind in the NAD^+ site.[131] In addition, we computationally docked 4-hydroxybenzaldehyde to a model of wild-type IMPDH in the open conformation of the apoenzyme as well as to IMP-bound and IMP/NAD^+-bound states. In both the apoenzyme and IMP-bound models, the majority of the top poses of 4-hydroxybenzaldehyde were found to occupy the NAD^+ binding site, approximately 20 Å from D243 (**Figure 18** and **Table 8**). Therefore, to relieve inhibition by 4-hydroxybenzaldehyde, the D243G substitution would need to perturb the structure or dynamics of the distant active site.

Understanding the mechanism of this perturbation required additional analysis. The hydrolysis of E-XMP* is the slow step in the IMPDH reaction, so E-XMP* is the predominant enzyme complex.[144] Therefore, a decrease in the affinity of 4-hydroxybenzaldehyde for E-XMP* can account for resistance to inhibition. Hydrolysis of E-XMP* requires a conformational change wherein a mobile protein flap folds into the cofactor binding site. We assessed the effect of the D243G mutation on the active site by performing MD simulations of wild-type and D243G mutant IMPDH in the covalently bound E-XMP* state using the closed conformation model, since the flap is disordered in crystal structures of the open conformation. In the simulations of the wild-type *E. coli* IMPDH, D243 forms stable hydrogen bonds with the side chains of K87 and R219 and also with the backbone of V220 (**Figure 19A**). In the absence of this hydrogen bonding network in the mutant (**Figure 19B**), G243 adopts two different conformations, one that resembles the wild type in which G243 is close to but not interacting with K87, R219, and V220, and another in which G243 is positioned farther away from these residues when a new hydrogen bond is formed with Q272 (**Figure 19C**). However, it was not obvious how these local changes around the mutation site propagate to the active site, which is located on the opposite side of the β-barrel.

To identify changes in protein dynamics resulting from the D243G mutation, we calculated root-mean-squared fluctuations (RMSFs) for chain A in both the wild-type and mutant. In both systems, high RMSFs were observed over the entire flap region **(Figure 20**). Upon inspection of specific interactions of flap residues, we found that the mutation leads to changes in hydrogen-bonding interactions with other residues on chain A or the adjacent chain D (**Figure 21** and **Figure 22**).

These interactions resulted in reorientation of a loop on the flap that could alter inhibitor binding. R386 interacts with D50 and D410 interacts with H432 on chain D more frequently in the wild type than in the mutant. Interestingly, other interactions between flap residues are observed more frequently (i.e. R386-E418, S399-D410, S405-D410) in the mutant. S405 and D410 also have additional hydrogen bonding interactions with residues on chain D that differ between the wild-type and mutant. S405 forms a hydrogen bond with N26 on chain D more frequently in the mutant than the wild type. In contrast, D410 forms a hydrogen bond with R439 on chain D more frequently in the wild type than the mutant. In addition, the covalent intermediate also forms a hydrogen bond with the side chain of E415 more frequently in the mutant than the wild type. Despite these changes in flap conformations and dynamics, the catalytic dyad remains in close proximity to the covalent intermediate, poised for catalysis (**Figure 23**).

RMSF analysis also revealed that helix α2 (residues 76-89) and helix α8 (residues 230-241) fluctuate more in the mutant than in the wild type (**Figure 20**). Helix α8 is downstream of the mutation site (**Figure 24A**). Therefore, the higher fluctuations of this helix and the adjacent helix α2 in the mutant are likely due to the loss of the hydrogen bonding network formed by D243 with K87, R219, and V220 as well as the formation of a new hydrogen bond between G243 and Q272 in the mutant. The N-terminal end of helix α2 and the C-terminal end of helix α8 are located near the NAD^+ binding site, which is also the predicted binding site for 4-hydroxybenzaldehyde based on docking to the open conformation model (**Figure 18**). Therefore, these perturbations to helices α2 and α8 represent another mechanism for the D243G mutation to affect inhibitor binding.

To understand the mechanism by which helix dynamics could affect inhibitor binding, we combined these MD results with our previous docking studies. In six of the top 10 docking poses, the phenolic hydrogen of the inhibitor forms hydrogen-bonding interactions with either the side chain or backbone of D248. The carbonyl oxygen of the inhibitor also interacts with the side chain of S250 (**Figure 24B**). D248 and S250 are located on the β-sheet (β11) downstream of the mutation site and are in close proximity to helices α2 and α8 as well as the loop on the flap that showed different interactions in the wild-type and mutant simulations. Thus, changes in the structure and dynamics of these regions around the NAD^+ binding site likely disrupt inhibitor access and binding.

Discussion

In this work, we have recapitulated the process of HGT and demonstrated the necessity for host adaptations to accommodate the *hca* pathway in both *E. coli* and *A. baumannii*. We identified a novel interaction between the newly introduced pathway and the endogenous metabolism, as well as the physiological and biochemical consequences of this interaction. Finally, we demonstrated how single point mutations to an essential host protein alter its conformational dynamics to prevent binding of the novel inhibitor while still preserving catalysis.

Highly similar *hca* pathways are present in various beta- and gamma-proteobacteria. Further HGT of this pathway would require either a host with an IMPDH homolog that is resistant to inhibition by 4-hydroxybenzaldehyde, or post-transfer selection for mutations that relieve inhibition. Understanding these types of limitations on HGT, and the mechanisms by which organisms evolve to avoid them, will aid in our ability to predict and manipulate horizontal gene transfer.[26, 27]

In combination, our results suggest that introduction of the *hca* pathway allowed only limited growth with coumarate because accumulation of 4-hydroxybenzaldehyde inhibited the native *E. coli* IMPDH. This inhibitory crosstalk results in nucleotide starvation and impairs growth and phenylpropanoid catabolism. Mutations to *guaB* prevent inhibition by 4-hydroxybenzaldehyde and allow growth with coumarate. There is no *a priori* reason to expect that a pathway for degradation of an aromatic compound would interact with a native pathway for nucleotide biosynthesis. Phenolic amides such as feruloyl amide have been shown to inhibit a different step in nucleotide biosynthesis,[145] but neither the substrate nor products of coumarate degradation are toxic at the relevant concentrations.[30, 31] These types of inhibitory cross-talk are likely to be common with heterologous engineered metabolic pathways, though they are rarely identified and alleviated.[125, 146, 147]

In particular, inhibition of microbial growth by aldehydes is commonly observed, though the mechanisms of toxicity can rarely be traced to a specific interaction.[148-150] Enzymatic pathways have frequently evolved to limit the release of free aldehydes, for example through enzymatic channeling.[151] It is unclear whether channeling between HcaA and HcaB limits the release of free aldehydes in either the native or heterologous hosts. Mutations that increase tolerance to free aldehydes generally do so either by increasing export of the toxic compound or by performing

redox chemistry to remove the aldehyde functionality.[152] In this work, we have shown an example of aldehyde toxicity that acts through a single protein and can be relieved by point mutations to the associated gene. For the D243G mutant, biochemical assays revealed mixed inhibition that was relieved through mutation. Other examples of nonspecific toxicity may prove to be similarly specific when characterized fully.

MD simulations provided insight into how a single amino acid substitution distant from the active site could relieve inhibition while maintaining catalysis. The mutation is located at the N-terminal end of β11, which is near the NAD^+ binding site where the inhibitor was predicted to bind based on docking calculations. The simulations showed local changes in the hydrogen bonding networks at the mutation site, which led to changes in the dynamics of the catalytic flap and helices α2 and α8 near the inhibitor binding site. In addition, the catalytic dyad showed only minor perturbations and remained poised for catalysis.

Across the replicate populations, many mutations were highly pleiotropic, including large insertions and deletions flanked by insertion sequences as well as mutations to core transcriptional machinery such as *rho* and *rpoB*. Duplications frequently spanned the insertion sites for engineered operons, suggesting that expression of the heterologous genes was limiting. By comparing across replicates, we were able to identify a set of point mutations that allowed growth with coumarate as the sole source of carbon and energy. However, a reconstructed strain containing these mutations does not grow as quickly with coumarate as the evolved isolates, suggesting that some of the remaining mutations provided additional fitness benefits.[141]

The two 4-HB monooxygenases, *praI* and *pobA*, are 60% identical at the nucleotide level, and the associated enzymes have 54% amino acid identity. We previously demonstrated that these enzymes required different optimization solutions to enable growth with 4-HB.[31] In contrast, in this experiment, the evolutionary solutions were very similar. Even in the optimized strain, JME131, the growth rate with coumarate was lower than the growth rate with PCA. We hypothesize that the conversion of coumarate into 4-HB is the rate-limiting step, and that the conversion of 4-HB into PCA by either enzyme was sufficient under these circumstances.

Multiple mutations were identified in the heterologous *cou* and *hca* pathways. In the *hca* pathway, these mutations served to increase expression of the pathway, either through pathway duplication or by intergenic mutations that affected translation, specifically increasing expression of the CoA ligase HcaC. The *cou* pathway mutations were coding mutations to a single gene, the *couL* that encodes a CoA ligase, and decrease expression of that enzyme.[141] These differential evolutionary responses could arise from different initial expression levels of the two CoA ligases, for example due to the placement of *couL* at the beginning of an operon and *hcaC* at the end. Further biochemical analysis will be required to precisely identify the consequences of these mutations.

We have described the use of experimental evolution to identify and alleviate deleterious interactions between engineered metabolic pathways for coumarate catabolism and native pathways for nucleotide biosynthesis and cofactor salvage. Many engineered pathways place a substantial burden on the production host yet understanding and accommodating these interactions remains challenging. Evolution can simplify this optimization process by directly selecting for mutations that eliminate the inhibition. As we did with *guaB*, researchers can then work backwards from the evolutionary solutions to understand the factors that were initially limiting productivity and the biochemical solutions to overcome those problems. By solving more problems of this sort, we will develop design rules for future forward engineering of metabolic pathways and better predictions of the likelihood of pathway transfer by HGT.

Appendix II

Tables

Table 7. Templates used to model the open and closed conformations of *E. coli* IMPDH

State	PDB ID	Organism	Resolution (Å)	E-value	% ID	Cofactors
Open	4X3Z	*Vibrio cholerae*	1.62	3.3e-34	86	NAD, XMP
	1ZFJ	*Streptococcus pyogenes*	1.90	9.8e-55	56	IMP
	5AHN	*Pseudomonas aeruginosa*	1.65	4.8e-59	66	IMP
	2CU0	*Pyrococcus horikoshii*	2.10	2.9e-47	49	XMP
	1VRD	*Thermotoga maritima*	2.18	8.3e-47	56	n/a
Closed	3TSB	*Bacillus anthracis*	2.60	N/A	55	PO_4

Table 8. Binding energies for the top five poses obtained from docking 4-hydroxybenzaldehyde to the apoenzyme, IMP-bound, and IMP/NAD^+-bound states of IMPDH

***interface_delta* (Rosetta energy units)**		
apoenzyme	**IMP-bound**	**IMP/NAD^+-bound**
-11.5	-12.4	-10.3
-11.0	-12.2	-9.4
-10.6	-12.1	-9.3
-10.6	-12.1	-8.9
-10.5	-11.4	-8.7

Figures

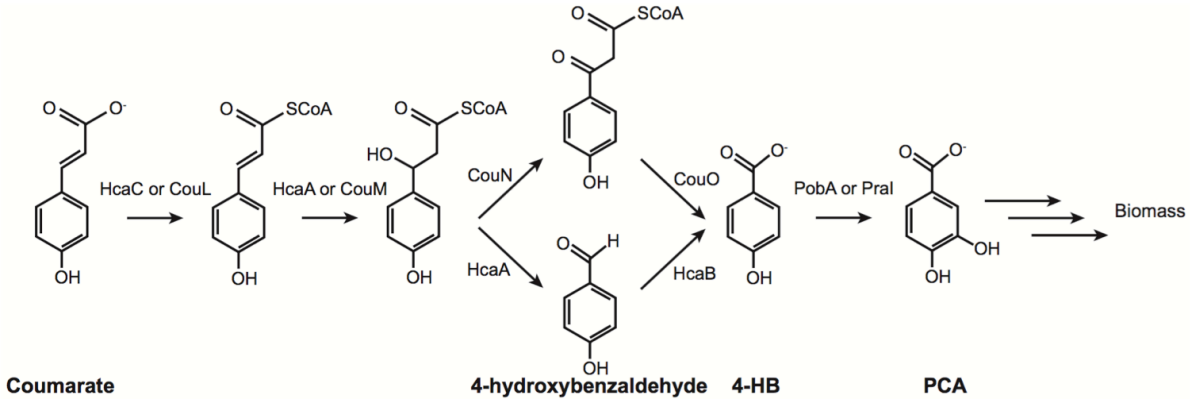


Figure 13. Two routes to convert the phenylpropanoid coumarate to 4-HB (the *hcaABC* pathway from *A. baylyi* ADP1 and the *couLMNO* pathway from *R. jostii*). 4-HB is then oxidized to PCA. For simplicity, cofactors and the resulting acetyl-CoA are not shown.

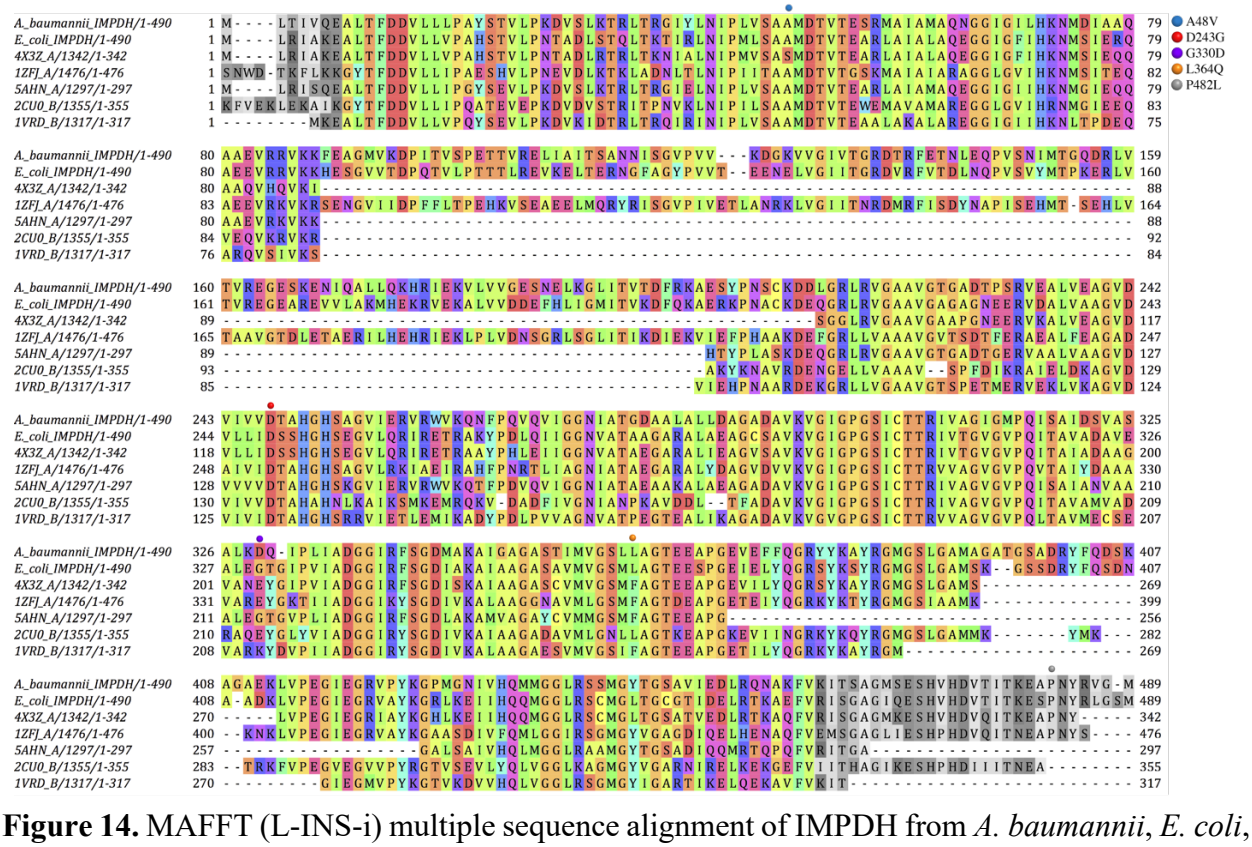


Figure 14. MAFFT (L-INS-i) multiple sequence alignment of IMPDH from *A. baumannii*, *E. coli*, and multiple template sequences used for structural modeling of the open conformation of *E. coli* IMPDH. Selected residues (gray) were trimmed at the N- and C-termini and were not included in the models.

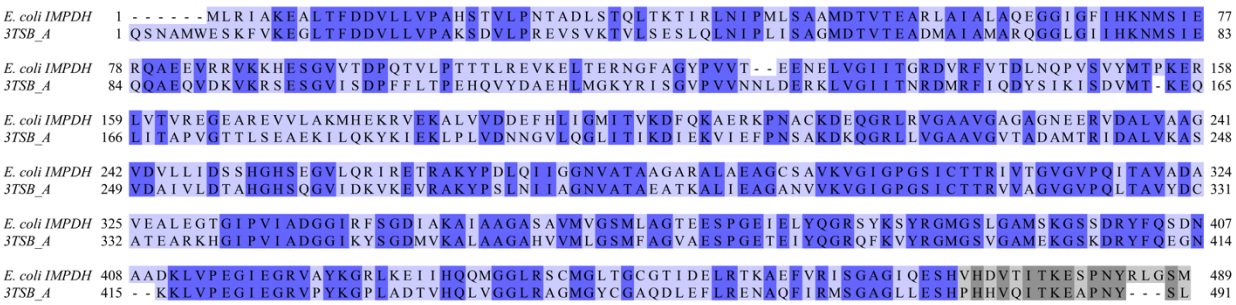


Figure 15. MAFFT (L-INS-i) alignment of *E. coli* IMPDH with the *B. anthracis* template sequence of the closed conformation that contains the catalytic flap (PDB entry 3TSB). Selected residues (gray) were trimmed at the C-terminus and were not included in the models.

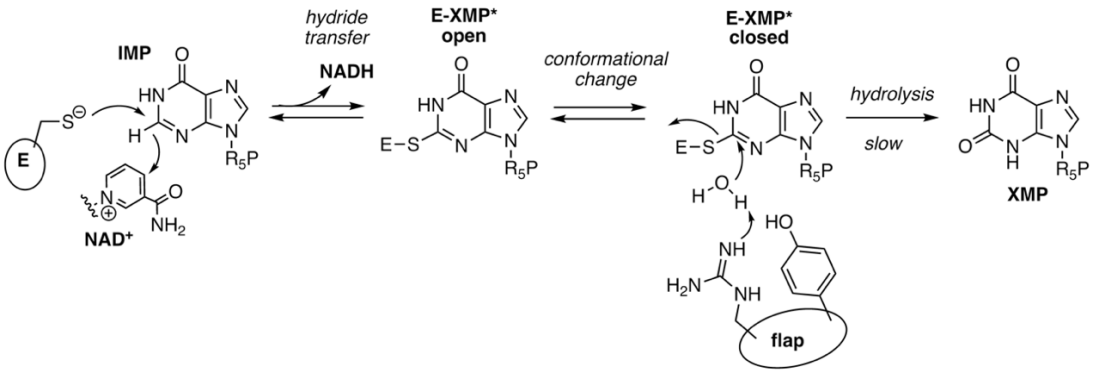


Figure 16. IMPDH mechanism. IMPDH transitions between two conformations, an open conformation for rapid substrate binding and hydride transfer and a closed conformation for the slower hydrolysis step. In the closed conformation, a mobile protein flap folds into the cofactor binding site.

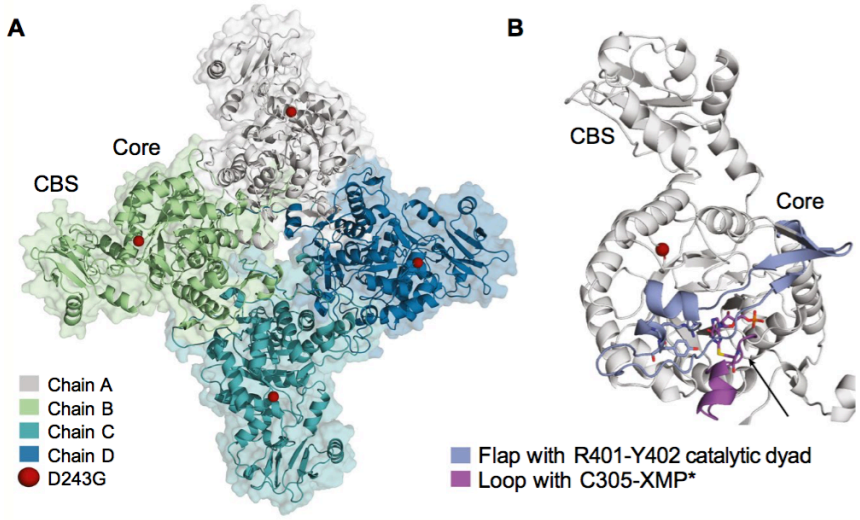


Figure 17. Beneficial mutations to IMPDH are distant from the active site. (**A**) Structural model of *E. coli* IMPDH colored by chain. (**B**) Chain A of IMPDH in the closed conformation, highlighting the loop containing C305-XMP* and the active site flap containing the R401-Y402 catalytic dyad.

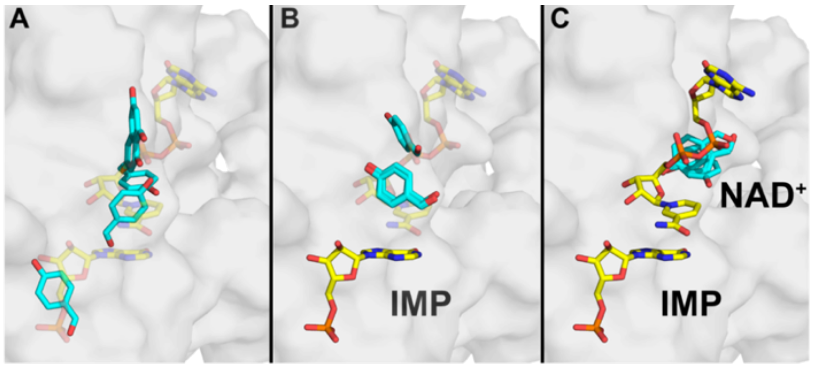


Figure 18. Top 5 predicted docking poses for 4-hydroxybenzaldehyde to the open conformation model of IMPDH in various enzyme states. (**A**) Apoenzyme, (**B**) IMP-bound, and (**C**) IMP/NAD^+-bound IMPDH. The top five docking poses are shown in each case. 4-hydroxybenzaldehyde carbons are shown in cyan. All other carbons are shown in yellow. Molecules in transparent representation are shown for reference but were not included in the docking. All hydrogens are omitted for clarity.

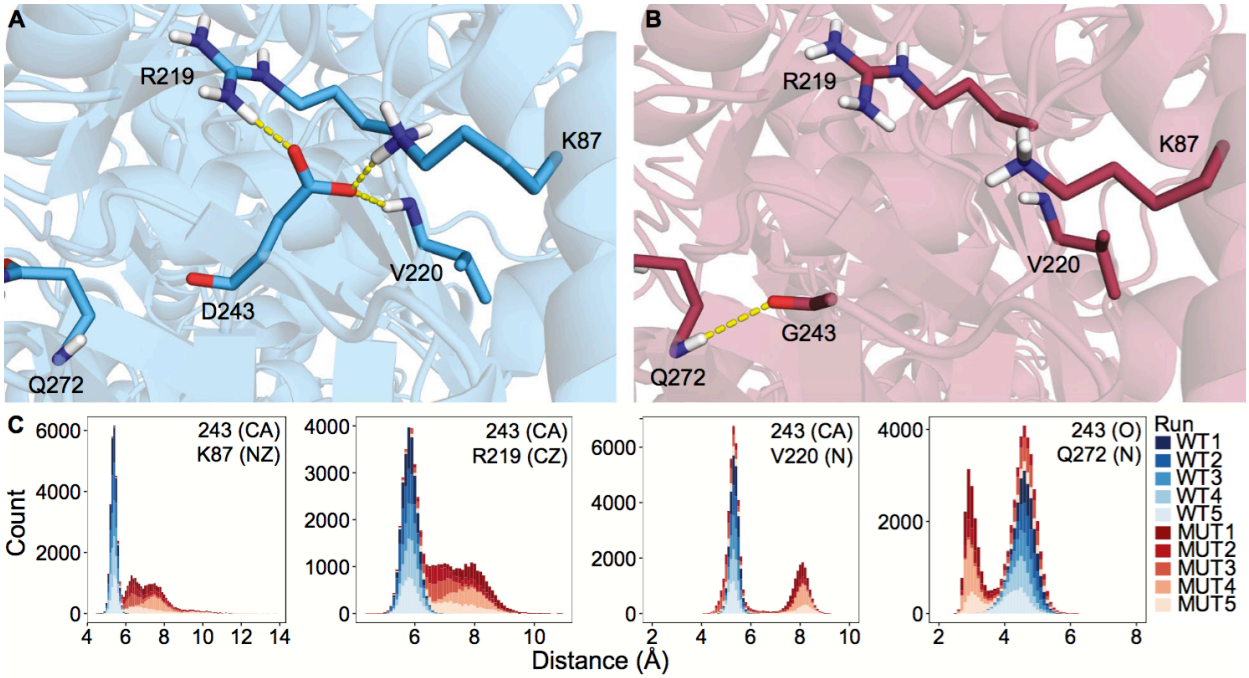


Figure 19. A beneficial mutation to IMPDH affects enzyme structural dynamics. Hydrogen bond network around the D243G mutation site for (**A**) wild-type and (**B**) mutant IMPDH from MD simulations. (**C**) Heavy atom distance distributions from five independent simulations.

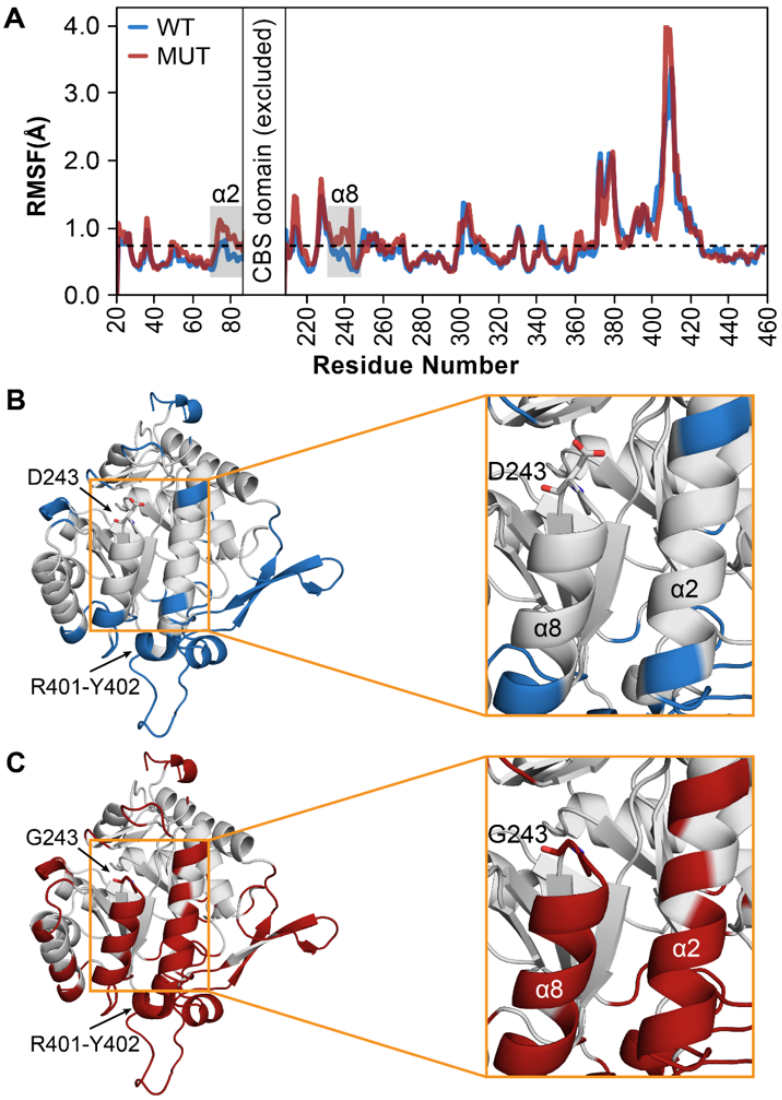


Figure 20. RMSF analysis of wild-type and mutant IMPDH. (**A**) RMSF of the chain A core domain of wild-type and mutant IMPDH from MD simulations. CBS domains and termini are highly dynamic and were excluded. The helix residues are indicated with grey boxes. Residues with RMSF values > 0.75 Å are shown on the model for (**B**) wild-type and (**C**) mutant.

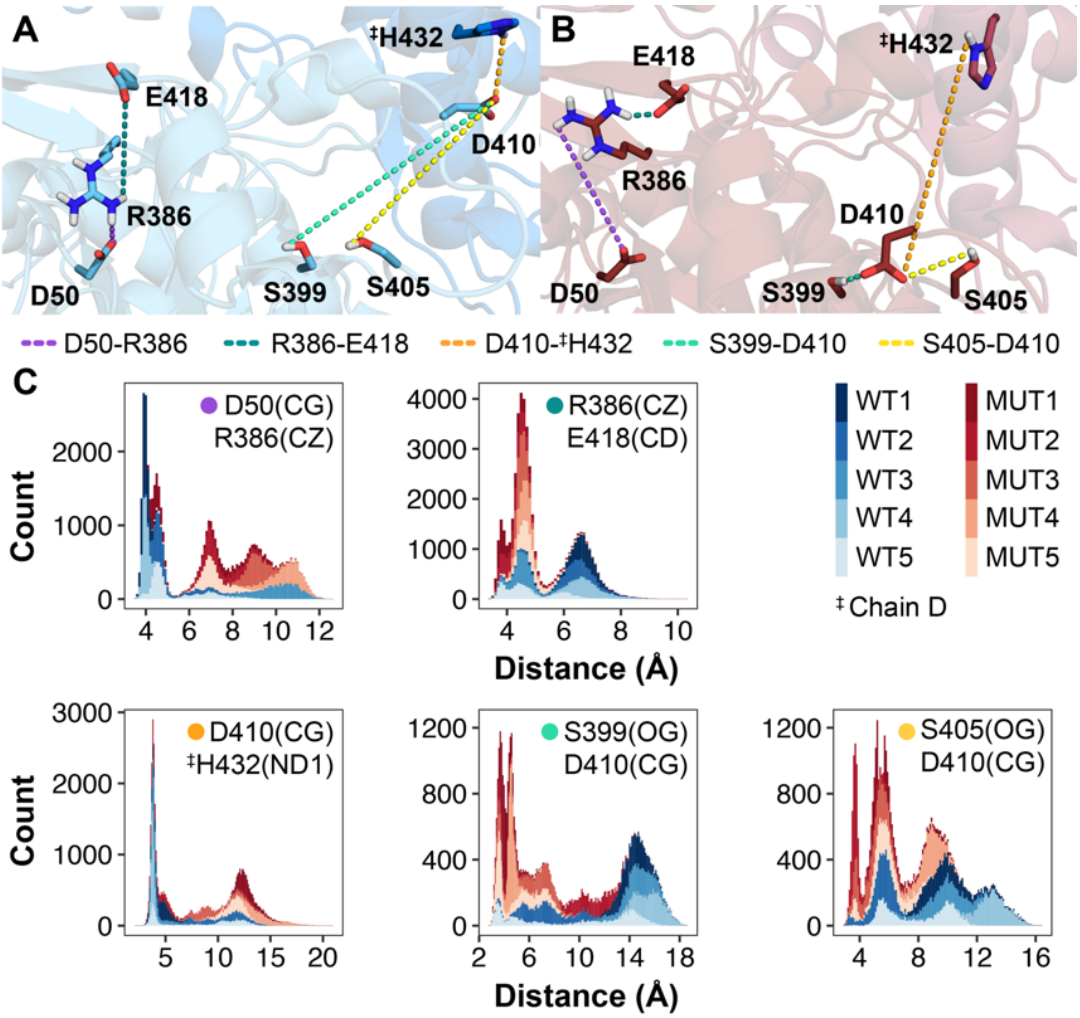


Figure 21. The D243G mutation alters the conformation of several residues on the catalytic flap. Hydrogen bond network around the catalytic flap for (**A**) wild-type and (**B**) mutant IMPDH from MD simulations. (**C**) Heavy atom distance distributions from the individual, independent simulation runs of the closed conformation model are shown.

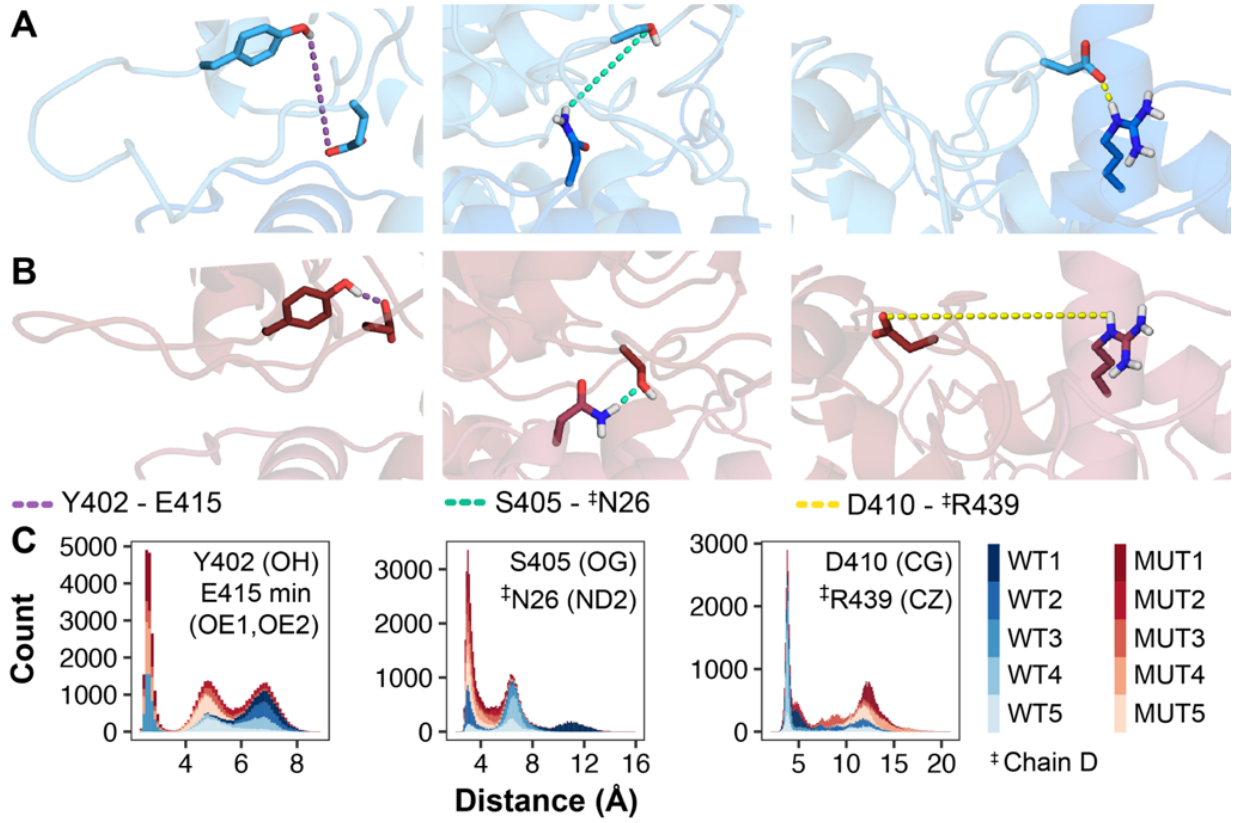


Figure 22. Changes in the conformation of residues on the catalytic flap in MD simulations of (**A**) wild type and (**B**) mutant IMPDH in the closed conformation. (**C**) Heavy atom distance distributions for residues near the catalytic dyad and C305-XMP* loop of chain A that have different interactions with other chain A residues or with residues on chain D between the wild-type and mutant.

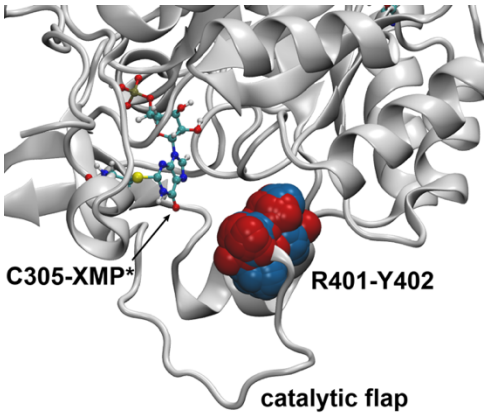


Figure 23. C-alpha conformations of the catalytic dyad in the wild-type (blue) and mutant (red).

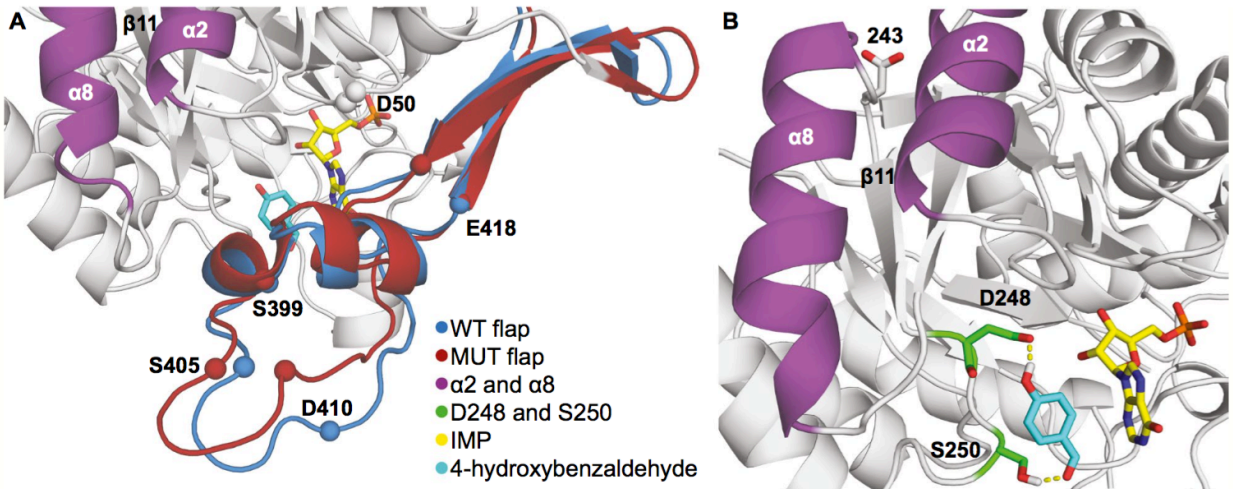


Figure 24. Summary of wild-type and mutant IMPDH simulations. (**A**) Selected snapshots of the flap from MD simulations of wild-type and mutant IMPDH in the closed conformation. Cα atoms of key residues whose interactions differ between wild type and mutant simulations (D50, R386, S399, S405, D410, E418) are shown as spheres and labeled. A selected docking pose is shown for 4-hydroxybenzadehyde in the IMP-bound open conformation after superposition with the closed conformation snapshots. (**B**) Selected docking pose showing 4-hydroxybenzaldehyde interactions with residues D248 and S250. Flap residues were omitted as they were not resolved in the templates used to model the open conformation.

CHAPTER III

MACHINE LEARNING-BASED PREDICTION OF ENZYME SUBSTRATE SCOPE: APPLICATION TO BACTERIAL NITRILASES

Text and figures are taken from the following:

Mou, Z.[#], Eakes, J[#]., Cooper, C.J., Foster, C.M., Standaert, R.F., Podar, M., Doktycz, M.J., Parks, J.M. Structure-based prediction of enzyme substrate scope with machine learning: Application to bacterial nitrilases. *In review.* DOI: https://doi.org/10.22541/au.158888180.03951231

[#] These authors contributed equally to this work.

Experiments were performed by J.E., C.M., M.J.D.; C.J.C. and J.M.P. built structural models; Z.M. performed ligand docking; C.J.C. and Z.M. calculated descriptors, performed machine learning analysis, and generated figures; Z.M., C.J.C., and J.M.P prepared the manuscript with input from all other authors.

Abstract

Predicting the range of substrates accepted by an enzyme from its amino acid sequence is challenging. Although sequence- and structure-based annotation approaches are often accurate for predicting broad categories of substrate specificity, they generally cannot predict which specific molecules will be accepted as substrates for a given enzyme, particularly within a class of closely related molecules. Combining targeted experimental activity data with structural modeling, ligand docking, and physicochemical properties of proteins and ligands with various machine learning models provides complementary information that can lead to accurate predictions of substrate scope for related enzymes. Here we describe such an approach that can predict the substrate scope of bacterial nitrilases, which catalyze the hydrolysis of nitrile compounds to the corresponding carboxylic acids and ammonia. Each of the four machine learning models (logistic regression, random forest, gradient-boosted decision trees, and support vector machines) performed similarly (average ROC = 0.9, average accuracy = ~82%) for predicting substrate scope for this dataset. The approach is intended to be highly modular with respect to physicochemical property calculations and software used for docking and modeling.

Introduction

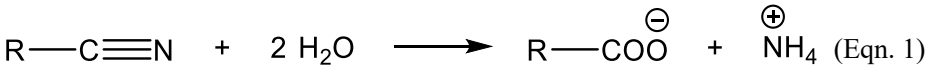
Many enzymes are capable of accepting multiple molecules as substrates. Knowledge of the repertoire of substrates for a given enzyme, often referred to as *substrate scope*, is informative for elucidating biochemical pathways and also for metabolic engineering. Standard sequence-based annotation methods are generally highly effective at identifying (super)family membership, conserved domains, sequence signatures, active site residues, and assigning gene ontology (GO) terms for sequences with detectable homology to proteins of known function but fall short of predicting substrate scope. The BRENDA enzyme database currently contains manually curated information on ~84,000 enzymes including classification nomenclature, biochemical reaction, substrate specificity, structure and other attributes, but is limited to experimentally verified systems.[32]

Beyond the primary amino acid sequence, protein structures provide insight into enzymatic function. The overall protein fold, domain architecture, and spatial arrangement of residues involved in substrate recognition and catalysis all provide useful clues to function. Homology modeling is often used to generate structural models of proteins when suitable templates are available. However, the accuracy of modeled structures depends on various factors, including the similarity between the query sequence and the template(s). Scoring functions and conformational sampling strategies also play a role in model accuracy.[1]

A combination of molecular docking of putative substrates to an available X-ray crystal structure, QM calculations of substrate reactivity, and experimental enzyme activity assays predicted substrate specificity of an enoyl-acyl carrier protein reductase (FabI).[33]In the absence of a crystal structure, homology modeling can be used in the context of ligand docking.[34-36] However, molecular docking studies often struggle to differentiate between ligands with similar scaffolds due to inaccuracies in the models and in scoring functions. In addition, docking is insufficient to predict enzymatic activity because it does not account for chemical reactivity.[37] Some of these limitations can be overcome by combining complementary information from modeling, docking and other sources. For example, a combined analysis of genomic context, homology modeling and metabolite docking was used to identify substrate specificities of multiple enzymes encoded in a bacterial gene cluster.[153]

Machine learning (ML) is widely applicable to a variety of problems from fields such as quantum mechanics, physical chemistry, biophysics, and physiology. For example, a Gaussian process model that incorporated information from protein sequence and contact maps derived from crystal structures was used in combination with directed evolution to engineer channelrhodopsin with high light sensitivity.[38] ML has also shown promise in predicting substrate specificity. For example, a support vector machines (SVM) approach was used to predict substrate specificity of adenylation domains in non-ribosomal peptide synthases from physicochemical properties of active site amino acids.[39]A related method extended this approach to predict specificity by incorporating active site structural information from sequence alignments to a template from a homologous structure.[40] Using SVM coupled with an active learning approach to prioritize compounds for experimental testing to provide maximal benefit to the model, substrates were predicted for four different enzymes with an accuracy of ~80%.[154] Enzymatic activity of 107 glycosyltransferase superfamily 1 (GT1) sequences from *Arabidopsis thaliana* was predicted with an accuracy of ~90% using a decision tree-based classifier that incorporated local sequence information, physicochemical properties of substrate donor and acceptor molecules, and experimental activity data.[41]

Nitrilases are a family of the carbon-nitrogen hydrolase superfamily that catalyze the hydrolysis of nitrile compounds to their corresponding carboxylic acids and ammonia (Eq. 1). They are an example of an enzyme family with broad scope and are found in a range of eukaryotic and prokaryotic organisms. Nitrilases play an important role in many biological processes, such as the degradation of toxic nitrile compounds, metabolism and generation of hormones, and synthesis of signaling molecules.[42] In the context of plant-microbe interactions, they are believed to play a role in hormone synthesis, nutrient assimilation, detoxification, and modulation of plant development and physiology, making them attractive for improved food crop production.[43]In addition, nitriles are desirable for their use in efficient chemo- and enantioselective synthesis of carboxylic acids, making them attractive for drug design.[44, 45] Typically, nitrilases are classified into three categories according to their substrate specificities: aliphatic, arylaceto-, and aromatic nitrilases.[43, 46] In terms of chemistry and reactivity, Enzyme Commission numbers have been assigned for aliphatic (EC 3.5.5.7) and arylacetonitrilases (EC 3.5.5.5). However, no broad category of aromatic nitrilases has been defined. Thus, existing sequence-based annotations are limited in their ability to classify nitrilases.



(Eqn. 1)

Various nitrilase activity assays have been described and are based on either fluorogenic or chromogenic substrates or pH indicator methods.[45, 47, 48] Recently, a chromogenic method was developed as a convenient means to screen recombinantly produced nitrilases in crude cell extracts.[49] Alleviating purification steps facilitates high-throughput screening and evaluation of diverse, potential substrates.

High-throughput methods are essential for evaluating the large number of putative nitrilases being identified through genome sequencing techniques. For example, functional screening of microbial metagenomes from a wide range of environments has led to the identification of a diverse collection of nitrilases. These efforts have facilitated characterization of the relationship between gene sequence and substrate specificity based on experimental evaluation of the hydrolysis of diverse nitrile substrates.[44]Three substrates, mandelic acid, phenyl lactic acid and 4-cyano-3-hydroxybutyric acid, were of particular interest due to their potential use in stereospecific pharmaceutical biosynthesis. Reactivity toward specific substrates as well as enzymatic stereoselectivity were found to be strongly correlated with the phylogenetic groupings of individual nitriles in sequence clades or clusters. Because most tested nitrilases were identified in metagenomic libraries and affiliation to specific organisms could not be determined, it is unknown if substrate specificity is linked to microbial taxonomy. More in depth analysis of some of the nitrilase subfamilies identified positive selective pressure for evolving novel substrate specificities and enantioselectivity, suggesting that these enzymes can undergo subtle site changes that alter their repertoire of accepted substrates.[155] Because shifts in substrate specificity and enantioselectivity were found associated with distinct sequences in specific subfamilies previously characterized for several substrates, we selected nine nitrilases from that study for in-depth enzymatic characterization and structural modeling. We also included two closely related putative nitrilases identified from bacterial genomes that potentially play roles in interactions with plant roots.[156]

Here we describe an integrated and modular approach in which we combine protein structural modeling, ligand docking, and physicochemical property calculation with experimental activity assays. We use this information to train several machine learning classifiers to predict enzyme activity for a set of bacterial nitrilases toward a library of 20 nitrile substrates. For this dataset, cross-validation revealed that that all four ML methods showed similar performance in predicting substrate scope.

Methods

For experimental details see:

Mou, Z., Eakes, J., Cooper, C.J., Foster, C.M., Standaert, R.F., Podar, M., Doktycz, M.J., Parks, J.M. Structure-based prediction of enzyme substrate scope with machine learning: Application to bacterial nitrilases. DOI: https://doi.org/10.22541/au.158888180.03951231

Phylogenetic analyses

Nitrilase sequences were selected for structural and enzymatic analyses based on prior substrate specificity data and were aligned along with related sequences from sequenced microbial genomes using Muscle v3.8[101] in Geneious v9.[102] Nitrilase sequences from plants were also included as an outgroup. A phylogenetic tree was constructed using FastTree v. 2.1.12.[103]

Structural modeling

The amino acid sequences of 12 target nitrilases were aligned with Clustal Omega (**Figure 26**).[157] The GREMLIN web server[91] was used to search the UniProt20 database for sequence homologs of each nitrilase, perform coevolution analysis, and identify potential structural templates from the Protein Data Bank. We used the 3.1 Å X-ray crystal structure of a bacterial nitrilase (Nit6803) from *Synechocystis* sp. PCC6803 (UniProt ID Q55949, PDB entry 3WUY) as a template and to generate a Rosetta symmetry file.[158] For all 12 putative nitrilases, the top template was 3WUY and the sequences were all covered well by the full *Synechocystis* sp. PCC6803 sequence (>81%). Structural modeling was supplemented with residue-residue contact restraints obtained from the coevolution analysis. We used *map_align* (https://github.com/gjoni/map_align) to align the contact maps to the top ten templates[22] identified by *hhsearch*. Due to the presence of inter-oligomeric contacts, dimer symmetry was defined based on 3WUY and this crystal structure was used as the master template for modeling. Fragments were obtained from the Robetta server.

RosettaCM[128] was then used to generate at least 5,000 models of each protein. We selected the top ten models based on the sum of the Rosetta energy and coevolution restraint score and aligned the models to the template dimer. For each protein, we selected the model with the lowest Rosetta score that had a low (< 3.5 Å) backbone RMSD to the 3WUY dimer and an "open" active site in which the volume of the active site (residues within 10 Å of Cα of the catalytic Cys) calculated with POVME 2.0[159] was greater than 50 $Å^3$.

Docking and docking descriptors

Three-dimensional structures of each nitrile were obtained from the ZINC database.[160] The geometry of each nitrile was optimized using density functional theory at the B3LYP/6-31G(d,p) level of theory in the gas phase.[161, 162] All quantum mechanical (QM) calculations were performed with Gaussian 16, revision A.03.[135] Restrained electrostatic potential (RESP) charges[163] were calculated at the HF/6-31G(d) level of theory in the gas phase. The optimized geometries and RESP charges were then used for docking with Rosetta Ligand.[133, 164] The REF2015 score function[165] was used for both homology modeling and docking. The center of mass of Sγ from Cys, Oε2 from Glu and Nζ from Lys in the catalytic triad was used as the initial docking site. We generated 5,000 docked models for each nitrile-nitrilase combination and selected the final docked pose based on the docking energy (*interface_delta*). Additional components of the Rosetta docking score were also included as descriptors for RF. These components included the following interfacial interaction energy terms: full-atom vdW attraction (*fa_atr*), electrostatics (*fa_elec*), vdW repulsion (*fa_rep*), hydrogen bonding terms (*hbond_bb_sc* and *hbond_sc*), and solvation energy (*fa_sol*).

Physicochemical descriptors

"Classical" 2D and 3D physicochemical descriptors were calculated with MOE.[166] QM descriptors included atomic partial charges computed from natural population analysis[167] and Merz-Singh-Kollman (MK) charges[168, 169] the C and N atoms of the cyano group, highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, molecular dipole moment, and molecular volume.

Active site descriptors

The active site of each enzyme-ligand pair was defined as all protein and ligand atoms within 10 Å of the Cα atoms of the catalytic triad. ProtDCal[170] was used to calculate active-site descriptors including thermodynamic indices of the folded and extended protein state, topographic indices, physicochemical and structural composition indices.

Machine learning and statistical analysis

The *scikit-learn* package (version 0.22) was used to perform the binary classification analysis using four ML methods including two decision tree-based ensemble methods: random forest (RF)[171] and gradient boosted decision trees (GBDT)[172], as well as a kernel-based method, support vector machines (SVM)[173], and logistic regression (LR). For this analysis, experimentally measured activities of < 2 mM ammonia were considered inactive and descriptors with high correlation to other descriptors (≥ 0.9) were removed. All statistical analyses and plotting were performed with Python 3.7 using Pandas, Numpy, and Matplotlib.

Results

We reasoned that protein structural modeling and ligand docking combined with physicochemical properties that describe the ligand and its reactivity could be used synergistically to predict substrate preferences. Structural modeling provides insight into overall protein folds and the arrangement of residues in the active site. Docking scores provide approximations of binding affinities but do not account for reactivity, which can be instead quantified by computing QM properties of the nitriles that depend on electron density and molecular orbitals. Additional molecular properties of the nitriles can be taken into account by calculating classical physicochemical descriptors (e.g., van der Waals surface area and related quantities). As a test case we selected bacterial nitrilases, which catalyze the hydrolysis of nitriles to form the corresponding carboxylates and ammonia (Eqn. 1). To create an effective training set, we selected a set of 12 nitrilase sequences (**Figure 27**) and evaluated their activity computationally and experimentally against a set of representative aliphatic, aromatic and arylaliphatic nitriles. The various descriptors and experimentally determined activity data were then used the machine learning classifiers to predict enzyme substrate scope.

Sequence selection and structural modeling

Standard sequence-based approaches generally cannot assign substrate preferences at the individual molecule level. Thus, we developed a structure- and property-based ML approach to predict substrate scope using bacterial nitrilases as a test case. Previously, 137 unique nitrilase sequences were identified by screening more than 600 environmental samples from terrestrial and aquatic environments.[44] The enzymes were then expressed heterologously and assayed for their ability to catalyze the enantioselective hydrolysis of three pharmaceutically relevant nitriles, 3-hydroxyglutaronitrile (3HGN), mandelonitrile (MA), and phenylacetaldehyde cyanohydrin (PAC), to form the corresponding carboxylic acids. Phylogenetic analysis of these sequences identified six distinct sequence clades that exhibited varying reactivities and enantioselectivities toward the three substrates. For example, nitrilase 1B15 hydrolyzed all three substrates with an enantiomeric excess for the corresponding R isomeric product ranging from 33 to 100%. In contrast, 1B16 exhibited S enantioselectivity toward 3HGN and PAC, but did not hydrolyze MA. From this set of 137 nitrilases, we selected a small representative set of nine enzymes from among three sequence clades. Greater emphasis was placed on two adjacent subclades (1A and 1B), but we also selected one sequence each from clades 2 and 3.

To date, only a few structures of nitrilases have been determined with X-ray crystallography. One such structure is that of Nit6803 from *Synechocystis* sp. PCC6803 (PDB entry 3WUY)[158], which is a member of sequence clade 1B (**Figure 27**). This enzyme hydrolyzes a broad range of nitriles, including aliphatic and aromatic mono- and dinitriles. In addition, we included two putative nitrilases identified in the genomes of plant rhizosphere-associated bacteria. These enzymes were selected on the basis of their similarity to sequences from subclade 1A and also to the structural template Nit6803. These 11 sequences have varying degrees of sequence identity to Nit6803 and range from 32-71% with a sequence coverage of at least 81%. We used the structure of Nit6803 as a template to generate homology models of a selected set of 20 nitriles (**Figure 25**) from among these substrate categories based on previous data sets[44, 49] and docked them to each enzyme model and also to the Nit6803 crystal structure (**Figure 28**). We then calculated various QM and classical physicochemical properties for each nitrile and additional active-site properties from the docked poses.

Enzyme activity assays

Target nitrilases were expressed heterologously in *E. coli* and were prepared as crude extracts.[174] These enzyme-containing extracts were added to solutions containing a selected nitrile and enzymatic activity was measured using a semi-quantitative colorimetric assay optimized for crude extracts based on a previously described method.[49]

All 12 enzymes were active toward at least one nitrile (**Figure 29**). In general, catalytically active enzymes tend to hydrolyze multiple nitriles with no obvious patterns in activities. Not surprisingly, docking scores do not correlate with enzymatic activity (**Figure 30**). We observed negligible activity (i.e., ≤ 2) toward all aliphatic nitriles except for 2-methylglutaronitrile. Interestingly, 1B15 and 1A8 were the only enzymes that did not display activity toward this nitrile. Furthermore, 1B15 was the only enzyme that had no activity toward aliphatic or aromatic nitriles. Thus, 1B15 is specific for arylaliphatic nitriles but is only moderately active for 3-phenylpropionitrile and cinnamonitrile. No appreciable activity was measured for any enzyme with 2-aminobenzonitrile or 2,6-dichloroaminobenzonitrile. 2A6 was active toward all arylaliphatic nitriles except cinnamonitrile and was the only enzyme that hydrolyzed mandelonitrile and α-methylbenzyl cyanide.

Prediction workflow

Having obtained the experimental activity assay data, structural models, docked ligand, and calculated descriptors, we trained various binary classification ML models to predict substrate scope for bacterial nitrilases. Because the activity assays are semi-quantitative, we used a binary classification approach to predict whether a given enzyme is active or inactive toward a given nitrile according to a chosen activity threshold. We considered four different activity thresholds (1, 2, 3 and 4 mM) for classifying nitrilase activity and selected a threshold of 2 mM ammonia to define enzyme-substrate pairs with negligible activity as being essentially inactive. Thus, activities below 2 mM were considered inactive.

To determine whether the use of oversampling techniques could be used to generate better models, a variety of synthetic minority oversampling technique (SMOTE)[175] methods were tested. For grid-search hyperparameter tuning and cross-validation we used an 80/20 training/test set split. We

further tested the robustness of the models by performing leave-*n*-protein-out tests, which were conducted by randomly and phylogenetically leaving out $n = 1, 2, 3, 4$ or 6 proteins during training and then using them as test sets.

We analyzed the performance of four different ML methods that are generally considered suitable for datasets of this size. These methods included random forest (RF), gradient-boosted decision trees (GBDT), logistic regression (LR), and support vector machines (SVM). For this dataset, which has a ratio of inactive:active substrates of 2:1 using a cutoff of 2 mM ammonia, oversampling did not significantly improve model performance. All four ML methods perform similarly as evaluated by performing tenfold cross-validation (**Figure 31A**). The average areas under the ROC curve (ROC_AUC) were all ~0.90 and the models had average accuracies of 79-83%. The methods also performed similarly for the test set with the exception of the recall metric, for which GBDT did not perform as well as the others (**Figure 31B**). Although the test set was used to assess classification predictions on completely unseen data, it only reflects a single, randomly chosen subset of the data. Thus, model performance from the test set does not necessarily reflect the overall robustness of the model.

We further assessed the robustness of the different ML methods by leaving out one enzyme at a time, training separate models on the remaining eleven enzymes, and then predicting the substrate scope for the left-out enzyme (**Figure 31C**). All four ML methods performed similarly for ROC_AUC, accuracy, and precision. However, RF performed the best for F1 and recall. We then randomly removed two, three, four, and six of the twelve proteins and observed that RF performance was similar the other methods and for some metrics outperformed GBDT, LR, and SVM.[174] In addition to randomly leaving out proteins, we also removed two, three, four, and six proteins according to their order and proximity in the phylogenetic tree to investigate the contribution of phylogenetic relationships on model performance. As observed for the random leave-out tests, RF generally performed similar to or outperformed the other methods in some metrics.

Discussion

Here we have developed an approach for predicting substrate scope for enzymes by combining structural modeling, docking, physicochemical properties and various machine learning methods

(**Figure 32**). Rather than generating a large training set, we sought to explore the limits of accuracy of the model by training the ML model on a relatively small amount of targeted in vitro enzyme assay data. The time and expense involved with generating and screening enzymes demands effective in silico approaches. Here, the use of crude extracts that contain heterologously produced enzymes combined with an automated, colorimetric activity assay facilitated construction of an effective training set. Our approach enables accurate predictions of substrate scope for a series of aliphatic, aromatic, and arylaliphatic nitriles by including descriptors for the enzymes, substrates and their interactions in ML models.

Given a phylogenetic tree and sparse activity data, it may be difficult to identify trends in substrate scope. In some cases, sequences that have high sequence identity show similar trends in substrate preference. For example, 1A1 and 1A2 are closely related (85% identical) and their substrate scopes differ only for the substrate 4-(dimethylamino)benzonitrile (**Figure 29**). 1B16 and 3WUY are also closely related (71% identical) and show similar patterns in activity (90% overlap in substrate scope). However, PMI28 and 1A8 are 88% identical but differ markedly in their respective substrate scopes. PMI28 displays activity toward 12 of the 20 nitriles spanning all three classes, making it one of the most active enzymes tested. In contrast, 1A8 is only active toward two aromatic nitriles. In other cases, distantly related sequences share similar substrate preferences. For example, 1A17 and 3WUY (51% identical) have the same substrate scope except that 4-nitrophenylacetonitrile is not hydrolyzed by 3WUY. Therefore, predictions of the substrate scope of an enzyme often cannot be made based on phylogenic analysis alone. In addition, subtle changes in the amino acid composition of the active site or in the chemical structure of the substrate may lead to differences in activity. In the present case, active enzymes tend to have high activity for many nitriles. However, in other cases it will not be known beforehand how much of the specificity space will be covered by the proteins or the substrate library. In such cases, active learning approaches in which the training data are augmented iteratively to optimize model performance, are expected to be particularly useful.[154]

Substituent effects play an important role in determining reactivity. For example, 2-aminobenzonitrile and 2,6-dichloroaminobenzonitrile are both aromatic nitriles with substituents that are ortho to the cyano group. In contrast to the other aromatic nitriles, these two molecules

were not hydrolyzed by any of the nitrilases tested. This large difference in reactivity may be due to the steric hindrance of the ortho functional groups or substituent effects. The two dinitriles were readily hydrolyzed by most enzymes, with the exceptions of 1A8 and 1B15 toward 2-methylglutaronitrile and 1A1, 1A2, and 1B15 toward isophthalonitrile. These dinitriles have high activities compared to the mononitriles, suggesting that both nitrile groups in the dinitriles were hydrolyzed. In a dinitrile, the conversion of one nitrile substituent to a carboxylate will alter the solubility and electrostatic properties of the resulting intermediate, which could affect the binding affinity and reactivity of the secondary substrate.

In a previously proposed catalytic mechanism for nitrilases[176], the first step of the reaction consists of a series of proton transfer steps involving the catalytic Cys, the cyano group, and an ordered water molecule, resulting in the formation of a thioimidate intermediate. Geometries of catalytic residues across a given enzyme family tend to be well conserved (i.e. RMSD < 0.5 Å) and it has been shown that incorporating this information in the form of geometric constraints can improve model quality.[177] Furthermore, docking results can potentially be improved by including additional restraints that account for specific interactions between the enzyme and putative substrates (i.e., selecting for catalytically relevant orientations). As enzymes preferentially bind transition states over ground states of substrates, it could be beneficial to include information about transition states in the docking calculations. Performing docking with a transition state mimic is a promising approach that can provide improved accuracy compared to ground state docking.[178] Most of the nitrile substrates considered in the present work are relatively rigid and extensive conformational sampling was not required. However, for other cases with more flexible ligands, conformational sampling may be critical and should therefore be included.

RF models performed as well as, or in some cases better than, the other three ML methods. Unlike kernel-based methods (i.e., SVM), decision tree-based methods (i.e., RF) allow for calculation of variable importance of each descriptor. For the nitrilase example, we used an 80:20 split of the data and calculated the variable importance for 20 independent runs initiated with different random seeds**.** Descriptors from all four categories were present in the top 10 most important descriptors over the 20 runs (**Figure 33**). Thus, including complementary information from each category indeed contributes to the predictive value of the model. QM descriptors do not appear frequently

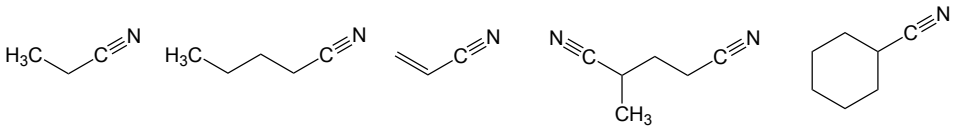
in the top 10, suggesting that descriptors intended to account for chemical reactivity are not as important as other properties for obtaining accurate predictions. In the present case, the QM descriptors are all similar among the 20 nitriles. For example, the natural population analysis (NPA) partial charge on the nitrile carbon ranges from 0.25 to 0.3. In contrast, MOE descriptors capture more global properties of the substrate molecules and are therefore more informative for classification. Although there are more ligand descriptors (MOE and QM) than those that contain information about the ligand in the context of the protein from the docked pose, docking and ProtDcal descriptors comprise the majority of the top 10 lists (**Figure 33**). Thus, for this system the descriptors that encode information from the structural models and docked poses are informative for accurately predicting substrate scope.

The approach developed here was designed to be highly modular, with readily swappable computational components. For example, protein modeling could be performed with other software such as I-TASSER[179-181], MODELLER[182],SWISS-MODEL[183], and others. Similarly, ligand docking could be performed with software such as Glide[184], AutoDock Vina[185], and many others. Alternatives for calculating physicochemical descriptors include Rcpi[186], PaDEL[187],Mordred[188] and essentially any quantum chemistry software. As expected, single amino acid substitutions can cause large changes in reactivity or specificity that would not be identified based on a phylogenetic analysis of the full sequence. In principle, our approach can capture these subtle effects if they lead to substantial changes in active site properties. Compared to sequence-based approaches[189], the modular, structure-based machine learning approach described here is more flexible, and should be readily extensible to enable prediction of substrate scope for many classes of enzymes. In addition, the experimental assays used are scalable for high-throughput applications. The application of advanced computational methods will lead to a better understanding of enzyme structure-function relationships and metabolic processes.

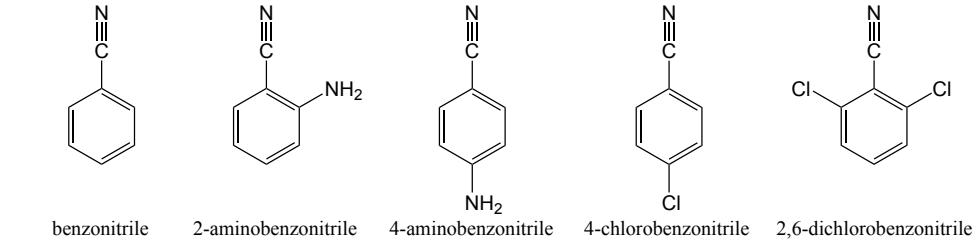
Appendix III

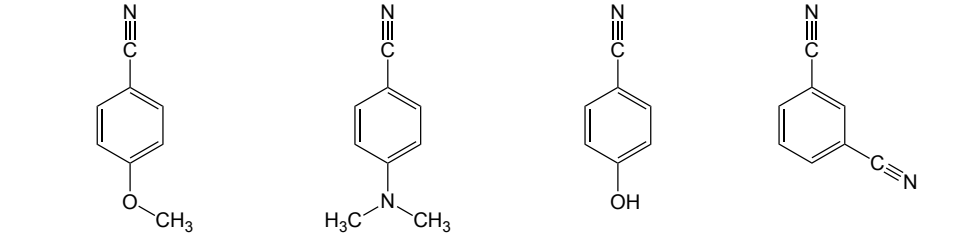
Figures

Aliphatic



Aromatic





Arylaliphatic

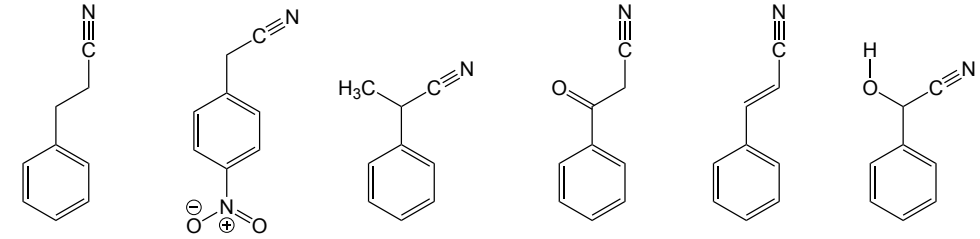


Figure 25. Nitriles used in this study to screen for nitrilase activity. See ref 174 for additional details on experimental methods.

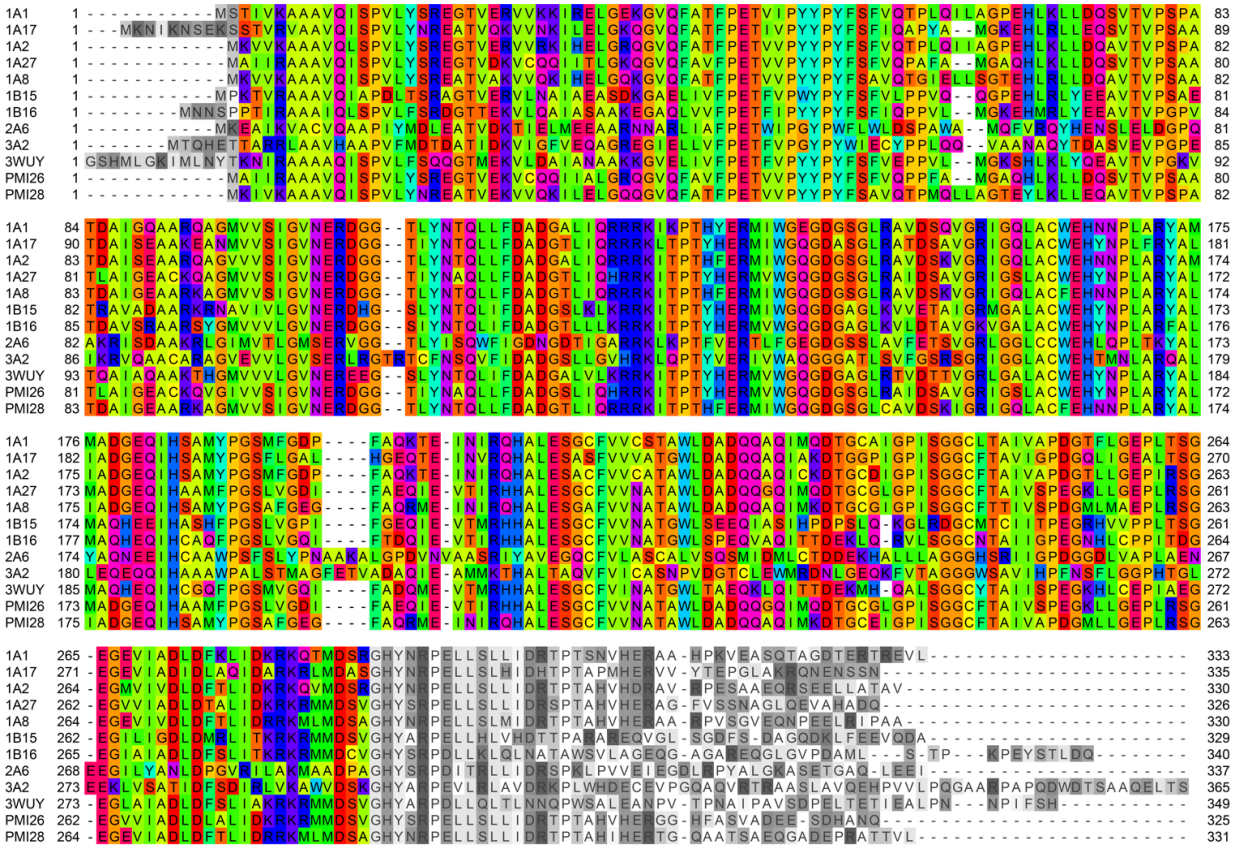


Figure 26. Clustal Omega alignment of target nitrilase sequences. Residues in gray were excluded from the models because they were not resolved in the template.

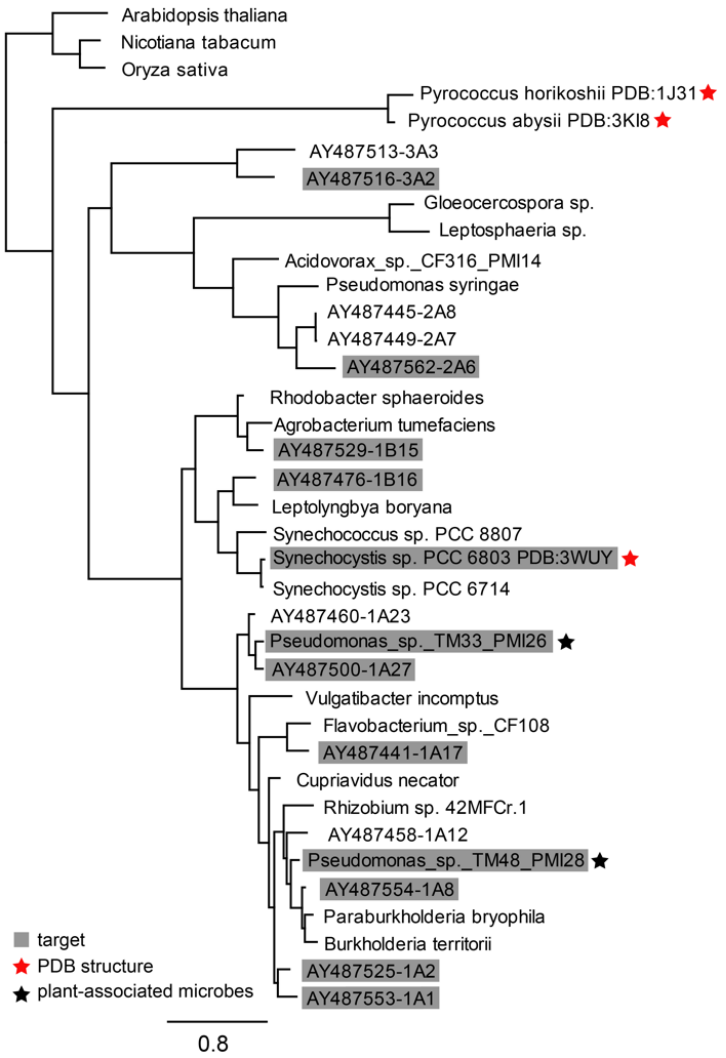


Figure 27. Phylogenetic tree of a family of nitrilases that encompass the enzymes used in this study (grey). The scale bar indicates the inferred number of substitutions per site. Enzymes for which an X-ray structure is available are indicated with a red star. Two putative nitrilases from plant root-associated bacteria are indicated with a black star.

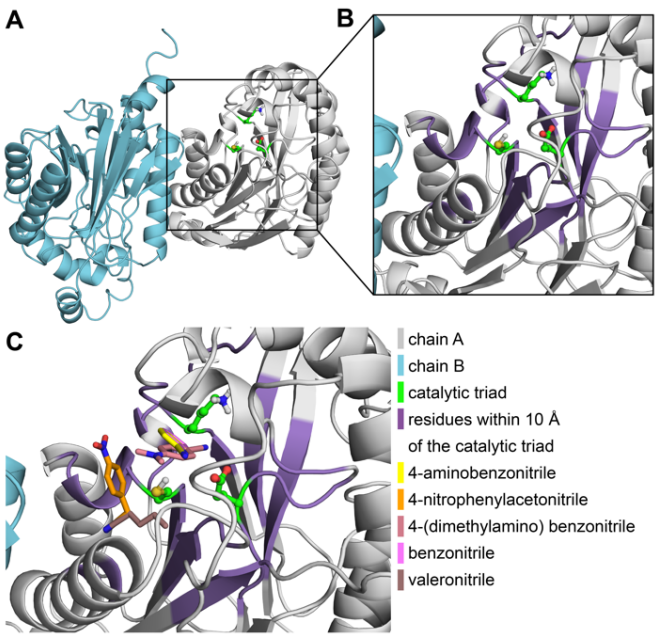
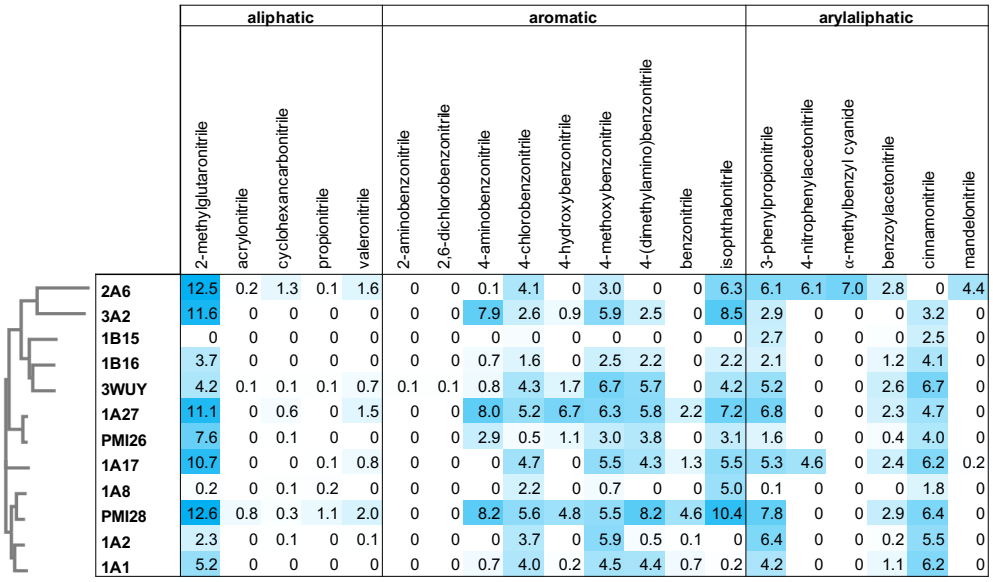


Figure 28. Nitrilase models and docking. (**A**) Structural model of a representative nitrilase (PMI26) with the catalytic triad of chain A shown as ball and stick and colored by element. (**B**) Residues within 10 Å of the catalytic triad. (**C**) Selected docked poses of nitriles are shown as sticks and colored by element with different colored carbons for each nitrile. Side chain carbons of the catalytic triad are shown in green.



	aliphatic					aromatic									arylaliphatic					
	2-methylglutaronitrile	acrylonitrile	cyclohexancarbonitrile	propionitrile	valeronitrile	2-aminobenzonitrile	2,6-dichlorobenzonitrile	4-aminobenzonitrile	4-chlorobenzonitrile	4-hydroxybenzonitrile	4-methoxybenzonitrile	4-(dimethylamino)benzonitrile	benzonitrile	isophthalonitrile	3-phenylpropionitrile	4-nitrophenylacetonitrile	α-methylbenzyl cyanide	benzoylacetonitrile	cinnamonitrile	mandelonitrile
2A6	12.5	0.2	1.3	0.1	1.6	0	0	0.1	4.1	0	3.0	0	0	6.3	6.1	6.1	7.0	2.8	0	4.4
3A2	11.6	0	0	0	0	0	0	7.9	2.6	0.9	5.9	2.5	0	8.5	2.9	0	0	0	3.2	0
1B15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.7	0	0	0	2.5	0
1B16	3.7	0	0	0	0	0	0	0.7	1.6	0	2.5	2.2	0	2.2	2.1	0	0	1.2	4.1	0
3WUY	4.2	0.1	0.1	0.1	0.7	0.1	0.1	0.8	4.3	1.7	6.7	5.7	0	4.2	5.2	0	0	2.6	6.7	0
1A27	11.1	0	0.6	0	1.5	0	0	8.0	5.2	6.7	6.3	5.8	2.2	7.2	6.8	0	0	2.3	4.7	0
PMI26	7.6	0	0.1	0	0	0	0	2.9	0.5	1.1	3.0	3.8	0	3.1	1.6	0	0	0.4	4.0	0
1A17	10.7	0	0	0.1	0.8	0	0	0	4.7	0	5.5	4.3	1.3	5.5	5.3	4.6	0	2.4	6.2	0.2
1A8	0.2	0	0.1	0.2	0	0	0	0	2.2	0	0.7	0	0	5.0	0.1	0	0	0	1.8	0
PMI28	12.6	0.8	0.3	1.1	2.0	0	0	8.2	5.6	4.8	5.5	8.2	4.6	10.4	7.8	0	0	2.9	6.4	0
1A2	2.3	0	0.1	0	0.1	0	0	0	3.7	0	5.9	0.5	0.1	0	6.4	0	0	0.2	5.5	0
1A1	5.2	0	0	0	0	0	0	0.7	4.0	0.2	4.5	4.4	0.7	0.2	4.2	0	0	1.1	6.2	0

Figure 29. Activity data (ammonia concentration in mM) for putative nitrilases with 20 nitrile substrates obtained from cell extracts at 50% dilution. Background color to the activity data values is added as a visual aid in estimating relative enzyme-substrate activity. See ref 174 for additional details on experimental methods.

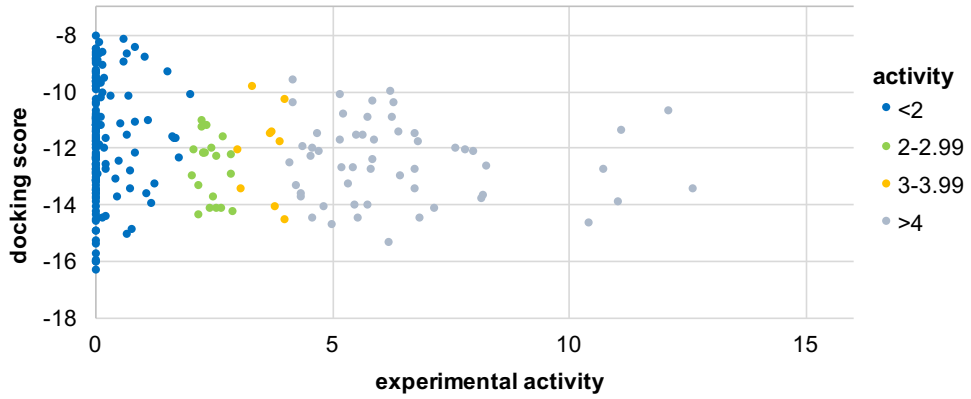


Figure 30. Experimental nitrilase activity (ammonia concentration in mM) versus Rosetta docking score. See ref 174 for additional details on experimental methods.

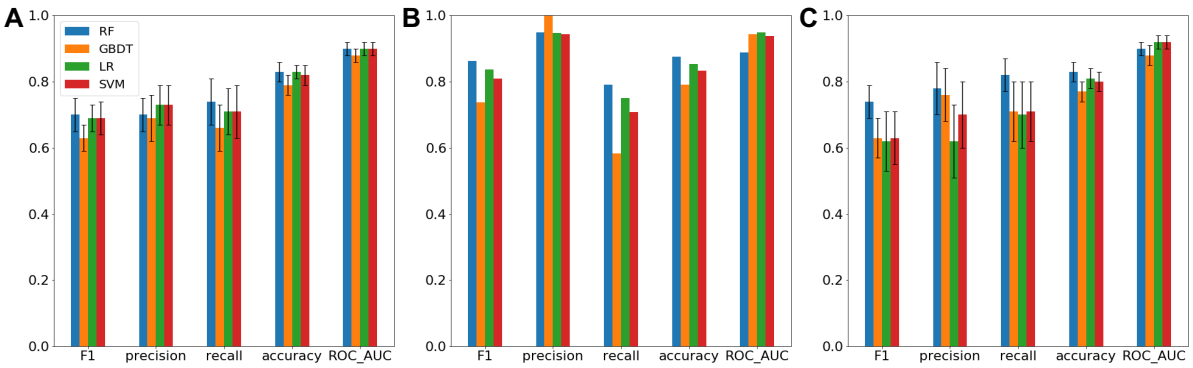


Figure 31. Machine learning model metrics. (**A**) Tenfold cross-validation (**B**) 80/20 test set and (**C**) leave-one-protein-out tests for a set of bacterial nitrilases and nitrile substrates. Error bars indicate the standard error of the mean (s.e.m.) with n = 10 for **A** and n = 12 for **C**.

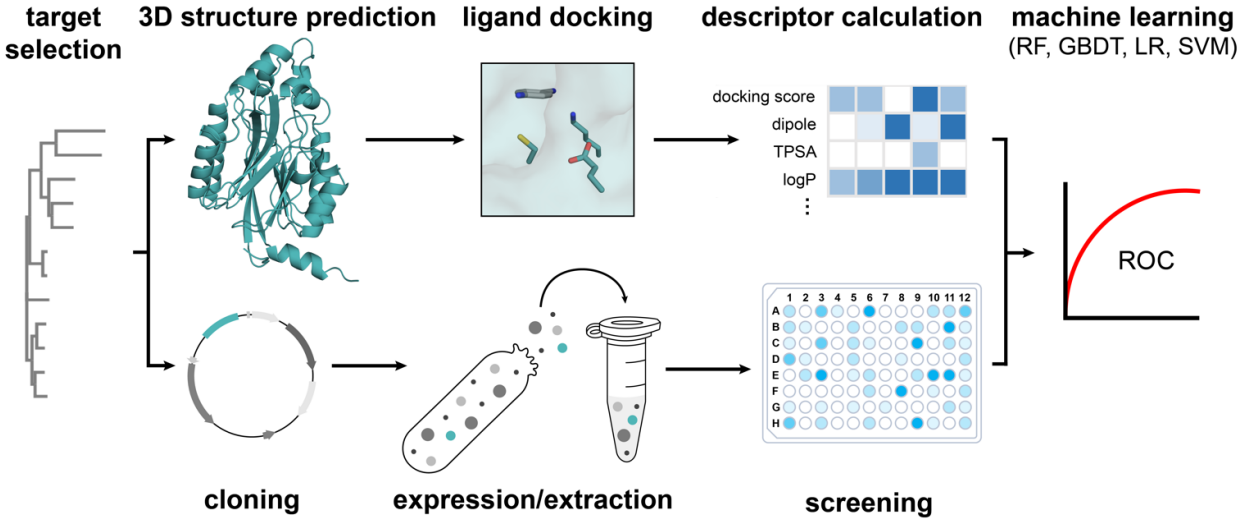


Figure 32. Graphical overview of the structure-based approach to predict the substrate scope of enzymes. After target selection, structural models are generated for docking and descriptor calculation and targets are cloned, expressed, and extracted for screening. The experimental activity data and calculated descriptors are then used to train an RF classification model that can then be used to predict substrate scope.

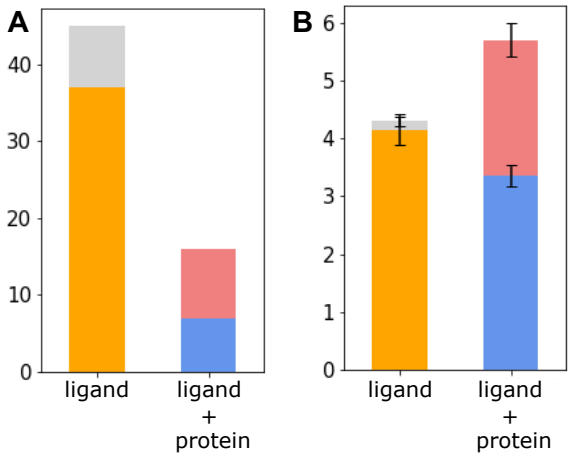


Figure 33. Analysis of descriptor categories. (**A**) Number of descriptors per category used for ML model building. (**B**) Descriptor counts for the top 10 features in 20 random seeds. Descriptors are colored by category (MOE = orange, QM = gray, docking = blue, ProtDCal = red). Error bars indicate the standard error of the mean (s.e.m.) with n = 20.

CHAPTER IV

MOLECULAR PROPERTIES THAT DEFINE THE ACTIVITIES OF ANTIBIOTICS AND IDENTIFICATION OF NOVEL EFFLUX PUMP INHIBITORS

Text and figures are taken from the following:
Cooper, S.J., Krishnamoorthy, G, Wolloscheck, D, Nguyen, J, Walker, J.K., Rybenkov, V.V., Parks, J.M. and Zgurskaya, H.I. Molecular properties that define the activities of antibiotics in *Escherichia coli* and *Pseudomonas aeruginosa*. *ACS Infect. Dis*. 2018, 4, 1223–1234.
Copyright (2020) American Chemical Society.

Experiments were performed in the lab of H.I.Z.; S.J.C. and J.M.P calculated physicochemical properties and performed RF analysis; S.J.C., J.M.P., and H.I.Z. wrote the manuscript with input from all other authors.

Green, A.T., Moniruzzaman, M., Cooper, C.J., Walker, J.K., Smith, J.C., Parks, J.M., Zgurskaya, H.I. Discovery of multidrug efflux pump inhibitors with a novel chemical scaffold. *BBA General Subjects.* 2020, 1864, 129546.

Experiments were performed in the lab of H.I.Z.; A.T.G. performed docking calculations; A.T.G., C.J.C., and J.M.P analyzed docking results. All authors contributed to preparing the manuscript.

Abstract

The permeability barrier of Gram-negative cell envelopes is the major obstacle in the discovery and development of new antibiotics. In Gram-negative bacteria, these difficulties are exacerbated by the synergistic interaction between two biochemically distinct phenomena, the low permeability of the outer membrane (OM) and active multidrug efflux. In this study, we used *Pseudomonas aeruginosa* and *Escherichia coli* strains with controllable permeability barriers, achieved through hyperporination of the OMs and varied efflux capacities, to evaluate the contributions of each of the barriers to protection from antibacterials. We analyzed antibacterial activities of β-lactams and fluoroquinolones, antibiotics that are optimized for targets in the periplasm and the cytoplasm, respectively, and performed a machine learning-based analysis to identify physicochemical descriptors that best classify their relative potencies. Our results show that the molecular properties selected by active efflux and the OM barriers are different for the two species. Antibiotic activity in *P. aeruginosa* was better classified by electrostatic and surface area properties, whereas topology, physical properties, and atom or bond counts best capture the behavior in *E. coli*. In

several cases, descriptor values that correspond to active antibiotics also correspond to significant barrier effects, highlighting the synergy between the two barriers where optimizing for one barrier promotes strengthening of the other barrier. Thus, both barriers should be considered when optimizing antibiotics for favorable OM permeability, efflux evasion, or both. Inhibition of multidrug efflux pumps is a promising approach for reviving the efficacy of existing antibiotics. Using existing physicochemical property guidelines in combination with computational ligand docking, we identified a new class of inhibitors of *E. coli* AcrAB-TolC. Six molecules with a shared scaffold were found to potentiate antibiotic activity.

Introduction

Gram-negative bacteria are notoriously more resistant to antibiotics than Gram-positive bacteria. The major reason for this resistance is that Gram-negative cell envelopes comprise two membranes of different compositions and functions.[55, 190, 191] The outer membrane (OM) is an asymmetric bilayer of lipopolysaccharides (LPS) and phospholipids with non-selective porins and substrate-specific channels embedded therein.[51, 52] The major function of the OM is protection from toxic molecules and enzymatic attacks in a hostile environment. The inner, or cytoplasmic, membrane is a phospholipid bilayer that is responsible for diverse physiological and metabolic functions. It also contains multidrug efflux pumps that protect intracellular functions by actively removing small, toxic molecules and peptides from the periplasm and cytoplasm.[53] The two barriers-the passive, low-permeability OM and active efflux in the inner membrane-act synergistically and are the major factors that are responsible for the intrinsic resistance of Gram-negative bacteria to a broad range of antimicrobial agents.[57, 192] In addition, the orthogonal to the OM sieving properties of the inner membrane are also thought to affect the intracellular accumulation of antibiotics [191].

The antibiotic resistance of Gram-negative pathogens has become particularly worrisome with the emergence of multidrug resistant strains in clinics, which often leave clinicians with no therapeutic options.[50] The discovery of new antibiotics that are active against these pathogens is hindered by low hit rates in screening efforts and by the lack of practical rules to maximize OM permeability and minimize efflux.[55, 56] The latter problem has been identified as a major bottleneck in addressing emerging multidrug resistance in clinics.[54]

To establish rules based on molecular properties that define antibiotic permeation, the two permeability barriers (OM and efflux) must be analyzed separately to define the factors contributing to each barrier.[57] For this purpose, we developed a hyperporination approach that facilitates control of OM permeability in Gram-negative cells through the inducible expression of a chromosomally encoded open pore (Pore) with a 2.4-nm internal diameter [58]. The expression of the Pore effectively and non-selectively allows influx of antibiotics and reduces the barrier constant, *B,* which is defined as the ratio of maximum attainable drug fluxes across the outer membrane into the cell and out of the cell via the efflux transporter.[192] The overexpression and

deletion of efflux pumps, on the other hand, allows manipulation not only of *B,* but also the efflux constant, K_E, which measures efflux efficiency for a given antibiotic.[192-195]

We focused on understanding interactions between the permeability barriers of Gram-negative bacteria and the antibacterial activities of β-lactams (BLs) and fluoroquinolones (FQs). Representatives of these antibiotic classes have been extensively developed and remain the major antibiotics administered in clinics. FQs target DNA replication by inhibiting DNA topoisomerases and, hence, to reach their targets must penetrate both the outer and inner membranes and evade efflux pumps. In contrast, transpeptidases, which are targeted by BLs, are located in the periplasm and these antibiotics are optimized to penetrate only across the OM and to evade efflux from the periplasm. Thus, the two classes differ significantly in their structures and physicochemical properties and contain determinants that are recognized by these different barriers.

Antibacterial activities were analyzed in two Gram-negative species that differ significantly in their permeability barriers: *P. aeruginosa* and *E. coli.* Although the lipid compositions of the OMs overall are similar between these two species [51, 59, 60], they differ in the composition and structure of their major general porins. The OM of *E. coli* contains ~200,000 copies per cell of OmpF/C porins, which have a molecular mass cutoff of ~600 Da. As a result, a significant number of antibiotics are active against this species.[61] In contrast, *P. aeruginosa* lacks such large general porins and instead utilizes substrate-specific porins of the Occ family to take up small compounds such as monosugars and amino acids.[62] Nevertheless, this species is susceptible to FQs and some BLs, suggesting alternative routes of permeation across the OM. Hyperporination of the OM through the expression of large non-specific pores negates the differences in permeabilities of the OMs in *P. aeruginosa* and *E. coli* and allows evaluation of the contributions of these barriers toward antibacterial activities. [57, 58]

Cryo-EM structures have been determined for the assembled AcrAB-TolC efflux pump, which consists of three main components, AcrB, AcrA and TolC (**Figure 34A** in **Appendix IV**). AcrB is a homotrimeric protein that consists of an α-helical integral membrane domain, a periplasmic porter domain that binds and extrudes substrates, and a docking domain that interacts with AcrA.[64, 65] AcrA is a membrane fusion protein that consists of four domains: α-hairpin, lipoyl, β-barrel, and

membrane-proximal (**Figure 34B**). TolC is a trimeric protein that consists of a β-barrel domain embedded in the OM and a periplasmic α-helical coiled-coil domain.[66] AcrB, AcrA, and TolC assemble in a 3:6:3 stoichiometry[196] to form a complex that spans the entire Gram-negative cell envelope.[63] In *E. coli,* inactivation of a single gene, *tolC,* leads to the complete loss of efflux across the OMs, because all efflux pumps capable of efflux across the OMs in this species depend on TolC. [197-199] In contrast, the major efflux pumps of *P. aeruginosa* are encoded in the same operons as specific *tolC* homologs, and each pump is functionally independent from the others. [200] Hence, multiple pumps must be inactivated to deplete the efflux capacity of *P. aeruginosa* and the differences between the barriers are further evident in the genetic make-up of their respective efflux pumps.

Identifying molecular properties that govern antibiotic activity in the presence and absence of the two barriers is expected to provide strategic guidelines for optimizing compounds against gram-negative bacteria. [191] Recently, random forest (RF) machine learning was used establish a set of rules for favorable accumulation of antibiotics in *E. coli*.[67, 201, 202] Liquid chromatography with tandem mass spectroscopy (LC–MS/MS) analysis revealed that small-molecule compounds containing amine functional groups were most likely to accumulate in *E. coli* cells, with primary amines having the highest accumulation.[67] Incorporation of a primary amine into the Gram-positive antibiotic deoxynybomycin (6DNM) resulted in a new antibiotic (6DNM-NH_3) that exhibited broad-spectrum activity against a panel of multidrug-resistant Gram-negative bacteria. In addition to containing an amine, antibiotics that tended to be successful at bypassing the OM permeability barrier were polar, amphiphilic, relatively rigid, and had low globularity.

We identify molecular properties of antibiotics that are associated with their activities, measured as minimum inhibitory concentrations (MICs), in *P. aeruginosa* and *E. coli* strains with controlled permeability of the OMs and variable efflux capacities. We also describe the characteristics of antibiotics that display activity when both the efflux and OM barriers are removed (PΔ6-Pore and ΔTolC-Pore), when only efflux (PΔ6 and ΔTolC) or the OM barrier is removed (PAO1-Pore and WT-Pore), and in the corresponding wild-type strains (PAO1 and WT). To establish these associations, we use RF classification to extract physicochemical properties of antibiotics that separate them based on the contributions of these two barriers. Our results show that molecular

properties selected by active efflux and the OM barriers are different for *E. coli* and *P. aeruginosa*. By combining existing physicochemical rules for OM permeability and efflux in *E. coli* in with computational docking, *in vitro* binding assays, and *in vivo* potentiation assays in bacterial strains with controllable permeability barriers we identified six molecules with a shared scaffold that potentiate the antibiotic activity of erythromycin and novobiocin in hyperporinated *E. coli* cells and in wild-type strains of both *A. baumannii* and *K. pneumoniae*.

Methods

For additional experimental details see:

Cooper, S.J., Krishnamoorthy, G, Wolloscheck, D, Nguyen, J, Walker, J.K., Rybenkov, V.V., Parks, J.M. and Zgurskaya, H.I. Molecular properties that define the activities of antibiotics in *Escherichia coli* and *Pseudomonas aeruginosa*. *ACS Infect. Dis*. 2018, 4, 1223–1234.
and
Green, A.T., Moniruzzaman, M., Cooper, C.J., Walker, J.K., Smith, J.C., Parks, J.M., Zgurskaya, H.I. Discovery of multidrug efflux pump inhibitors with a novel chemical scaffold. *BBA General Subjects.* 2020, 1864, 129546.

Experimental MIC and MPC *measurements*

All strains[201] were grown in Luria-Bertani broth at 37˚C with shaking. Susceptibilities of cells to different classes of antibiotics were determined by 2- and 4-fold dilutions as described previously.[57, 58] Therefore, MIC ratios of 2-4-fold changes are within error of the assay. For RF classification of MICs and MIC ratios, the lowest MIC in the range was used. Antibiotics were purchased from MicroSource Discovery Systems, Inc. and Sigma-Aldrich. All minimal inhibitory concentration (MICs) determinations were done at least twice. For EPIs, minimal potentiating concentration (MPC) was defined as a concentration of a compound that decreases the MIC of an antibiotic by four (MPC_4) or more fold.

Physicochemical property calculation

Three-dimensional structures of antibiotics used in RF classification were obtained from the ZINC database.[132, 203] Marvin calculator plugins[204] were used to calculate the most likely tautomeric and protonation states at pH 7.4. Geometries were optimized with the Amber12:EHT molecular

mechanics force field[205] implemented in MOE version 2015.[206] MOE was then used to calculate >300 2D and 3D molecular descriptors[207], and the resulting descriptor values were analyzed with respect to the MIC and MIC ratio data. Descriptors with standard deviations equal to zero were discarded. Redundant descriptors (i.e., correlation coefficients > 0.85) were identified and removed using the *findCorrelation* function in the R package *caret.*[208] A total of 143 descriptors were used for *P. aeruginosa*, and 142 descriptors for *E. coli* (**Table 9** in **Appendix IV**). Prior to analysis, the descriptor values were scaled and centered so they all had the same variance.

RF analysis of MICs and MIC ratios

RF combines the information from numerous decision trees to obtain a consensus classification of "high" or "low" activity (MICs) or barrier effects (MIC ratios) from molecular descriptor values. At each node of a tree, RF determines which descriptor from a randomly selected subset of descriptors best separates the antibiotics in the training set classified as "high" from those that are classified as "low". Each time a descriptor is selected as the best splitter, a best split value, or threshold (T), is obtained based on the descriptor values for that subset of antibiotics. The threshold values for each descriptor were averaged over all occurrences in each model (T_{avg}) to obtain general guidelines for desirable descriptor values.

R version 3.3.2[209] was used to perform the RF analysis of molecular descriptors. RF scripts were adapted from Richter et al., and classification was performed with *caret* using tenfold cross validation repeated ten times on a set of 2000 decision trees. The RF classification models were assessed with receiver operating characteristic (ROC) curves and confusion matrices. The top 20 descriptors for each set of response variables (MICs or MIC ratios) were determined by quantifying the overall variable importance of the machine learning model using the out-of-bag error, i.e., the decrease in classification accuracy when a single descriptor is removed. Scatter plots containing the $T_{avg,}$ T_{min}, and T_{max} values were generated with *ggplot2* for the top 20 descriptors with natural log-transformed MICs and MIC ratios.

Ensemble docking of primary amines to AcrA

We used the Tranche Browser to search a subset of the ZINC 15 database [160] for 3D representations of in-stock primary amines with "standard" reactivity, molecular charge in the range of -2 to 2 at pH "ref" or "mid", log P between 2.5 and 3, and no molecular weight cutoff, resulting in ~1.8M

small molecules (as of November 26, 2018). Alternate protonation and tautomeric states at pH 7.4 were included for each molecule. We then selected for primary amines, resulting in 22,842 compounds. From this list we selected molecules that are relatively rigid (number of rotatable bonds ≤ 4), polar (dipole moment ≥ 5.5 D), amphiphilic (amphiphilic moment ≥ 4.0 $Å^2$), and have low globularity (≤ 0.14).[67, 201] Molecular properties were calculated with MOE 2016.[206]

Previously, we generated a full-length model of AcrA from *E. coli*, performed a 50-ns molecular dynamics simulation of the model, and extracted 29 representative conformations using RMSD clustering.[210] In the present work, ensemble docking[211] to each of these 29 conformations was performed at four sites (E67 (site I), K241 (site II), I343 and I252 (site III), and F81 and F254 (site IV), **Figure 34B**) with VinaMPI using a 25 Å x 25 Å x 25 Å docking search space and an exhaustiveness of 10.[212] Of the resulting ~50 top compounds selected based on docking score at any site, commercially available compounds were purchased from ChemBridge (San Diego, CA).

Results and Discussion

MICs and MIC ratios

To determine how active efflux and the OM permeability barrier contribute to the activities of antibiotics, we selected 64 representatives of the BLs (cephalosporins (CEFs), penicillins (PENs), and meropenem) and FQs. These antibiotics differ significantly in their structures and properties, ranging in molecular mass from less than 300 Da to 650 Da, with log $D_{7.5}$ values from -3 to ~4 and log $P_{(o/w)}$ values varying from -2 to 3.5. In addition, we included a few representatives belonging to other classes of antibiotics that have been analyzed previously[57, 58]: two macrolides (azithromycin and erythromycin), the activities of which were strongly affected both by active efflux and OM permeability, chloramphenicol, which was weakly affected by both efflux and the OM barrier, and gentamicin, the activity of which was not affected by the OM.

MICs of antibiotics were measured in *E. coli* WT, the efflux-deficient variant ΔTolC, and the pore-producing derivatives WT-Pore and ΔTolC-Pore.[58] For *P. aeruginosa,* four strains were also analyzed: the wild type PAO1, strain PΔ6 lacking six efflux pumps (Δ*mexAB-oprM*, Δ*mexCD-oprJ*, Δ*mexXY*, Δ*mexJKL*, Δ*mexEF-oprN,* Δ*triABC*), and their pore-producing derivatives, PAO1-Pore and PΔ6-Pore, respectively.[213] All strains were previously shown not to have significant

growth defects, and to differ dramatically in their susceptibilities to various classes of antibiotics. All strains were previously shown not to have significant growth defects, and to differ dramatically in their susceptibilities to various classes of antibiotics.

For *E. coli* WT cells, MICs could be measured for all tested antibiotics, whereas the MICs of ~30% of the antibiotics were too high to be determined in *P. aeruginosa* PAO1 cells (**Figure 35** and **Figure 36**). However, the MICs of all antibiotics could be determined in PΔ6-Pore, highlighting the large contribution of the permeability barriers in this species toward antibiotic activities. To normalize to the differences in biochemical potency among compounds, our key measured parameters were efflux ratios and OM barrier ratios, defined as $MIC_{parent}/MIC_{mutant}$, for efflux mutants and hyperporinated mutants, respectively.

Cephalosporins

These antibiotics were relatively potent against *E. coli*, with the lowest MICs against WT in the low nanomolar range, but not against *P. aeruginosa,* for which the most potent representatives had MICs in the mid to low micromolar range (**Figure 35** and **Figure 36**). This gap in CEF potency can be attributed to both the species-specific differences in permeability barriers and the expression of chromosomal BLs in *P. aeruginosa* strains.[214, 215] The combination of these two factors resulted in about half of the analyzed CEFs lacking appreciable activities against the wild-type PAO1 strain. However, in both species the activities of almost all CEFs were potentiated by hyperporination of the OM, inactivation of efflux, or both, albeit to different degrees. As a result, besides the BLs, all CEFs had a measurable MIC in the minimal barrier PΔ6-Pore strain with the most potent activities in the mid nanomolar range. This result suggests that in *P. aeruginosa* strains, the permeability barriers are synergistic with BLs and contribute significantly to resistance against these antibiotics.

Interestingly, in both species CEFs were modestly (≤ fourfold) affected by efflux deletions. In *P. aeruginosa,* the exceptions were ceftibuten (8-fold), cefotaxime and cefepime (16-fold) and cefmenoxime (64-fold), whereas in *E. coli,* cefuroxime was potentiated 16-fold upon efflux activation (**Figure 37**). In contrast to efflux, the effect of hyperporination was species-specific. With a few exceptions, the hyperporinated *E. coli* cells were only slightly more susceptible to

CEFs (≤ fourfold) than WT cells (**Figure 35** and **Figure 37**). Cefonicid, cefoperazone, and cefuroxime were among the most limited (16-fold) by the *E. coli* OM barrier. In *P. aeruginosa* on the other hand, the effect of hyperporination was more drug specific. Some CEFs were not significantly affected by hyperporination in *P. aeruginosa* (e.g., cefdinir and cefalonium), whereas others were significantly limited by the OM barrier (e.g., cefotaxime, and cefmenoxime).

The increased potency of CEFs in the barrierless strains highlights the synergistic effect of active efflux and the OM barrier. In most cases, *P. aeruginosa* PΔ6-Pore cells were ≥ 16-fold more susceptible to CEFs than the wild-type PAO1 cells, with the exceptions of cefalonium, cefoperazone, cefprozil, and cefuroxime. Ceftriaxone and cefmenoxime were the most potentiated CEFs (≥ 256-fold). In contrast, most CEFs were not affected significantly (≤ fourfold) by removal of the OM barrier and inactivation of efflux in *E. coli* (e.g., cefepime and cefalonium). Other CEFs such as cefdinir and cefuroxime had activities that were potentiated greater than 32-fold upon removal of both barriers (**Figure 36** and **Figure 37**).

Penicillins

Like CEFs, these antibiotics differ significantly in their activities against *E. coli* and *P. aeruginosa* and are affected by the OM, active efflux, and BLs. In general, the MICs of PENs in both species were in the micromolar and millimolar ranges. However, unlike CEFs, the effects of active efflux and hyperporination on the activity of PENs spanned orders of magnitude in both species.

PENs were either poor substrates of *P. aeruginosa* efflux pumps (e.g., penicillin and amoxicillin) or excellent substrates (e.g., ampicillin and methicillin) (**Figure 35** and **Figure 37**). Likewise, in *E. coli* PENs were either poor substrates of efflux pumps in *E. coli* (e.g., amoxicillin and ampicillin), or excellent substrates (e.g., cloxacillin and dicloxacillin), the activities of which increased by 512-fold upon the removal of efflux compared to WT (**Figure 36** and **Figure 37**). In both *P. aeruginosa* and *E. coli*, most PENs were significantly limited by the OM barrier. Some PENs, such as piperacillin, azlocillin, and nafcillin were minimally affected by hyperporination of PAO1-Pore. In contrast, activities of other PENs including ampicillin, methicillin, and cloxacillin were all potentiated by ≥ 16-fold. In *E. coli*, greater than fourfold increases in potentiation were observed in hyperporinated cells compared to WT cells for all PENs except ampicillin,

carbenicillin, and hetacillin, with maximal increases in activity of 64-fold for piperacillin and azlocillin.

PΔ6-Pore cells were in general ≥ 16 times more susceptible to all PENs (**Figure 37**). The activities of several PENs were highly limited by both barriers in *P. aeruginosa,* including carbenicillin, nafcillin, and azlocillin (≥ 10^3-fold). The susceptibility of ΔTolC-Pore cells increased at least eightfold compared to WT cells for all PENs, with maximum increases of 512-fold for cloxacillin, dicloxacillin, nafcillin, and oxacillin. Thus, the structural differences of PENs and CEFs lead to dramatic effects in both the antibiotic permeation across the OM and active efflux avoidance.

Fluoroquinolones

Unlike BLs, FQs are highly potent against both species, but on average *P. aeruginosa* PAO1 was 16-fold less susceptible to these antibiotics than *E. coli* WT. Accordingly, the FQ MICs in PAO1 were in the low micromolar to high nanomolar range, whereas in *E. coli,* with a few exceptions, the MICs of FQ were in the sub micromolar range.

FQs were relatively good substrates of efflux pumps in both species. In *P aeruginosa,* most FQs were potentiated by at least 16-fold upon efflux deletion, except for difloxacin (fourfold). In *E. coli,* FQs such as sparfloxacin and nadifloxacin were good substrates of efflux pumps, as evidenced by the potentiation of their activities in ΔTolC cells (64-fold and 256-fold, respectively) (**Figure 35**). Other FQs such as prulifloxacin and sarafloxacin only showed a fourfold increase in susceptibility upon removal of efflux capabilities. Unlike BLs, most FQs were not significantly affected by the removal of the OM barrier (≤ fourfold) in either species. However, the susceptibility of PAO1-Pore cells toward moxifloxacin and nadifloxacin increased 16-fold compared to PAO1, whereas WT-Pore was 16-fold more susceptible than WT for nadifloxacin (**Figure 36** and **Figure 37**).

However, even such small changes in the permeation across the OM contributed significantly to the FQ potency when synergized with active efflux. In general, PΔ6-Pore cells were ≥ 64 times more susceptible to FQs than WT. Furthermore, the activities of certain FQs were potentiated by more than a thousand-fold in PΔ6-Pore cells compared to PAO1 cells. For example, among the

FQs flumequine and nadifloxacin were potentiated ≥ 10^3-fold. In *E. coli* ΔTolC-Pore, most FQs displayed eightfold or greater potentiation, although norfloxacin, pefloxacin, pazufloxacin, prulifloxacin and sarafloxacin were potentiated by only fourfold upon removal of both barriers.

Taken together, these results show that in *P. aeruginosa* and *E. coli* strains the MICs of tested antibiotics ranged from millimolar to sub-nanomolar values, and their potencies varied dramatically depending on the presence or absence of one or both permeability barriers.

Molecular property fingerprints of MICs and MIC ratios

We used RF classification to dissect the specific effects of OM permeability and efflux that limit antibiotic activity. Specifically, we identified the most important physicochemical properties, i.e. those that resulted in the largest decrease in accuracy upon removal, for classifying the relative potency of the antibiotics in mutant strains and wild-type strains (as measured by MIC) and the dependence of the antibacterial activities on efflux, the OM barrier, or both (as measured by MIC ratios). We performed RF classification on 143 descriptors for *P. aeruginosa* and 142 descriptors for *E. coli* (**Table 9**). *E. coli* MICs ≤ 4 μM were classified as low, i.e. active. However, because *P. aeruginosa* shows greater resistance to antibiotics, antibiotics classified as low had MICs ≤ 20 μM. All other MICs were classified as high (i.e., inactive). For both species, MIC ratios ≤ 4 were classified as low, or having no significant barrier effect, and MIC ratios > 4 were considered to show a significant barrier effect.

The most important molecular descriptors identified by RF classification provide a "fingerprint" that describes the molecular characteristics that best distinguish between high and low classifications for each set of MICs or MIC ratios. The descriptors belong to eight aggregate categories: charge, connectivity, molecular topology number of selected atom or bond types, physical properties, potential energy, shape, and surface area (**Table 9**).

Charge Properties

These descriptors quantify electrostatic properties in a molecule or portion of a molecule. Partial charges calculated with the partial equalization of orbital electronegativity (PEOE) method[216] are the most prevalent charge descriptors identified from the RF analysis. Several of these descriptors

map charges to specific van der Waals surface area (VSA) regions (e.g., fractional negative, or total negative VSA). This category is prominent for both *P. aeruginosa* and *E. coli* MICs and MIC ratios. Charge properties are abundant in the top descriptors for *P. aeruginosa*, comprising 8-10 of the top 20 descriptors for each strain. However, for *E. coli* MICs, charge descriptors are less abundant (≤ 7 of the 20). Compared to MICs, the MIC ratios feature fewer charge descriptors (4-8 of the top 20 for *P. aeruginosa* MIC ratios, and 2-5 for *E. coli*). For both *E. coli* and *P. aeruginosa*, efflux ratios (PAO1/PΔ6, PAO1-Pore/PΔ6-Pore, WT/ΔTolC, and WT-Pore/ΔTolC-Pore) include the most charge descriptors, suggesting that electrostatic properties may help in distinguishing between antibiotics that are significantly limited by the efflux barrier and those that are not.

Atom connectivity, shape, and molecular topology

Connectivity and topological descriptors represent molecules as graphs in which vertices correspond to atoms, and edges correspond to bonds. These descriptors are not abundant in the fingerprints of either species. ΔTolC-Pore is the only strain that contains a connectivity descriptor, along with the MIC ratios WT-Pore/ΔTolC-Pore and PAO1/PΔ6-Pore. This category is absent from the top 20 list in *P. aeruginosa* MICs and appears only sparingly among all *E. coli* strains except for ΔTolC-Pore. ΔTolC/ΔTolC-Pore, PAO1/PΔ6-Pore, PAO1/PAO1-Pore, and PAO1-Pore/PΔ6-Pore also contain shape descriptors. In contrast to shape and connectivity descriptors, most MIC or MIC ratio fingerprints contain topological descriptors. Five of the top 20 descriptors for WT-Pore are topological, incorporating measures of the partition coefficient, partial charges, or polarizability. However, the descriptor fingerprints for other *E. coli* strains contain two topological descriptors. This category is highly abundant in the *E. coli* ratios WT/ΔTolC-Pore and WT/WT-Pore, (8 and 7 of the top 20, respectively), indicating that molecular topology may be relevant for distinguishing antibiotics that are severely limited by hyperporination in the presence and absence of efflux compared WT.

Atom and bond counts

Examples of these descriptors include the numbers of hydrogen bond donor atoms (a_{don}), aromatic rings, and oxygen atoms, as well as measures of flexibility in the form of total and fractional rotatable bond counts. The fingerprints for all strains of *P. aeruginosa* contain a single atom or

bond count descriptor, and all *E. coli* strains include 2-4 descriptors in this category. These descriptors are present in the fingerprint of all ratios except for WT/WT-Pore and PAO1/PΔ6-Pore.The top descriptor lists for ΔTolC-Pore and WT-Pore ΔTolC-Pore each include four atom or bond count descriptors. Thus, certain descriptors in this category may be useful for classifying antibiotic activity in the absence of efflux in *E. coli*.

Physical Properties

Physical properties such as molecular weight, solubility coefficient, and partition coefficient, are commonly considered in drug design. Descriptors of this type are present for all MICs except PΔ6. *E. coli* MIC ratios all list 2-4 descriptors physical property descriptors, and the ratios WT/ΔTolC-Pore and WT/WT-Pore have the most physical descriptors. The trend is similar for *P. aeruginosa*, with PAO1/PΔ6-Pore and PAO1/PAO1-Pore having the most physical descriptors among *P. aeruginosa* MIC ratios. Thus, common metrics used in rational drug design classify antibiotics in the OM ratios better for *E. coli* than for *P. aeruginosa*.

Potential Energy Descriptors

Potential energy descriptors quantify energetic contributions from, for example, van der Waals (VDW) effects or solvation. Similar to physical properties, these descriptors appear in the fingerprints for all MICs and MIC ratios in both species at least once. The wild-type and hyperporinated strains contain two potential energy descriptors in both species, but the efflux-deficient strains contain only a single descriptor in this category.

Surface Area Properties

Surface area (SA) descriptors are the second most abundant descriptor category, with all MIC and ratio fingerprints including 2-8 occurrences. Many of these descriptors are based on either the total or subdivided VSA of a molecule combined with another property such as lipophilicity (SlogP and log $P_{(o/w)}$), hydrophobicity, shape, or connectivity. Compared to *E. coli*, the respective ratios in *P. aeruginosa* have more descriptors in this category, highlighting the differences between the two species. The descriptor fingerprints for hyperporinated strains in both species contained five SA descriptors, suggesting that these descriptors can distinguish active from inactive antibiotics in the presence of only the efflux barrier.

P. aeruginosa and E. coli permeability barriers select for different molecular properties

To provide a set of descriptor guidelines that are favorable for antibiotic potency in the presence and absence of efflux and OM barriers, we selected from among the top descriptors those that trend with MICs and/or MIC ratios. To identify the "optimal" descriptor values for active antibiotics, we determined the average best split value or threshold (T_{avg}) that separates high and low MICs or ratios in the RF analysis. As in previous studies, antibiotics were grouped on the basis of how their antibacterial activities were affected by the OM barrier and active efflux barriers. [57]

The OM barrier

Antibiotic properties that are favorable for OM permeation were captured in the MIC fingerprints of the efflux-deficient PΔ6 and ΔTolC strains and in the fingerprints of the MIC ratios for cells with hyperporinated and intact OM barriers, i.e., PΔ6/PΔ6-Pore, PAO1/PAO1-Pore, ΔTolC/ΔTolC-Pore and WT/WT-Pore. Antibiotics with high OM barrier ratios were classified as being strongly affected by permeation across the OM. This group includes BLs, predominantly CEFs.

The properties that trend with changes in MICs in both ΔTolC and PΔ6 strains are rigidity and VDW potential energy (E_{vdw}) (**Table 10**). The most potent antibiotics have high values for E_{vdw}, with average threshold values of 4.1 kcal mol^{-1} for ΔTolC and 4.6 kcal mol^{-1} for PΔ6. Molecular rigidity was quantified here as the fraction of rotatable bonds ($b_{rotR} = b_{rotatable}/b_{total}$). Rigid antibiotics (i.e., b_{rotR} below T_{avg} of 0.2) were more active against both *E. coli* and *P. aeruginosa* in the absence of efflux, suggesting that less flexible molecules more readily overcome the OM barrier in both species. Rigidity was previously found to be important for increasing accumulation in wild-type *E. coli*.[217]

As mentioned previously, charge properties are more abundant in the *P. aeruginosa* fingerprints. Several charge descriptors trend with antibiotic potencies in PΔ6 but not in ΔTolC cells. In general, the most effective antibiotics have larger dipole moments (T_{avg} = 5.5 D), suggesting that molecules with greater charge separation more readily permeate the OM in *P. aeruginosa*. Both fractional positive water accessible SA (FASA+) and fractional negative VSA ($PEOE_{VSA\ FNEG}$) provide

information about favorable charge distributions in these molecules. Antibiotics with a high FASA+ (T_{avg} = 0.5) and a low $PEOE_{VSA\ FNEG}$ (T_{avg} = 0.5) generally have low MICs. However, these descriptors trend with MICs in other *P. aeruginosa* strains as well (**Table 10**), suggesting their importance in antibiotic permeation across both barriers.

From the high PAO1/PAO1-Pore ratios, the OM barrier of *P. aeruginosa* counterselects for antibiotics with a high principal moment of inertia in the Y direction (pmiY, T_{avg} = 2930). PmiY is not present in the WT/WT-Pore fingerprint and, therefore, is specific for the *P. aeruginosa* OM barrier. Antibiotics that show a significant barrier effect in PΔ6/PΔ6-Pore generally have high negative and fractional negative VSA ($PEOE_{VSA\ NEG}$, T_{avg} = 141 Å^2 and $PEOE_{VSA\ FNEG}$, T_{avg} = 0.4), and low VSA with SlogP values in the range of 0 to 0.1 ($SlogP_{VSA3}$, T_{avg} = 62.8 Å^2). The same trends are evident for $PEOE_{VSA\ NEG}$ (T_{avg} = 210 Å^2) and $SlogP_{VSA3}$ (T_{avg} = 20.5) in the corresponding *E. coli* ΔTolC/ΔTolC-Pore ratio.

Thus, several descriptors that trend with MICs or MIC ratios are common for the OM barriers of both *E. coli* and *P. aeruginosa*. For example, rigid molecules more readily permeate the OM in both species. On the other hand, charge and shape descriptors are characteristic only for the OM barrier in *P. aeruginosa*.

The active efflux barrier

Descriptors that quantify the effect of the active efflux barrier are present in the fingerprints for PAO1-Pore and WT-Pore MICs. They are also present in MIC ratios PAO1-Pore/PΔ6-Pore, PAO1/PΔ6 in *P. aeruginosa*, and WT-Pore/ΔTolC-Pore and WT/ΔTolC in *E. coli*. In both species, FQs are the dominant antibiotics limited by this barrier. Meropenem is known to cross the OM of *P. aeruginosa* using the amino acid-specific channel OprD[218]. This carbapenem is a substrate of the *P. aeruginosa* efflux pumps. In contrast, in *E. coli* meropenem potency is strongly limited by OM permeation, but not by efflux.

For PAO1-Pore, active antibiotics have a balance of fractional positive and negative SA, with high FASA+ (T_{avg} = 0.5) and low $PEOE_{VSA\ FNEG}$ (T_{avg} = 0.5) and a high dipole moment (T_{avg} = 5.6 D). As

observed with the OM barrier, antibiotics with the greatest activity against PAO1-Pore are more rigid (b_{rotR} T_{avg} = 0.2).

The top descriptors, that trend with the PAO1/PΔ6 ratio, are primarily charge descriptors. Interestingly, relative positive partial charges ($PEOE_{RPC+}$, T_{avg} = 0.1) trend with both the PAO1/PΔ6 and PAO1-Pore/PΔ6-Pore efflux ratios, but not with the OM ratios. Thus, *P. aeruginosa* efflux pumps may select for antibiotics with high positive partial charges. In addition, rigid antibiotics have higher PAO1/PΔ6 ratios, suggesting that *P. aeruginosa* efflux pumps may favor rigid molecules. Lipophilicity trends with PAO1-Pore/PΔ6-Pore ratios. The more lipophilic molecules tend to have higher values for this ratio. Thus, a metric such as SlogP (T_{avg} = -0.53), which is often considered for maximizing membrane permeability, may also promote efflux in *P. aeruginosa*.

In *E. coli*, active antibiotics in WT-Pore have E_{vdw} values above ~4 kcal mol^{-1}. However, charge descriptors are mostly absent for this strain and do not show any notable trends with MICs or ratios in *E. coli* (**Table 10**). Antibiotics that are significantly limited by active efflux in the presence of the OM barrier in *E. coli* (WT/ΔTolC) have log $P_{(o/w)}$ and SlogP values greater than 1.3 and -0.7 respectively, indicating that lipophilicity plays a role in efflux pump specificity. Similarly, antibiotics with high WT-Pore/ΔTolC-Pore ratios have log $P_{(o/w)}$ and SlogP values greater than the average thresholds of 1.5 and -0.5, respectively.

Thus, the lipophilic properties identified by RF for active efflux are similar in both species. These descriptors may be useful for guiding the prediction of antibiotic potencies and the effects of efflux. In addition, partial positive charges in antibiotics are selected by active efflux in *P. aeruginosa*.

OM and active efflux synergy

Some antibiotics are strongly affected both by hyperporination and efflux inactivation. In both species, FQs, macrolides and BLs are all included in this group. Antibiotics with activities that were significantly affected by the removal of both barriers (PAO1/PΔ6-Pore and WT/ΔTolC-Pore) generally have positive log $P_{(o/w)}$ values (**Table 10**). Interestingly, the average threshold for log $P_{(o/w)}$ in WT/ΔTolC-Pore (T_{avg} = 0.7) is higher than for PAO1/PΔ6-Pore (T_{avg} = -0.9). Lipophilicity

(SlogP) values greater than the average threshold are associated with high PAO1/PΔ6-Pore and WT/ΔTolC-Pore ratios with values of -0.6 and -0.9, respectively. In *E. coli*, b_{rotR} is an important feature for describing the differences in antibiotic effectiveness between the maximal and minimal barrier strains. In general, rigid molecules have lower WT/ΔTolC-Pore ratios ($T_{avg} = 0.2$), and thus antibiotics below this threshold value were often limited by both barriers.

PΔ6-Pore and ΔTolC-Pore are minimal-barrier strains in which both barriers have been removed. The positive trends of MICs in these strains with the lipophilic descriptors $logP_{(o/w)}$ and SlogP and the rigidity descriptor b_{rotR} are consistent with the trends in the respective MIC ratios described above (**Table 10**). In addition, FASA+ is also a top descriptor for PΔ6-Pore, with active antibiotics having higher values than $T_{avg} = 0.5$.

Taken together, these results show that several top descriptors identified by RF classification trend with MICs and/or MIC ratios in *P. aeruginosa* and *E. coli* and that these descriptors vary between species and barriers. In several cases, optimizing for one barrier promotes strengthening of the other barrier, suggesting a synergistic relationship between the OM and efflux barriers. For example, rigid antibiotics (e.g., FQs) are more active against PΔ6 and thus, more readily permeate the OM in *P. aeruginosa*, but these antibiotics are also often excellent substrates of efflux pumps. The activity of antibiotics in *P. aeruginosa* is primarily captured by charge and surface area descriptors, whereas properties identified for the *E. coli* barriers point to the role of topology, physical properties, and atom or bond counts. The calculated threshold values of these descriptors provide guidelines that may be useful for selecting or designing antibiotics with favorable properties to overcome these barriers.

Chemical structure and descriptor relationships

The top descriptors described above are sensitive to small changes in the chemical structures of antibiotics. For example, the CEFs ceftriaxone and cefepime (**Table 11**) differ by the cyclic substitutions at position 3 of the cephalosporin nucleus. These two antibiotics were active against both *E. coli* and *P. aeruginosa*, but differed significantly in the effects of efflux and the OM barrier (**Figure 35** and **Figure 36**). Cefepime, with a methylpyrrolidine substitution at position 3, was not sensitive to efflux or the OM barrier in *E. coli* but was affected strongly by both barriers in *P.*

aeruginosa. In the case of *P. aeruginosa,* cefmenoxime, which has a methyltetrazole group at this position, was further affected by both barriers with a thousand-fold increase in activity compared to WT. Although cefepime and cemenoxime have similar values for FASA+, $PEOE_{RPC+}$, a_{don}, and b_{rotR}, cefmenoxime has a higher fractional and total negative SA, whereas cefepime has a higher dipole moment, E_{vdw}, and lipophilicity (**Figure 38**).

The aminopenicillins amoxicillin and ampicillin differ by the presence of a hydroxy group at C4 in the phenyl ring (**Table 11**). These two aminopenicillins had high MICs in PAO1, but both are relatively active against WT (**Figure 35** and **Figure 36**). In *E. coli,* the activity of these antibiotics was mainly limited by the OM barrier. For *P. aeruginosa* either inactivation of efflux or hyperporination was required to obtain a measurable MIC for ampicillin. In contrast, both inactivation of efflux and hyperporination was needed to obtain a measurable MIC for amoxicillin. These two antibiotics have similar descriptor values for several properties such as b_{rotR}, $SlogP_{VSA3}$, and FASA+, but ampicillin has a higher dipole moment, lipophilicity, E_{vdw}, and pmiY (**Figure 38**). Amoxicillin has a higher a_{don} as a result of its additional hydroxyl group, which may decrease the effect of the OM barrier, at least in *E. coli.*

For FQs, difloxacin differs from sarafloxacin by the presence of a methyl group at N4 in the piperazinyl ring and sarafloxacin is a stronger base and is more hydrophilic (**Table 11**). Difloxacin has a higher total and fractional negative SA, SlogP, and pmiY, and a lower dipole moment than sarafloxacin (**Figure 38**). In *P. aeruginosa* both of these antibiotics were affected by the removal of both barriers, but only sarafloxacin was significantly limited by efflux alone. In both species hyperporination did not greatly limit antibiotic activity ($\leq$ 4-fold change) (**Figure 35**). However, in *E. coli* active efflux provided a significant barrier to overcome for difloxacin but not sarafloxacin (**Figure 36**).

The activities of most FQs were mainly limited by active efflux. The exceptions are moxifloxacin and nadifloxacin, which also showed OM barrier limitations and an even greater increase in activity in the minimal-barrier strains, reflecting the synergy of these barriers (**Table 11**). Several of the top descriptors differ between these two antibiotics. For example, moxifloxacin has a higher dipole moment, pmiY, and $SlogP_{VSA3}$, but nadifloxacin has a higher negative SA and an additional

hydrogen bond donor. Moxifloxacin, a fourth generation FQ belonging to the 6-hydrogenquinolones, bears a cyclopropyl group at N1 coupled with a C8 methoxy and a pyrrolopyridine at C7. The bulky heterocyclic group and overall lipophilicity strongly affect both the permeation across the OM and the effect of active efflux. Chemically, nadifloxacin has a lipophilic tricyclic benzoquinolizine core, with a 4-hydroxypiperidino moiety at the C8 position. This singular moiety lacks a distal basic functionality, which is unusual for a side chain of a quinolone, as all marketed quinolones bear side chains with a basic functionality, thereby providing two or three ionizable groups compared with only one for nadifloxacin ($pK_a = 6.8$). Thus, even small changes in the chemical structures of antibiotics such as a single substituent change can significantly affect one or more descriptors that capture the behavior of the OM barrier, active efflux, or both.

Identification of EPIs using physicochemical property filters

An important feature of OM-permeable compounds is the presence of a cationic amine, with primary amines being the most permeable. However, primary amines are relatively rare in chemical databases. For example, only ~0.1% of the ChemBridge Microformat Set contains this functional group.[67] Of the limited number of amines, antibacterials should also be relatively rigid, polar, amphiphilic, and have low globularity. To generate a focused library of molecules that could potentially serve as efflux pump inhibitors, we searched a subset of the ZINC database and performed additional filtering for compounds with appropriate properties. We then docked the resulting ~1,400 compounds to an ensemble of conformations of monomeric AcrA at four different potential binding sites and selected compounds based on docking score. Of the resulting ~50 top predicted binders, 34 commercially available compounds were purchased and tested experimentally.

To qualify as EPIs, compounds must satisfy at least three criteria: (i) they must enhance the activities of antibiotics that are effluxed in strains containing functioning pumps, (ii) they must not significantly potentiate the activities of antibiotics in strains that lack efflux pumps, and (iii) must interact with AcrA or AcrB.[219, 220] MPC_4 values to measure antibiotic potentiation of novobiocin and erythromycin and surface plasmon resonance (SPR) were used to identify six compounds that meet these criteria. Interestingly, the six compounds that potentiate antibiotics and bind AcrA all

are substituted 4(3-aminocyclobutyl)-pyrimidin-2-amine compounds. Each of these six compounds has low globularity (0.05-0.12), few rotatable bonds (2-4), relatively high dipole moment (5-10 D) and relatively high amphiphilic moment (~4.5-7). The most favorable docking score for each compound corresponds to binding at either site II or site III, which flank the β-barrel domain of AcrA (**Figure 39**).

We also tested the six top hits for their ability to potentiate the activity of novobiocin and erythromycin in four other Gram-negative pathogens. These compounds did not potentiate antibiotic activity in *P. aeruginosa* or *E. cloacae*. However, some of these compounds increased the efficacy of novobiocin and erythromycin as measured by MPC_4 in wild-type cells of both *A. baumannii* (up to 8-fold) and *K. pneumoniae* (up to 2-fold). This result suggests that these compounds have broad-spectrum activity and permeate the OM of *A. baumannii* better than the OM of *E. coli*.

Conclusions

The OM barrier and efflux synergistically limit antibiotic activity in Gram-negative bacteria. In the present study, we have used *P. aeruginosa* and *E. coli* strains with controlled OM permeability and varying efflux capacity to identify trends in the antibacterial activities of CEFs, PENs and FQs. In addition to high-affinity target binding, OM permeation should be maximized and efflux should be minimized to obtain optimal antibacterial activity. Using RF classification, we have identified properties that distinguish between antibiotics with high and low antibiotic activities (MICs) in the various strains and identify the properties related to the effect of altering the OM permeability and efflux barriers (MIC ratios).

The top descriptors for *P. aeruginosa* are dominated by electrostatic properties. Active antibiotics have a higher dipole moment and antibiotics with low dipole moments are significantly limited by the OM barrier. On the other hand, partial positive charges trend specifically with the efflux ratios. Rigid antibiotics and antibiotics with high VDW energies (E_{VDW}) were more active against both *E. coli* and *P. aeruginosa*. Lipophilicity (log $P_{(o/w)}$ and SlogP) trend positively with efflux ratios in both species and classify antibiotics in the OM ratios better for *E. coli* than for *P. aeruginosa*, likely the result of the more complex barriers in *P. aeruginosa*.

The physicochemical properties identified here reflect chemical bias toward CEFs, PENs and FQs. However, performing a similar predictive analysis on diverse molecular libraries is expected to be beneficial for identifying new antibiotics or EPIs. The differences in rules between *P. aeruginosa* and *E. coli* suggest that different properties may need to be targeted for optimization of antibiotics against different species. Furthermore, we have shown that the properties selected as important for the OM barrier ratios differ significantly from those for the efflux ratios and that these two barriers can work together to develop resistance to antibiotics. Therefore, we recommend that antimicrobials should be optimized to evade efflux and enhance OM permeability simultaneously.

Using existing physicochemical guidelines as filters in combination with ensemble docking, in vitro binding studies, and *in vivo* potentiation assays in bacterial strains with controllable permeability barriers, we identified a new class of EPIs with activity against several Gram-negative bacteria. Six molecules with a shared scaffold were found to potentiate the antibiotic activity of erythromycin and novobiocin in hyperporinated *E. coli* cells and in wild-type strains of both *A. baumannii* and *K. pneumoniae*.

Appendix IV

Tables

Table 9. Descriptor categories and number of descriptors in each category

Descriptor category	Number of Descriptors *P. aeruginosa*	*E. coli*
charge	33	32
connectivity	1	1
molecular topology	18	18
n_{atoms}, n_{bonds}	17	16
physical properties	14	14
potential energy	9	10
shape	9	9
surface area	42	42
total	**143**	**142**

Table 10. Average threshold values for top descriptors that trend with MICs and MIC ratios

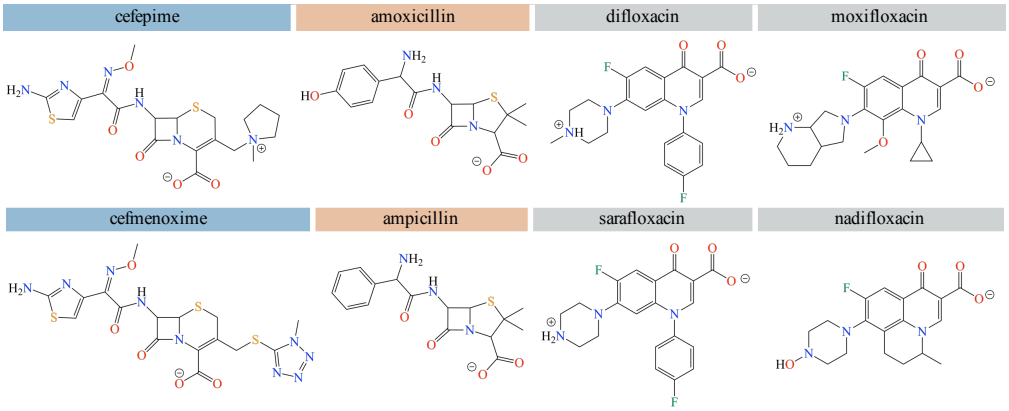
	MICs								MIC ratios									
	P. aeruginosa				***E. coli***				***P. aeruginosa***					***E. coli***				
Descriptors	PΔ6-Pore	PΔ6	PAO1-Pore	PAO1	ΔTolC-Pore	ΔTolC	WT-Pore	WT	PAO1/PΔ6-Pore	PΔ6/PΔ6-Pore	PAO1/PAO1-Pore	PAO1-Pore/PΔ6-Pore	PAO1/PΔ6	WT/ΔTolC-Pore	ΔTolC/ΔTolC-Pore	WT/WT-Pore	WT-Pore/ΔTolC-Pore	WT/ΔTolC
Dipole		5.5[a]	5.6[a]	5.6[a]						5.5[a]								
FASA+	0.5[a]	0.5[a]	0.5[a]	0.5[a]									0.5[b]					
PEOE_RPC+												0.1[b]	0.1[b]					
PEOE_VSA-2	21.9[b]	17.0[b]	16.9[b]	15.3[b]	30.5[b]	21[b]	24.9[b]	20.4[b]			17.9[c]	26.7[a]						
PEOE_VSA_FNEG		0.5[b]	0.5[b]	0.4[b]					0.3[b]	0.4[b]		0.39[b]						
PEOE_VSA_NEG				170[b]	171[b]					141[b]		167[c]	170[c]		210[b]	210[b]		
a_don								1.5[b]		2.3[c]		2.3[a]						
b_rotR	0.2[b]	0.2[b]	0.2[b]			0.2[b]		0.2[b]					0.2[a]	0.2[a]			0.2[a]	
logP(o/w)					0.2[a]				-0.9[b]				0.8[b]	0.7[b]	-0.1[c]		1.5[b]	1.3[b]
SlogP					-0.9[a]				-0.6[b]			-0.5[b]		-0.9[b]			-0.5[b]	-0.7[b]
E_vdw		4.6[a]	6.1[a]	6.6[a]		4.1[a]	3.8[a]	5.4[a]									8.8[b]	
pmiY							2140[c]				2930[b]							
SlogP_VSA3			39.4[a]	48.5[a]						62.8[a]					20.5[a]			

[a] Descriptor negatively trends with MICs (i.e., active antibiotics generally have descriptor values above T_{avg}) or MIC ratios (i.e., significant barrier effects are often observed in antibiotics with descriptor values below T_{avg}).

[b] Descriptor positively trends with MICs (i.e., active antibiotics generally have descriptor values below T_{avg}) or MIC ratios (i.e., significant barrier effects are often observed in antibiotics with descriptor values above T_{avg}).

[c] No trend between descriptor and MIC or MIC ratio values.

Table 11. Example CEFs, PENs, and FQs highlighting how small changes in antibiotic structure contribute to differences in MIC ratios and molecular descriptors. See ref 201 for additional details on experimental methods



		MIC Ratios						Descriptors												
Descriptor Category: charge; connectivity; molecular topology; n_{atoms}, n_{bonds}; physical; potential energy; shape; surface area		*P. aeruginosa*			*E. coli*															
	Antibiotic	PAO1/PΔ6-Pore	PAO1/PΔ6	PAO1/PAO1-Pore	WT/ΔTolC-Pore	WT/ΔTolC	WT/WT-Pore	dipole	FASA+	PEOE_RPC+	PEOE_VSA-2	PEOE_VSA_FNEG	PEOE_VSA_NEG	a_don	b_rotR	logP(o/w)	SlogP	E_vdw	pmiY	SlogP_VSA3
CEF	cefepime	64	16	16	4	2	2	5.5	0.60	0.07	30.6	0.26	118	2	0.23	-0.3	-1.3	4.7	4136	72.6
	cefmenoxime	1024	64	16	8	1	1	3.4	0.49	0.10	40.1	0.40	175	2	0.25	-1.1	-1.8	-5.1	3456	34.5
PEN	amoxicillin	64	1	1	8	1	8	3.8	0.49	0.10	19.8	0.53	177	3	0.19	0.6	-1.2	2.7	3021	0
	ampicillin	≥64	16	≥64	8	1	4	5.2	0.47	0.12	19.8	0.56	182	2	0.19	0.9	-0.9	11.4	3182	0
FQ	difloxacillin	16	4	4	16	16	4	6.6	0.60	0.11	0	0.43	161	0	0.09	2.6	1.7	14.1	2104	73.8
	sarafloxacin	64	16	1	4	4	1	11.6	0.59	0.12	0	0.39	138	0	0.10	2.3	0.3	13.4	1949	73.8
	moxifloxacin	256	64	16	64	16	4	12.4	0.69	0.12	0	0.33	119	0	0.12	2.0	-0.2	12.6	5239	55.3
	nadifloxacin	4096	64	16	1024	256	16	6.9	0.61	0.13	0	0.45	145	1	0.07	2.0	0.8	15.0	1145	36.9

Figures

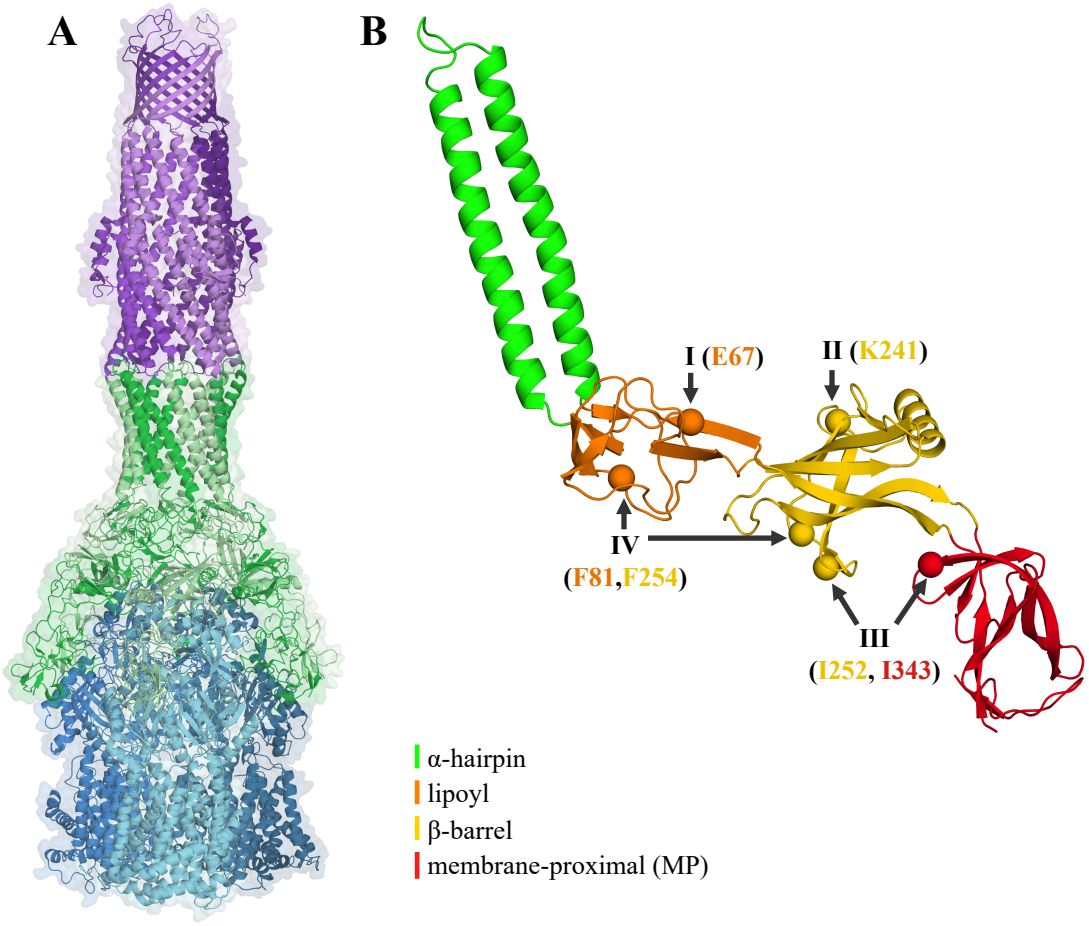


Figure 34. Structure of the *E. coli* efflux pump AcrAB-TolC. (**A**) Cryo-EM structure of the AcrAB-TolC complex (PDB Entry 5NG5) shown in cartoon and surface representations. Individual subunits of TolC (purple), AcrA (dark green), and AcrB (blue) are shown. (**B**) Sites I-IV on AcrA used for docking are color coded by domain: α-hairpin (light green), lipoyl (orange), β-barrel (yellow) and membrane-proximal (red). Residues used to define the center of each site are shown as spheres and colored by domain.

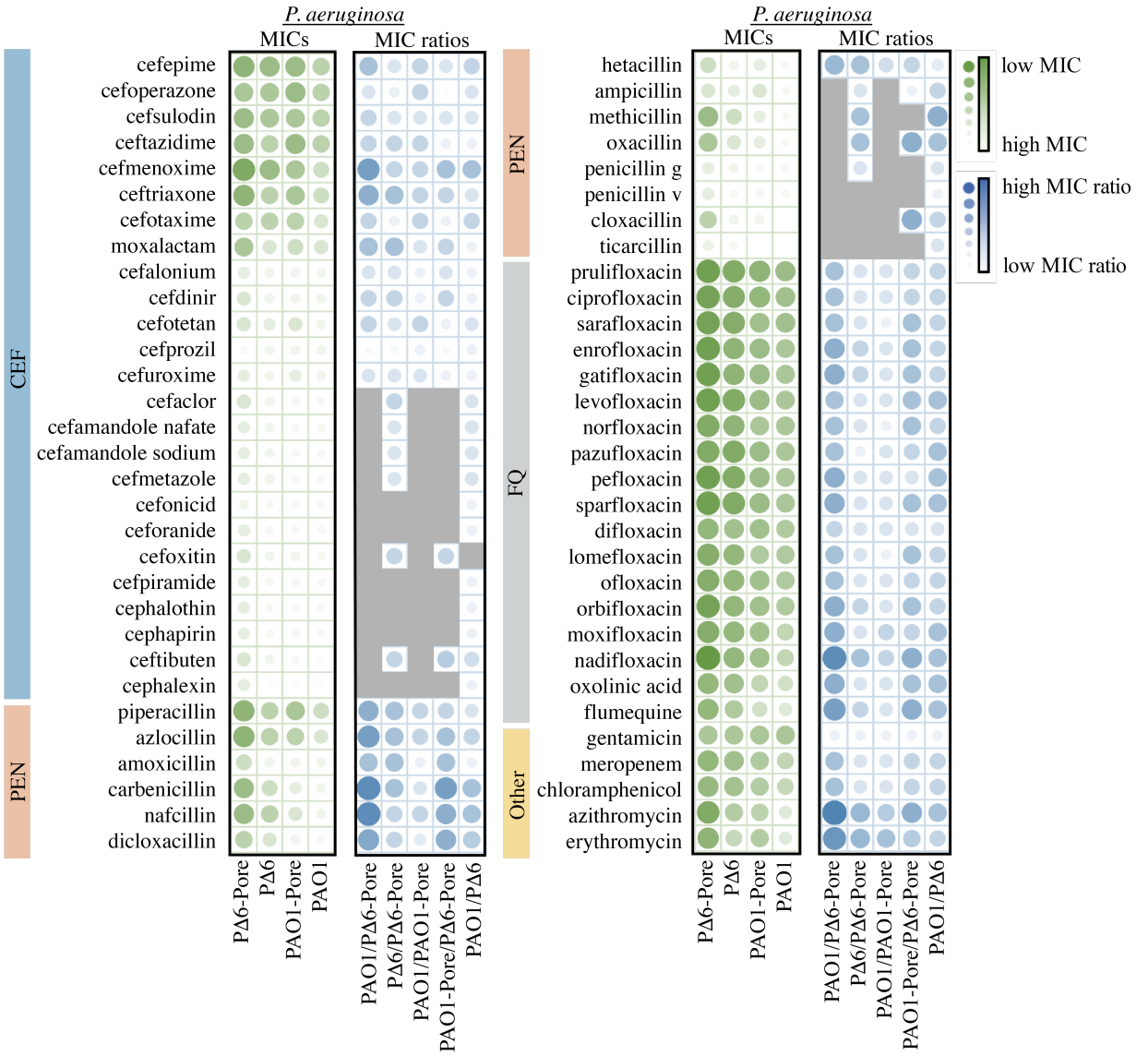


Figure 35. Scaled MICs and MIC ratios for antibiotics in *P. aeruginosa* sorted by antibiotic class and PAO1 MIC from lowest to highest within each class. Values were natural log-transformed and then scaled between 0 and 1. Gray squares indicate MIC ratios that are outside of the measurable range. MICs (green) report on relative potency and MIC ratios (blue) report on the dependence of antibiotic activity on efflux, the OM barrier, or both. See ref 201 for additional details on experimental methods.

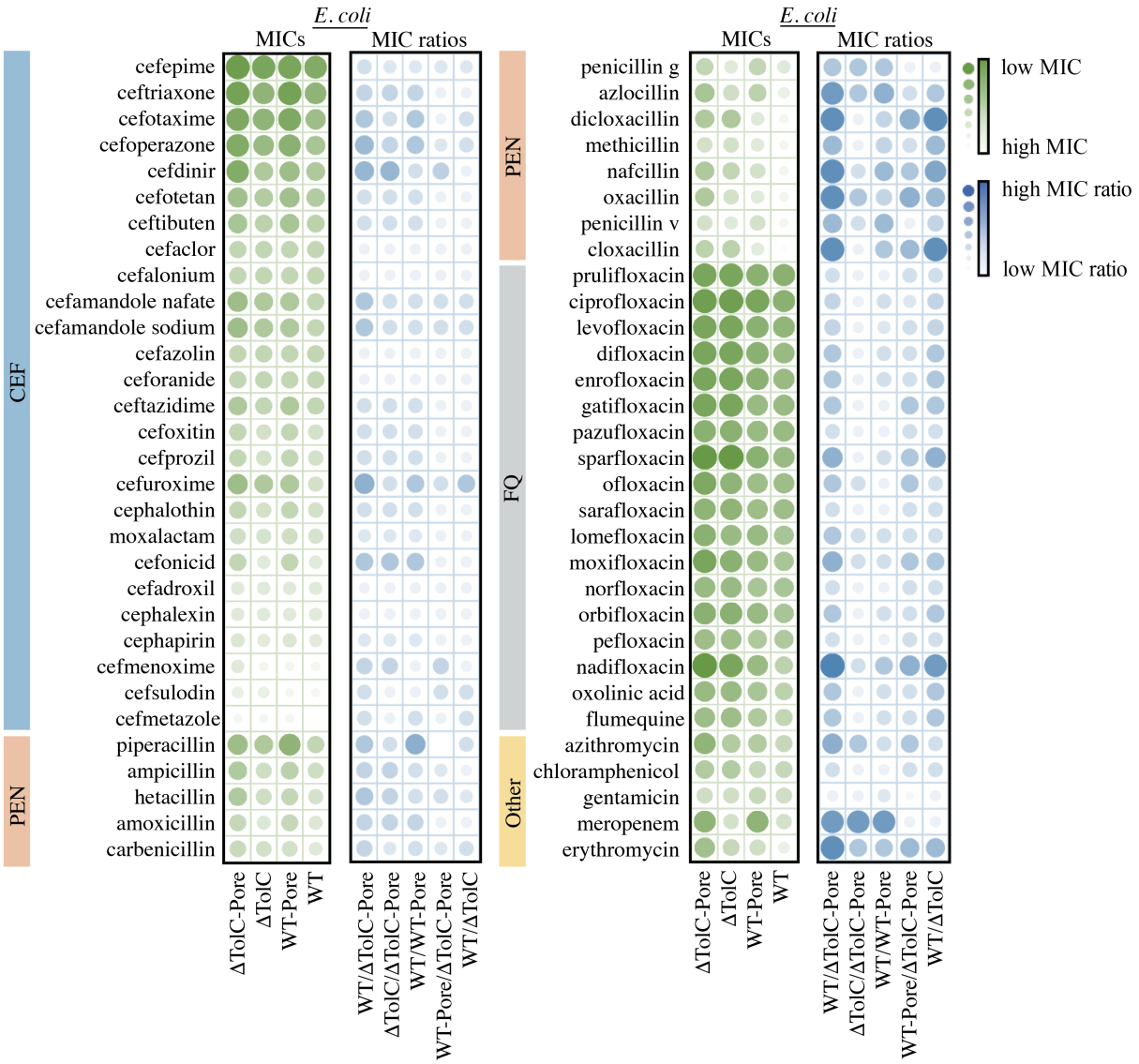


Figure 36. Scaled MICs and MIC ratios for antibiotics in *E. coli* sorted by WT MIC from lowest to highest. Values were natural log-transformed and then scaled between 0 and 1. MICs (green) report on relative potency and MIC ratios (blue) report on the dependence of antibiotic activity on efflux, the OM barrier, or both. See ref 201 for additional details on experimental methods.

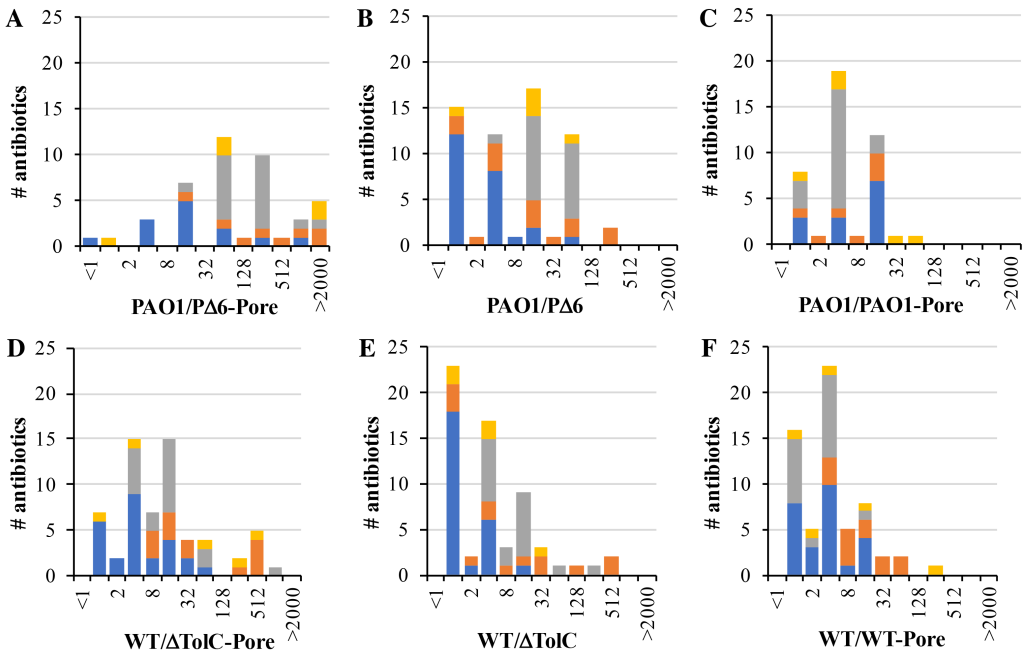


Figure 37. *P. aeruginosa* and *E. coli* MIC ratios colored by class (CEF=blue, PEN=orange, FQ=gray, other=gold) for (**A**) the "barrierless" ratio (PAO1/PΔ6-Pore), (**B**) only efflux pump deletion (PAO1/PΔ6), and (**C**) only hyperporination (PAO1/PAO1-Pore), (**D**) the "barrierless" ratio (WT/ΔTolC-Pore), (**E**) only efflux pump deletion (WT/ΔTolC), (**F**) only hyperporination (WT/WT-Pore). Only MIC ratios with measurable values are shown. The fold changes in MICs are shown on the X-axes and the number of antibiotics with the corresponding fold changes in MICs are on Y-axes. See ref 201 for additional details on experimental methods.

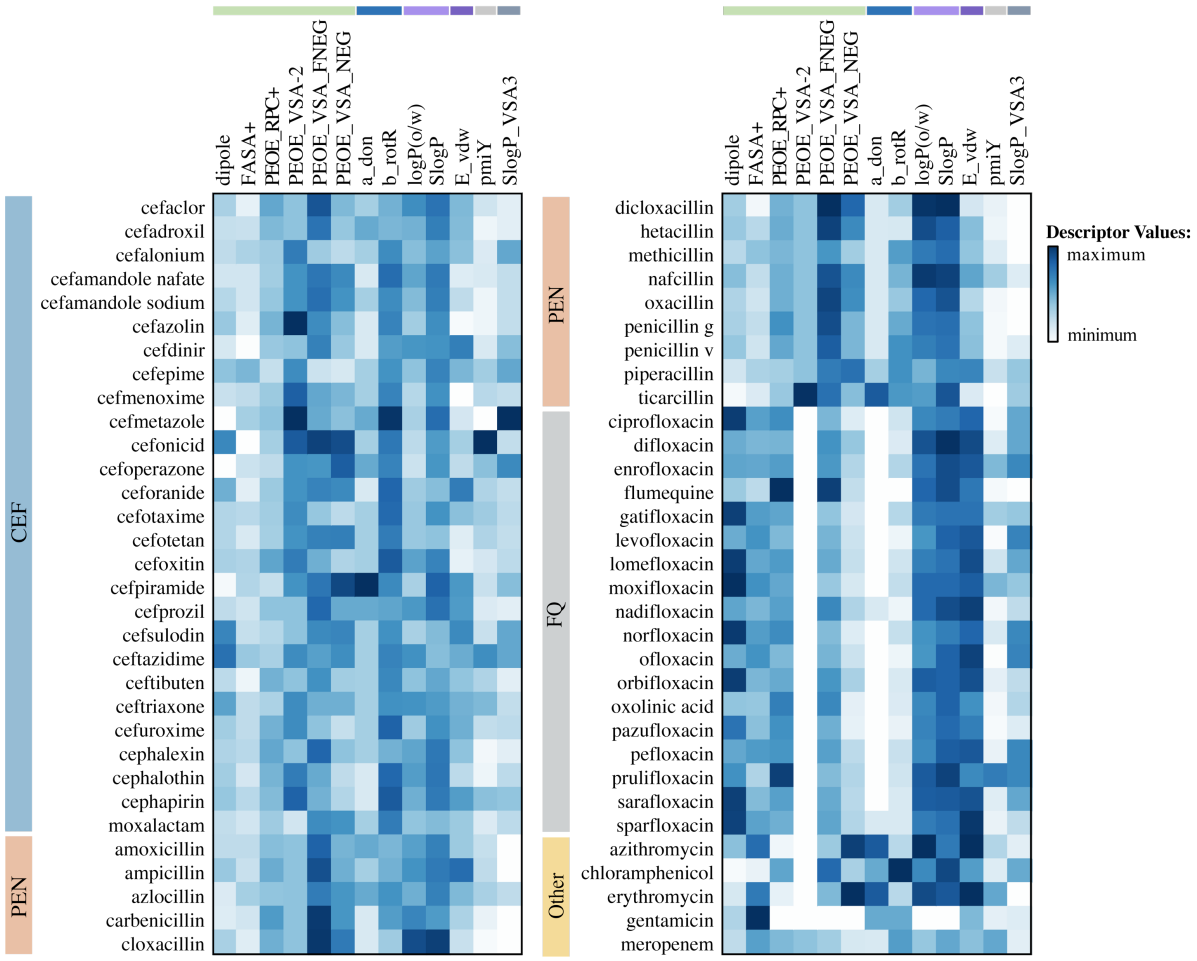


Figure 38. Selected top molecular descriptors from RF classification of antibiotics with activity against *P. aeruginosa* and *E. coli*. Descriptor values were scaled between 0 and 1 and colored blue, with darker blue indicating higher descriptor values. Minimal values for a given descriptor are shown in white.

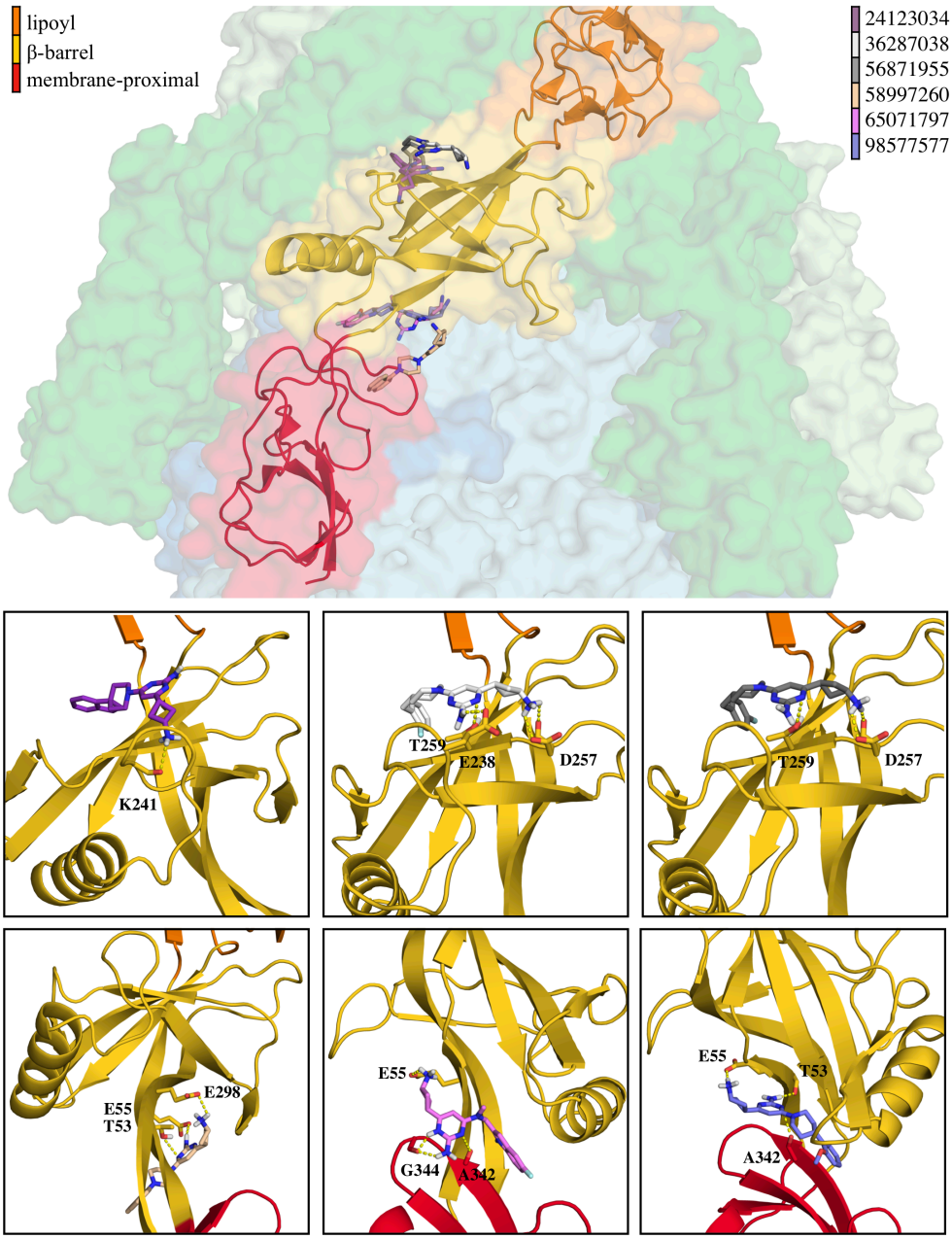


Figure 39. Highest-scoring docked pose for the six compounds that bind to AcrA and potentiate antibiotic activity (top). Individual snapshots of AcrA were aligned to chain A (colored by domain) in the cryo-EM structure of the AcrAB-TolC complex (PDB entry 5NG5), which is shown in cartoon and surface representations (green = AcrA, blue = AcrB). Individual docking poses for each of the six compounds highlighting interactions with residues in AcrA colored by domain (bottom). See ref 201 for additional experimental details.

CHAPTER V

SUBSTRATE BINDING INDUCES CONFORMATIONAL CHANGES IN A CLASS A β-LACTAMASE THAT PRIME IT FOR CATALYSIS

Text and figures are taken from the following:

Langan, P.S., Vandavasi, V.G., Cooper, S.J., Weiss, K.L., Ginell, S.L., Parks, J.M., and Coates, L. Substrate binding induces conformational changes in a class A β-lactamase that prime it for catalysis. *ACS Catal.* 2018, 8, 2428-2437.

Copyright (2020) American Chemical Society.

P.S.L., V.V.G., K.L.W., and L.C. performed experiments and crystallography; S.J.C. and J.M.P. performed QM/MM MD simulations; All authors contributed to preparing the manuscript.

Abstract

The emergence and dissemination of bacterial resistance to β-lactam antibiotics via β-lactamase enzymes is a serious problem in clinical settings, often leaving few treatment options for infections resulting from multidrug-resistant superbugs. Understanding the catalytic mechanism of β-lactamases is important for developing strategies to overcome resistance. Binding of a substrate in the active site of an enzyme can alter the conformations and pK_as of catalytic residues, thereby contributing to enzyme catalysis. Here we report X-ray and neutron crystal structures of the class A Toho-1 β-lactamase in the apo form and an X-ray structure of a Michaelis-like complex with the cephalosporin antibiotic cefotaxime in the active site. Comparison of these structures reveals that substrate binding induces a series of changes. The side chains of conserved residues important in catalysis, Lys73 and Tyr105, and the main chain of Ser130 alter their conformations, with Nζ of Lys73 moving closer to the position of the conserved catalytic nucleophile Ser70. This movement of Lys73 closer to Ser70 is consistent with proton transfer between the two residues prior to acylation. In combination with the tightly bound catalytic water molecule located between Glu166 and the position of Ser70, the enzyme is primed for catalysis when Ser70 is activated for nucleophilic attack of the β-lactam ring. Quantum mechanical/molecular mechanical (QM/MM) free energy simulations of models of the wild-type enzyme show that proton transfer from the Nζ of Lys73 to the Oε2 atom of Glu166 is more thermodynamically favorable than when it is absent. Taken together, our findings indicate that substrate binding enhances the favorability of the initial proton transfer steps that precede the formation of the acyl-enzyme intermediate.

Introduction

Since their fortuitous discovery and introduction into the clinic, β-lactam antibiotics have revolutionized medicine.[68, 69] With their ability to imitate units of the bacterial cell wall during cell wall synthesis, they inhibit cell wall regeneration during the autolysis and rebuilding process in the cell, thus causing cell death.[70, 221] However, the development of acquired and evolved resistance by bacteria is inevitable. Despite the wide variety of β-lactam antibiotics available today, there are constantly emerging threats to public health from resistant strains. Four mechanisms are used individually or in combination by resistant bacteria to overcome β-lactam antibiotics: (i) mutations in the active site of penicillin binding proteins (PBPs) that result in reduced affinity of antibiotics for PBPs, (ii) decreased expression of outer membrane proteins that give β-lactams access to the cell wall building region in the periplasm, (iii) expression of efflux pumps that expel β-lactams and other substances that are harmful to the cell, and (iv) the production of β-lactamases, which inactivate β-lactam antibiotics.[222]

β-lactamases are divided into four classes (A-D) on the basis of sequence homology.[70] Apart from the class B metalloenzymes, which require two Zn^{2+} ions in the active site to function[70, 223, 224], β-lactamases are serine-reactive hydrolases. Typical class A β-lactamases include sulfhydryl variable (SHV), Temoniera (TEM), and extended-spectrum β- lactamase (ESBL) cefotaxime-resistant (CTX) M-type enzymes. CTX-M-type β-lactamase enzymes are often encountered in bacterial intraabdominal and urinary tract infections. CTX-M ESBLs can inactivate first-, second-, and third- generation cephalosporins and monobactam antibiotics.[70-72] In combination with their broad substrate profile, these β-lactamases create challenges for clinical treatment and increase mortality rates.

Due to its potent activity against extended-spectrum cephalosporins, drugs that reach the spinal fluid in a high enough concentration to treat meningitis, Toho-1 is a class A CTX-M-type ESBL β-lactamase of particular interest.[225-227] Common to other class A β-lactamases, Toho-1 is composed of two highly conserved domains, α/β and α, the interface of which forms the active site cavity.[73] Like all class A β- lactamases, Toho-1 employs an active site serine nucleophile (Ser70) to cleave the β-lactam bond of the substrate in a two- step acylation-deacylation reaction cycle that leads to overall hydrolysis (**Figure 40** in **Appendix V**). Various detailed mechanisms have been

proposed for the formation of the acyl-enzyme intermediate during catalysis.[74-77] Differentiating between these mechanisms can be facilitated by unambiguously identifying key protonation states and hydrogen-bonding interactions of the catalytically important residues and the substrate. Neutron crystallography is ideally suited to experimentally determine protonation states. Our previous studies have shown that both Glu166 and Lys73 can undergo changes in protonation state upon binding of transition state analogues[228] and during the formation of the acyl-enzyme intermediate.[229, 230]

Quantum mechanical/molecular mechanical (QM/MM) calculations can provide key mechanistic insights that are complementary to experiments. For example, this approach allows detailed inspection of short-lived intermediates and transition states and the quantification of reaction energetics of enzymatic reactions. To obtain meaningful results, solvent effects must be properly taken into account. Long-range electrostatic effects are known to contribute significantly to the structure and properties of biomolecules.[231] Most often in QM/ MM simulations, a nonperiodic water droplet model is used, and this approach has been shown to capture long-range electrostatic interactions sufficiently well in comparison to periodic simulations with QM/MM-Ewald approaches.[232] Configurational sampling is required to provide information about free energies and can be obtained using umbrella sampling. The computational cost of performing QM/MM umbrella sampling with density functional theory (DFT) is quite high. Thus, a common approach is to perform the simulations with a computationally efficient semiempirical QM method. However, semiempirical methods are generally less accurate than DFT. Their accuracy can be improved by computing potential energies with DFT-based QM/MM calculations and then accounting for thermal and entropic effects by performing configurational sampling with semi-empirical-based QM/MM simulations.

Previous QM/MM studies of class A β-lactamases have focused on the acylation[78-81]and deacylation steps[82, 83], and some of these studies have helped establish likely mechanisms for β-lactam inactivation. However, the specific contributions of the substrate in modulating proton transfer free energies has not been investigated. Thus, in the present work, we have combined X-ray and neutron crystallography with QM/MM simulation to identify key factors that contribute to catalytic rate enhancement by a class A β-lactamase. By generating a Ser70Ala mutant of Toho-1

β-lactamase and obtaining a crystal structure with the substrate cefotaxime, we have captured the preacylation complex. This approach allowed us to scrutinize the conformations and protonation states of key active site residues as β-lactam hydrolysis is poised to occur. Our findings provide evidence in support of a concerted base hypothesis originally proposed by Mobashery and co-workers.[78] Specifically, they reveal concerted changes in the conformations of several residues upon substrate binding and the presence of a hydrogen bond network capable of facilitating cleavage of the β-lactam bond. Furthermore, our QM/MM free energy simulations show that the presence of the cefotaxime substrate alters the relative proton affinities of key catalytic residues, facilitating proton transfers prior to acylation. Recent high-resolution crystal structures of CTX-M14, another class A β-lactamase, in complex with a conjugated penicillin[233] and boronic acid inhibitors[234] have indicated that changes in the protein microenvironment upon binding of small molecules can induce protonation state changes.

Methods

For experimental details see:

Langan, P.S., Vandavasi, V.G., Cooper, S.J., Weiss, K.L., Ginell, S.L., Parks, J.M., and Coates, L. Substrate binding induces conformational changes in a class A β-lactamase that prime it for catalysis. *ACS Catal.* 2018, 8, 2428-2437.

Simulation system setup

The program antechamber from the AMBER 14 suite of programs[235] was used to assign general Amber force field[236] parameters for cefotaxime. AM1-BCC atomic partial charges[237, 238] were calculated with the program sqm.[239]

The 2.1 Å neutron structure of Arg274Asn/Arg276Asn Toho-1 β-lactamase[240] was used to generate the apoenzyme model. The 1.1 Å X-ray structure of Ser70Ala/Arg274Asn/Arg276Asn Toho-1 β-lactamase with cefotaxime determined in the present study[241] was used to generate the cefotaxime-bound model with Ala70 converted back to Ser. The ff14SB force field[242] and TIP3P water model[138] were used to describe the protein and solvent, respectively. Each system was solvated in a periodic box with a 20 Å margin between the protein and the sides of the box. The charge of the cefotaxime system was neutralized by adding a single sodium ion. No ions were

added to the apoenzyme system. Thus, although crystallization was carried out in high ionic strength buffers, all simulations were performed under low ionic strength conditions.

All protonation states and active-site hydrogen-bonding patterns in the apoenzyme and cefotaxime-bound models were assigned directly from the corresponding neutron structures. The protonation states and hydrogen-bonding patterns are consistent in both structures. Unfortunately, it was not possible to trap the substrate without mutating a catalytic residue, in this case Ser70. Although it is possible that mutating Ser70 to Ala could alter active site pK_a values, we do not expect that it would change the protonation states.

Classical MD simulations

Initial relaxation of the system consisted of 250 steps of steepest descent minimization followed by 750 steps of conjugate gradient minimization. During all MD simulations, a time step of 2 fs was used, and covalent bonds to hydrogen atoms were constrained to their equilibrium distances with the RATTLE algorithm.[243] A cutoff of 8 Å was used for van der Waals interactions, and long-range electrostatic interactions were computed with the particle mesh Ewald (PME) method.[244, 245] Each system was heated to 300 K with a Langevin thermostat over 50 ps in the canonical (NVT) ensemble with a 5.0 kcal mol-1 $Å^{-2}$ harmonic restraint applied to all heavy atoms of the protein, substrate, and the active site water molecule. Next, a 50 ps equilibration was performed in the isothermal-isobaric (NPT) ensemble to adjust the pressure of the system to 1 atm, again with the same 5 kcal mol-1 $Å^{-2}$ harmonic restraint. An additional 200 ps equilibration was then performed with a 1 kcal mol-1 $Å^{-2}$ restraint on all Cα atoms, cefotaxime heavy atoms, and the active site water molecule. Production MD simulations were then performed in the NPT ensemble for >10 ns with no restraints. All classical MD simulations were performed with pmemd.

QM model calculations

To identify a suitable level of theory with which to describe the QM subsystem in the QM/ MM calculations, we performed a set of model calculations on selected amino acid side chains. The M06-2X global hybrid density functional[246] and 6-311++G(2d,2p) basis set were previously shown to provide an accurate description of proton transfers between amino acids, yielding a mean unsigned error of 1.0 kcal mol^{-1} and maximum error of 1.4 kcal mol^{-1} for proton affinities at 0 K in comparison to benchmark CCSD(T)/CBS values (**Table 12** in **Appendix V**).[247] In that work,

geometries were optimized at the CPCM/MP2/6-311++G(d,p) level of theory. We followed the same procedure but reoptimized the geometries of the Asp, Ser, and Lys side chains at the CPCM/ M06-2X/6-31G(d) level of theory and then calculated gas-phase single-point energies with the larger 6-311++G(2d,2p) basis set. The resulting errors relative to the benchmark values were 0.2, 1.4, and 1.3 kcal mol^{-1} for Asp, Ser, and Lys, respectively, indicating that M06-2X/6-31G(d) geometries and M06-2X/6-311++G(2d,2p) single-point energies provide quite accurate proton affinities and are expected to be useful for quantifying proton transfer energetics in the present enzyme system.

QM/MM system preparation

A spherical water droplet model was used for all QM/MM simulations. For the apo and cefotaxime-enzyme systems, a single snapshot was chosen near end of each classical MD simulation trajectory, and all water molecules greater than 36 Å from Cα of Ser70 were removed. The entire protein was enclosed within the spherical solvent shell. Prior to performing QM/MM calculations, the energy of the system was minimized for 20000 steps.

For both the apoenzyme and cefotaxime-enzyme model, the side chains of Ser70, Lys73, Glu166, and one active site water molecule were included in the QM subsystem. Hydrogen link atoms were used to saturate the valences of the covalent bonds at the QM/MM boundary. The substrate was included in the MM region for the cefotaxime-bound model. The total charge of the QM subsystem was zero for both models. No cutoffs were used for the nonbonded interactions in the QM/MM calculations.

Potential energy profiles

We performed a series of restrained geometry optimizations using a generalized reaction coordinate approach.[248] Geometries were considered converged when the RMS gradient dropped below 10^{-2} kcal mol^{-1} $Å^{-1}$. The reaction coordinate corresponded to the Lys73 catalytic base mechanism proposed by Mobashery and co-workers.[78] They also considered an alternative pathway in which Glu166 serves as the catalytic base that deprotonates Ser70 through a water molecule. However, this pathway was found to have a 4 kcal mol^{-1} higher energy barrier that the Lys73 general base pathway. Thus, we simply chose to focus on the Lys73 pathway in our work. The same atoms were used to define the reaction coordinate for both the apoenzyme and

cefotaxime-enzyme models (**Figure 41**).

The restraint energy, U, was defined as

$U = k(w_1|\boldsymbol{d_1}| + \ldots + w_n|\boldsymbol{d_n}|-\boldsymbol{d_0})^2$ (Eqn. 2)

where k is the force constant in kcal mol^{-1} Å^{-2}, w_n are the restraint weights, d_n are the interatomic distances, and d_0 is a distance offset parameter that is adjusted at each step along the reaction path. A force constant of 2000 kcal mol^{-1} Å^{-2} was used, and the reaction coordinate was adjusted in increments of 0.1 Å. We modified the program sander to allow up to six interatomic distances to be included in the generalized reaction coordinate. All atoms greater than 24 Å from Cα of Ser70 were restrained during the potential energy profile calculations. All restrained geometry optimizations were performed sequentially in both the forward and reverse directions until smooth, continuous paths were obtained.

The following interatomic distances and weights were used:

d_1 = Glu166(Oε)−water(H), w_1 = -1.0

d_2 = water(H)−water(O), w_2 = 1.0

d_3 = water(O)−Ser70(Hγ), w_3 = -1.0

d_4 = Ser70(Hγ)−Ser70(Oγ), w_4 = 1.0

d_5 = Ser70(Oγ)−Lys73(Hζ), w_5 = -1.0

d_6 = Lys73(Hζ)−Lys73(Nζ), w_6 = 1.0

QM/MM free energies

Performing Born-Oppenheimer MD simulations with a DFT description of the QM subsystem requires considerable computational resources. However, analogous simulations with a semiempirical Hamiltonian are much less computationally demanding and can provide reasonably accurate free energies in many cases. The PM6 Hamiltonian was developed to provide accurate geometries and energies of many types of molecules.[249] Although proton affinities computed with PM6 exhibit errors on the order of 2-4 kcal mol^{-1} for the side chains of Lys, Ser, and Glu,[250] we expect that performing umbrella sampling along a predefined reaction coordinate should provide meaningful estimates of thermal and entropic contributions. Thus, to combine the accuracy of DFT for the underlying energetics with the computational efficiency of PM6 for sampling, we first

computed DFT/MM potential energies and then added a free energy correction computed with PM6/MM at each reaction coordinate value:

$$G_i = E_i\ (M062X/MM) + G_i^{corr}\ (PM6/MM)$$

where

$$G_i^{corr}(PM6/MM) = G_i^{US}\ (PM6/MM) - E_i^{PE}\ (PM6/MM)$$

For the umbrella sampling simulations, PM6/MM geometries obtained from restrained optimizations were used as initial structures for umbrella sampling simulations, which were performed at the same level of theory. We used the same reaction coordinate as for the restrained optimizations, but in this case the force constants for the harmonic restraints were set to 300 kcal mol^{-1} $Å^{-2}$ for each simulation window. Atoms greater than 24 Å from Cα of Ser70 were restrained with a force constant of 5 kcal mol^{-1} $Å^{-2}$. The time step was set to 1 fs, and the temperature was maintained at 298 K with a Langevin thermostat and a collision frequency of 10 ps^{-1}. After 5 ps MD equilibration, we performed 25 ps production sampling at 298 K for each replica. The potential of mean force (PMF) for each system was obtained with the weighted histogram analysis method (WHAM)[251] implemented in the program WHAM. Statistical errors were estimated using bootstrapping with 100 samples. All classical MD and QM/MM simulations were performed with AMBER 14.[235] For the QM/MM calculations with DFT, the AMBER 14/Gaussian 09[135] interface was used.

Results and Discussion

Substrate-free structures

Previously, we have used neutron and X-ray crystallography to determine the side- chain orientations, water positions, protonation states, and hydrogen-bonding networks in the active sites of some of the reaction states of Toho-1 with both substrate-free and substrate-bound enzyme.[240, 252] In comparison to previously determined structures of the wild-type apoenzyme, the X-ray and neutron structures in this study show no perturbation to the active site as a result of the Ser70Ala mutation. In the X-ray and neutron structures of the apoenzyme, Lys73 appears in a single conformation oriented toward Ser130 and forming hydrogen bonds with the carbonyl group of Ser130 (2.79 Å, ESD 0.02 Å), Oγ of Ser130 (2.95 Å, ESD 0.02 Å), and OD1 of Asn132 (2.66 Å, ESD 0.02 Å) (**Figure 42**). The omit nuclear density maps in the neutron structure show a definitively protonated Lys73 in the ND_3^+ form in this conformation, indicating that this residue

is the donor group in all three hydrogen bonds. However, we note that the absence of the Ser70 side chain in the Ser70Ala mutant could potentially alter the pK_a of Lys73 and thus its protonation state relative to the wild-type enzyme. In addition, the pH of crystallization was slightly acidic (6.1), which may also have affected the protonation state of Lys73.

Enzyme-substrate complex

The enzyme-substrate complex is identical to the substrate-free complex with the following exceptions. In the X-ray structure of the enzyme-substrate complex, cefotaxime has a refined occupancy of 66% and Lys73 is present in two conformations, one of which is identical to its conformation in the substrate-free enzyme (**Figure 42**). The refined occupancy for the Lys73 conformation corresponding to that found in the substrate-free complex (A conformation) is 36%, whereas the refined occupancy for the second Lys73 conformation (B conformation) is 64%. In the A conformation, Lys73 forms hydrogen-bonding interactions with the carbonyl of Ser130 (2.74 Å, ESD 0.05 Å), Oγ of Ser130 (2.95 Å, ESD 0.07 Å), and Oδ1 of Asn132 (2.74 Å, ESD 0.05 Å) and is presumed to be fully protonated, as in the neutron structure of the substrate-free enzyme.

In contrast, in the B conformation Lys73 interacts with the catalytic water molecule (2.89 Å, ESD 0.03 Å), Oδ1 of Asn132 (2.76 Å, ESD 0.03 Å), and Oγ of Ser130 (3.06 Å, ESD 0.06 Å). The catalytic water molecule forms hydrogen bonds with Lys73 (B) (2.89 Å, ESD 0.03 Å), Oδ1 of Asn170 (2.81 Å, ESD 0.03 Å), and Oε2 of Glu166 (2.56 Å, ESD 0.03 Å) (**Figure 42**). The catalytic water is therefore likely to be coordinated such a way that it accepts hydrogen in a hydrogen bond with Lys73 and it donates hydrogen in a hydrogen bond with Oε2 of Glu166, thus providing a possible proton transfer pathway between Lys73 and Glu166. The B factors are particularly low for all atoms in the side chain of Lys73 in both X-ray structures (all are less than 10 Å^2), enabling these discrete conformations to be visualized.

Protonation states of the key catalytic residues Lys73 and Glu166 in the precovalent state of class A β-lactamases have been debated for decades. Lys73 has been proposed to be either cationic ($-NH_3^+$) or neutral ($-NH_2$)[76], which lead to different mechanisms for the formation of the acyl-enzyme intermediate (**Figure 40**). The results from this study allow us to resolve between these alternatives.

In the unprotonated form, Lys73 would act as the general base to activate Ser70 directly for the attack on the β-lactam ring of the substrate. A QM/MM study of TEM-1 β-lactamase with penicillanic acid substrate investigated the identity of the general base in the acylation step.[78] That study provided evidence for a concerted base mechanism in which a proton is transferred from Lys73 to Glu166 through an active site water molecule and Ser70. The authors proposed that substrate binding alters the pK_as of Lys73 and Glu166. The proton transfer results in a Michaelis complex in which all key residues, Ser70, Lys73, and Glu166, are neutral. A neutral Lys73 would then deprotonate Ser70 to activate it for nucleophilic attack on the β-lactam ring. Experimental verification of these results has been challenging due to the transient nature of the Michaelis complex and the high reactivity of β-lactamases. Substrate binding can alter the protein microenvironment and thus the pK_as of catalytic residues. In β-lactamases, these effects could potentially help drive proton transfer in the precovalent Michaelis complex (**Figure 40**), facilitating the acylation step of the reaction.

As the separately refined occupancies of the cefotaxime substrate (66%) and the B conformation of Lys73 (64%) are nearly equivalent, it is highly likely that the conformation of Lys73 is directly altered by the binding of cefotaxime in the active site cavity. Superimposing the substrate-free structure of Toho-1 β-lactamase Arg274Asn/Arg276Asn, which possesses Ser70, with the cefotaxime- bound structure described in the present work yields an RMSD for main chain atoms of 0.28 Å, indicating that the two structures are nearly identical. In the cefotaxime-bound structure, Nζ of Lys73 in the B conformation lies just 2.39 Å away from Oγ of Ser70 in the 2ZQ8 structure, whereas this distance increases to 3.30 Å in the A conformation. Thus, upon binding of cefotaxime, Lys73 moves closer to Ser70 to enable proton transfer from Lys73 to Ser70.

The catalytic water molecule in the cefotaxime-bound structure of the Ser70Ala mutant is located 2.56 Å away from Oε2 of Glu166 and 2.36 Å from Oγ of Ser70 (PDB ID 2ZQ8). This orientation provides a clear pathway for a proton to be transferred from Ser70 to a negatively charged Glu166 via the catalytic water. The B conformation of Lys73 is also ideally positioned to transfer a proton to Ser70 simultaneously with the protonation of Glu166, which would result in the formation of a precovalent complex in which Ser70, Lys73, and Glu166 are all neutral. The protonation state of

Glu166 in the 15 K X-ray substrate-free crystal structure was deduced by comparing the difference in carboxyl bond lengths with their estimated standard deviations (ESDs). The unrestrained carboxyl bond lengths for Glu166 refined to essentially equal bond lengths, i.e., 1.26 Å (ESD = 0.02 Å) for the Cδ−Oδ1 bond and 1.25 Å (ESD = 0.02 Å) for the Cδ−Oδ2 bond, indicating that an anionic Glu166 is poised to accept a proton from the catalytic water.

To determine whether the presence of the substrate alters proton transfer energetics prior to acylation, we performed QM/MM free energy simulations with models of the substrate- free and cefotaxime-bound enzyme and compared the free energy profiles for the proton transfers in each system. We generated a model of the cefotaxime-bound structure by replacing Ala70 in the Ser70Ala mutant with the native Ser. For the substrate-free model, we used the neutron structure of substrate-free Toho-1 β-lactamase published previously.[240,45] It has been shown that proton transfer from Lys73 to Glu166 via Ser70 and an active site water molecule is energetically favorable for TEM-1 β-lactamase with penicillanic acid as the substrate.[78,38] Therefore, this reaction pathway was considered in the present simulations.

The QM/MM optimizations indicate that the proton transfers are concerted and synchronous in both the apoenzyme and the cefotaxime-bound system (**Figure 43**), as evidenced by a single free energy barrier in each case (**Figure 44**). For the substrate-free model the estimated proton transfer free energy barrier is 5 kcal mol^{-1}, with the product state (i.e., neutral Lys73 and neutral Glu166) being 2.5 kcal mol^{-1} higher in energy than the reactant state (i.e., cationic Lys73 and anionic Glu166). These relative free energies are consistent with the neutron crystal structure of the apoenzyme, in which Lys73 is cationic and Glu166 is anionic. For the cefotaxime- bound model we obtained a computed proton transfer barrier of 2.8 kcal mol^{-1} and a reaction free energy of -6.2 kcal mol^{-1}. Thus, the presence of the cefotaxime substrate clearly alters the relative pK_as of Lys73 and Glu166 to facilitate proton transfer from Lys73 to Glu166, as proposed previously.[78] Of course, acylation cannot occur in the absence of substrate, but this comparison enables quantification of the role of the substrate toward catalyzing its own hydrolysis.

In the substrate-free structure, Tyr105 also appears in a single conformation in which its phenolic hydroxy group forms a hydrogen bond with a water molecule (2.60 Å) that also forms a hydrogen bond with the main chain carbonyl of Tyr129 (2.76 Å). In the substrate-bound structure with

cefotaxime, Tyr105 is present in two conformations. In the first conformation (A), which has a separately refined occupancy of 34%, the hydroxy group of Tyr105 interacts again with a water molecule (2.61 Å) that also forms a hydrogen bond with an adjacent water molecule (2.80 Å). However, in the second conformation (B), which has a refined occupancy of 66%, the hydroxy group of Tyr105 forms a direct hydrogen bond with the main-chain carbonyl of Tyr129 (2.57 Å). In this conformation, Tyr105 forms several close contacts with the main-chain atoms of Tyr129 and Ser130, which in turn induce a slight change in the conformation of the main chain around the carbonyl group of Ser130 that helps trigger the formation of the B conformation of Lys73. The B factors are low for most of the atoms in the side chain of Tyr105 in the cefotaxime-bound X-ray structure (~12 $Å^2$ on average), which enables these discrete conformations to be visualized. The movement of Tyr105 is likely to be driven by the binding of the cefotaxime substrate, which has an identical separately refined occupancy value of 66%, due to multiple hydrophobic interactions between the phenol ring of Tyr105 and the dihydrothiazine ring of cefotaxime. This finding agrees with several earlier studies which have shown that, while not a catalytic residue, Tyr105 is important for catalytic efficiency.[252]

Conclusions

By using a combination of neutron and X-ray crystallography, we have determined the protonation states of Lys73 and Glu166 in the active site of the precovalent complex. Using occupancy refinement, we have observed how the binding of the cefotaxime substrate initiates several conformational changes from Tyr105 to Ser130 and then Lys73. This conformational change in the side chain of Lys73, which is directly induced by cefotaxime binding, places the enzyme into a catalytically competent state. Our findings are consistent with the concerted base hypothesis originally proposed by Mobashery and co-workers.[78] They proposed that substrate binding triggers proton transfer from Lys73 to Glu166 through Ser70 and an active site water molecule. The resulting neutral Lys73 then deprotonates Ser70 to facilitate nucleophilic attack on the β-lactam ring. We performed QM/MM free energy simulations of the initial proton transfer steps to quantify the role of the substrate cefotaxime in facilitating the reaction. We found that proton transfer from Lys73 to Glu166 is more thermodynamically favorable when the substrate is bound. The present study fills in considerable detail on the structure and energetics of the proton relay network in the active site of a class A β-lactamase in the precovalent complex. Our structures also reveal concerted changes in the conformations of several residues upon substrate binding and the

presence of a hydrogen bond network capable of facilitating cleavage of the β-lactam bond. Although we did not directly observe catalysis in action, these three crystal structures two X-ray and one neutron supported by extensive QM/MM calculations, provide a compelling case for the most likely catalytic route.

Appendix V

Tables

Table 12. Proton affinities[a] (kcal mol^{-1}) for selected amino acid side chains calculated at the M06-2X/6-311++G(2d,2p)//CPCM/M06-2X/6-31G(d) level of theory[b] and benchmark CCSD(T)/CBS//CPCM/MP2/6-311+G(d,p) values.

Amino acid	M06-2X	CCSD(T)	Error
Asp^-	349.61	349.80	0.20
Ser^-	376.62	377.90	1.40
Lys	226.68	228.03	1.30
MUE			1.00

[a] zero-point-exclusive proton affinity at 0 K

[b] a single dihedral angle was constrained during geometry optimization in some cases to prevent intramolecular hydrogen bonding

[c] Error = CCSD(T)–DFT

Figures

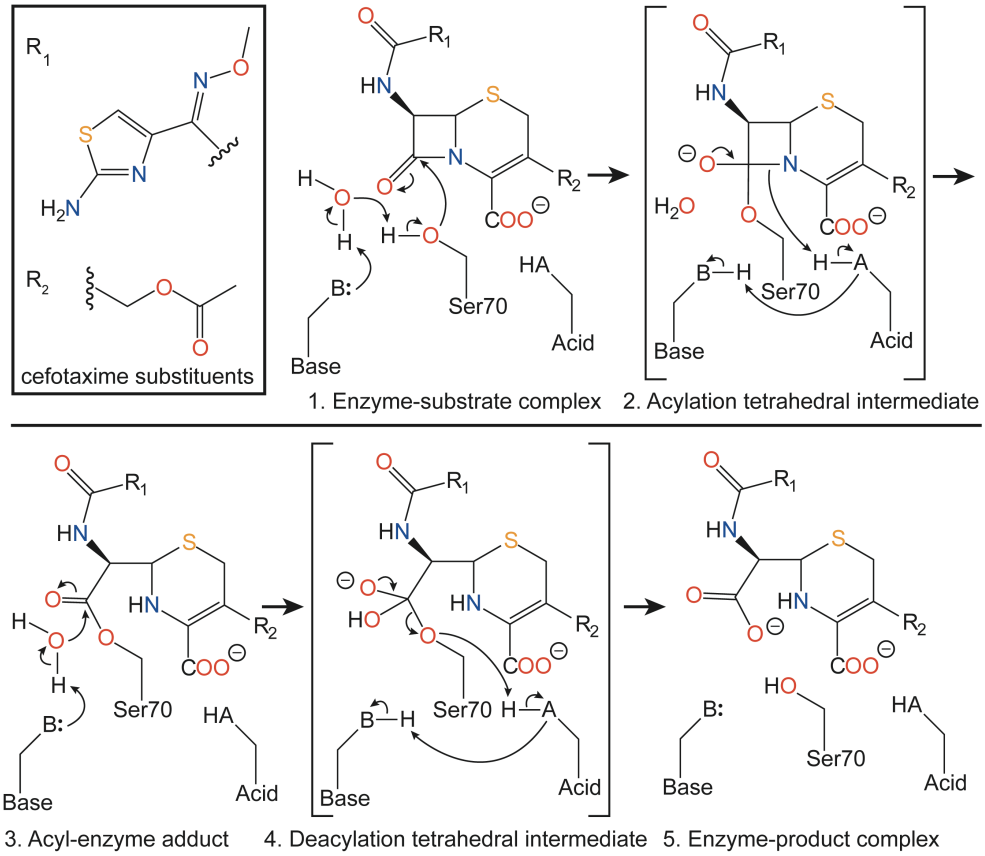


Figure 40. Catalytic mechanism of class A β-lactamase inactivation of a β-lactam substrate. A serine nucleophile cleaves the β-lactam bond of the substrate in two steps, acylation and deacylation, which lead to hydrolysis: First, the pre-covalent enzyme-substrate complex is formed and the acylation reaction is initiated (1). General base-catalyzed nucleophilic attack on the β-lactam carbonyl by the serine hydroxy group proceeds through a tetrahedral intermediate (2) and forms a transient acyl-enzyme adduct (3). The acyl-enzyme adduct (3) undergoes general base-catalyzed attack by the hydrolytic water molecule and forms a second tetrahedral intermediate during deacylation (4), which subsequently collapses to form a post-covalent complex (5) prior to release of the hydrolyzed product.

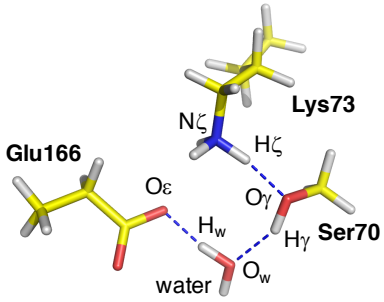


Figure 41. Atoms used to define the generalized reaction coordinate in the QM/MM simulations.

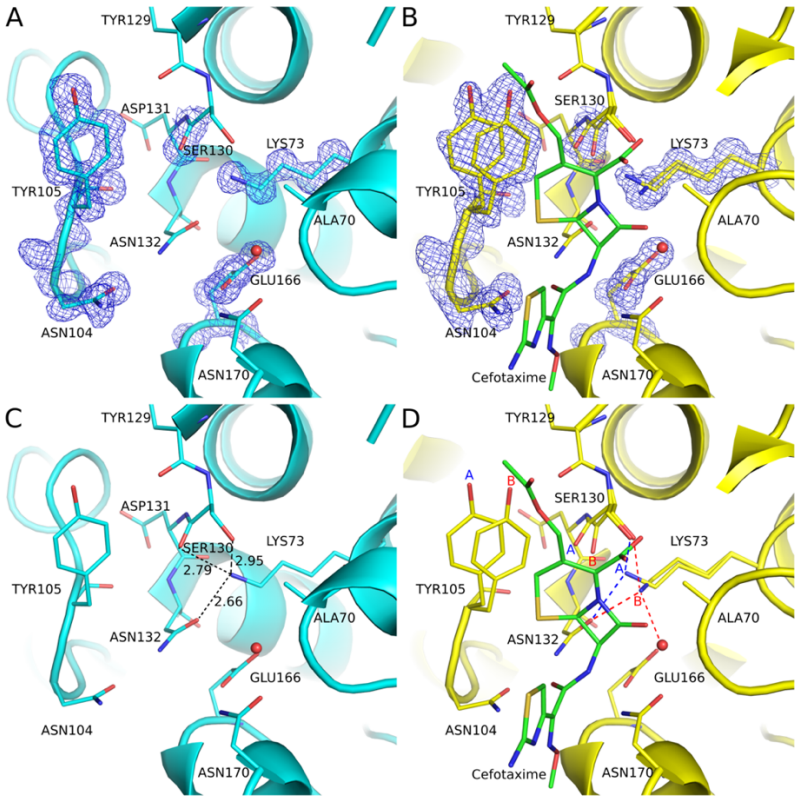


Figure 42**.** Structures of substrate-free and cefotaxime-bound active sites of the Ser70Ala Toho-1 mutant. Electron density $2F_o$-F_c maps are represented at a σ level of 1.0. (**A, C**) cefotaxime-free structure shown in cyan. Nitrogen atoms are shown in blue, carbon in cyan, sulfur in gold, and oxygen in red with the catalytic water molecule shown as a red sphere. (**C**) Distances (in Å) between Lys73 and close contacts are labeled. (**B, D**) The cefotaxime-bound structure is shown in yellow. Cefotaxime carbons are shown in green and protein carbons are shown as yellow sticks. (**D**) Both conformations are shown for Tyr105, Ser130, and Lys73. Conformations A and B are shown with the B conformations accommodating cefotaxime within the active site. For clarity, electron density is shown only for certain active sites residues. See ref 241 for details on experimental methods.

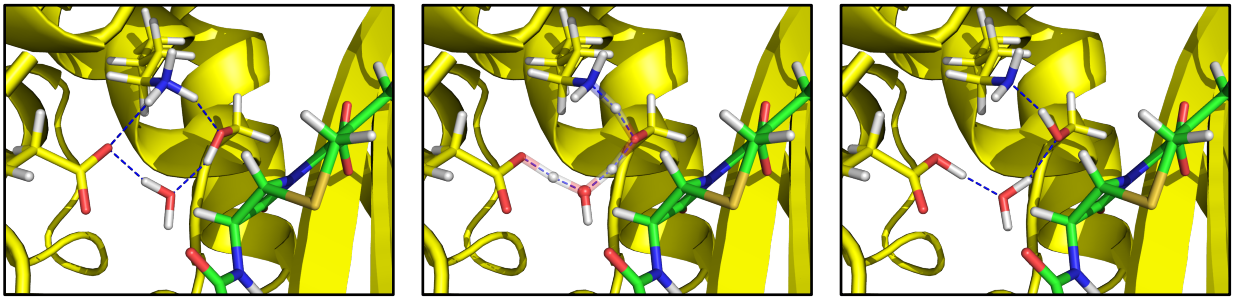


Figure 43**.** Reaction path for QM/MM simulations. Reactant (*left*), transition state (*middle*) and product (*right*) states for the transfer of a proton from Lys73 to Glu166 through Ser70 and the active-site water molecule in the cefotaxime-bound enzyme.

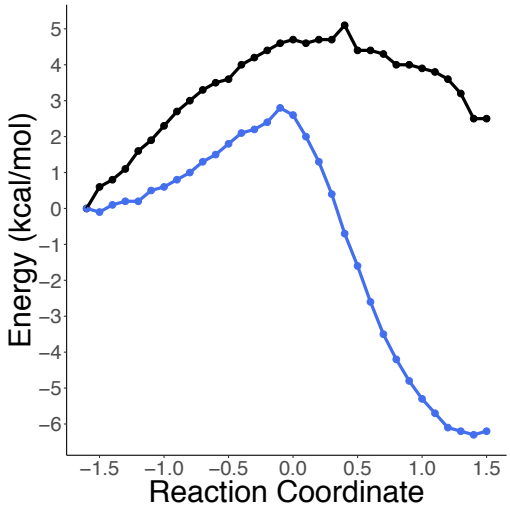


Figure 44. Computed free energy profiles for proton transfer from Lys73 to Glu166 through Ser70 and the active-site water molecule in the substrate-free (black) and cefotaxime-bound enzyme (blue).

CONCLUSIONS

Overview

The 3D structure of a protein can be fundamentally useful to understanding its function. The most common way to obtain a structure of a protein in the absence of an X-ray, cryo-EM, or NMR structure is to use homology modeling, or the mapping of the target sequence onto a closely related homolog with an available structure. However, despite recent advances and efforts in structural biology, the 3D structure of many protein families remains unknown. Recent advances in genomic and metagenomic sequencing combined with coevolution analysis and protein structure prediction have allowed for highly accurate structural modeling of proteins previously considered intractable to model due to the lack of suitable templates.[6, 22] Models generated by any or all of these approaches can then be further studied with other computational tools such as molecular dynamics (MD) simulations, docking of small molecule ligands to identify substrates, machine learning (ML) and quantum mechanical/molecular mechanical (QM/MM) free energy simulations to study reaction mechanisms. Here these various computational approaches were combined with predicted or experimentally determined structures to better understand the structure, dynamics, functions, and mechanisms of various bacterial proteins (**Figure 1**).

Structural modeling of the HgcAB complex provides insights into the mechanism of bacterial mercury methylation

By using metagenome-based protein structure calculations to generate models of the individual domains of HgcA and of HgcB valuable insights were obtained into the biochemical mechanism of Hg methylation in anaerobic microorganisms (Chapter I). HgcA is predicted to consist of two domains, the cobalamin binding domain (CBD) and a transmembrane domain (TMD). HgcB is predicted to have a dicluster ferredoxin fold. UV-visible spectroscopy of HgcA and HgcB heterologously expressed in *E. coli* confirmed that that these proteins do in fact bind vitamin B_{12} corrinoid and iron-sulfur cofactors, respectively (**Figure 2**), as predicted from previous bioinformatics analyses. These cofactors were then incorporated into the structural models and used coevolution-restraints to predict how these domains assemble to form the HgcAB complex from *D. desulfuricans* ND132 (**Figure 8**).

Based on coevolution analysis, the two domains of HgcA were not predicted to interact with each other, but rather both interact with HgcB (**Figure 3**). In addition, the TMD of HgcA did not have any detectable sequence homology with available crystal structures (**Table 1**). Furthermore, the interdomain contacts are not all fully satisfied, suggesting that domain motion may occur (**Figure 7**). The [4Fe-4S] clusters are located far away (~15 Å) from the Co center of the corrinoid cofactor. Therefore, the CBD of HgcA may move closer to the [4Fe-4S] clusters for efficient electron transfer. The corrinoid/iron-sulfur protein (CFeSP), the closest known homolog of the CBD, has been shown to undergo large-scale conformational rearrangements. These rearrangements were observed in crystal structures upon binding to a $(\beta/\alpha)_8$ triosephosphate isomerase (TIM) barrel protein that acts as a methyltransferase.[115] In the model of the HgcAB complex the CBD is oriented towards the membrane surface and would need to rearrange to accommodate a TIM barrel protein as a methyl donor. HgcB includes a pair of cysteine residues (Cys94 and Cys95) located at its C-terminus. Pairs of cysteines are commonly observed in proteins and enzymes involved in metal trafficking and detoxification. For example, the mercuric reductase catalyzes the reduction of Hg^{II} to Hg^0. This protein contains two Cys residues at its C-terminus that acquire Hg^{II} for transfer to another the pair of Cys residues in the active site active site.

The mechanism of MeHg formation by HgcAB has been proposed to involve reduction of the corrinoid cofactor by HgcB, methylation of the Co center, and methyl transfer to a Hg substrate. Based on insights from our model, we propose that Cys94 and Cys95 from HgcB acquire Hg^{II} and deliver it to the corrinoid cofactor for methylation (**Figure 12**). Assuming that the reaction proceeds through radical ligand exchange, formation of a crosslinked HgcB-Cys94/95(Sγ)–Co^{III}–(Sγ)Cys93-HgcA intermediate would occur. Both thiolate ligands would then be released upon reduction of the Co center and either of the C-terminal cysteines (Cys94/95) would be able to deliver $[CH_3Hg]^+$ to Cys73 from HgcB. To complete the reaction cycle an exogenous tholate (i.e. cysteine residue on a protein) would then liberate $[CH_3Hg]^+$ from HgcB.

These mechanistic insights obtained from our structural model are consistent with known experimental data and will facilitate the development of hypotheses that address more detailed structural and functional questions which can then be tested experimentally. For example, polar residues identified in the model to be located at the interface of individual domains of HgcA and

HgcB could be mutated to determine if in vivo mercury methylation activity is reduced. When an in vitro methylation system becomes available, similar experiments could be performed to measure the effects of mutation on interprotein binding, electron transfer, or methylation kinetics. In addition, this work demonstrates how coevolution-based analysis can be used to predict the structures of protein-protein complexes. Similar approaches could be applied to identify additional binding partners of HgcA and HgcB (e.g., the electron donor, methyl donor, other membrane-associated proteins). We hypothesize that an efflux pump may associate with HgcAB to enable rapid export of methylmercury. In the absence of sufficient sequences for coevolution analysis as described here to predict binding partners, deep-learning based approaches could be used for contact prediction (i.e. RaptorX-contact[253, 254]) and structural modeling (i.e. AlphaFold[255] or DMPFold[256]), as these methods have been shown to produce accurate structures from multiple sequence alignments with fewer sequences (~100) than with conventional coevolution-based approaches.

Subtle changes in the dynamics of the D243G mutant of IMPDH relieves inhibition and maintains catalysis

Introduction of the *hca* pathway in *E. coli* resulted in only limited growth with coumarate because accumulation of 4-hydroxybenzaldehyde inhibited the native inosine monophosphate dehydrogenase (IMPDH). Engineered pathways can put a substantial burden on the host (i.e. inhibiting an enzyme involved in purine biosynthesis). Directed evolution can be used to select for mutations that alleviate deleterious interactions between engineered metabolic pathways. Here, a series of single point mutations in IMPDH were identified that were able to relieve inhibition by 4-hydroxybenzaldehyde (Chapter II). Biochemical assays also confirmed that inhibition is relieved in the D243G mutant. This mutation is located at the N-terminal end of β11, which is near the NAD^+ binding site where the inhibitor was predicted to bind based on docking to a homology model of IMPDH in the open conformation (**Figure 18** and **Table 8**). Surprisingly, this mutation is located ~20 Å from the active site and did not appear to affect catalysis.

To investigate how a single point mutation to an essential host protein can affect inhibitor binding at the distant active site while also maintaining catalysis we built a homology model of *E. coli* IMPDH in the closed conformation and ran MD simulations of the wild-type and D243G mutant.

In the wild-type simulations, the side chain of D243 forms stable hydrogen bonds with residues K87, R219, and V220 (**Figure 19**). In the mutant, G243 can no longer form these hydrogen bonds and instead interacts with Q272. Meanwhile, the catalytic dyad located on the opposite side of the mutation containing β-barrel remained poised for catalysis by showing only minor perturbations in dynamics (**Figure 23**). Thus, it was still not clear how the mutation propagates changes to the active site.

To further identify changes in protein dynamics resulting from the D243G mutation, root-mean-squared fluctuations (RMSFs) were calculated in both the wild-type and mutant (**Figure 20**). In both systems high RMSFs were observed over the entire region containing the catalytic flap. Upon inspection of specific interactions of flap residues, the mutation was found to lead to changes in hydrogen-bonding interactions with various nearby residues that results in reorientation of a loop located on the flap that could be responsible for altering inhibitor binding (**Figure 21** and **Figure 22**). In addition, helices α2 and α8 were found to fluctuate more in the mutant than in the wild-type (**Figure 20**). Helix α8 is downstream of the mutation. Therefore, the increased fluctuations in this region are likely due to the loss of the hydrogen bonding network D243 can form. In addition, N-terminal end of helix α2 and the C-terminal end of helix α8 are located near the NAD^+ binding site. Hydrogen bonding interactions were observed between 4-hydroxybenzaldehyde and residues D248 and S250 located on a β-sheet just downstream of the mutation site and in close proximity to helices α2 and α8 (**Figure 24**). Taken together, changes in the structure and dynamics of these regions likely disrupts inhibitor access and binding to the NAD^+ binding site. Follow-up MD simulations could be performed in the presence of the inhibitor, which would require straightforward force field parameterization of 4-hydroxybenzaldehyde. Additional computational studies could also use MD simulations to investigate the changes in dynamics of other point mutations that relieved inhibition and similar kinetics experiments could be used to measure enzyme activity in these mutants.

Structure-based prediction of enzyme substrate scope in bacterial nitrilases

Protein structures provide insight into substrate scope, or the repertoire of substrates for a given enzyme. 3D structures also provide information about the overall fold of the protein as well as domain architecture and special arrangement of residues that can be useful in determining

biochemical pathways and providing clues about enzyme function. Here ML was used to predict substrate scope for a series of bacterial nitrilases by combining structural modeling, docking, and physicochemical property calculations with experimental in vitro enzyme assays (**Figure 32**). Using different machine learning models our approach obtained accurate predictions of substrate scope for a series of aliphatic, aromatic, and arylaliphatic nitriles by including descriptors for the enzymes, substrates and their interactions in the models (Chapter III).

Given a phylogenetic tree and sparse activity data (**Figure 27** and **Figure 29**), it may be difficult to identify trends in substrate scope. In some cases, highly identical sequences can have similar substrate scopes (i.e., 1A1 and 1A2). However, sequence similarity is not always a good indicator of overlap in substrate scope, as seen in the markedly different activities of PMI28 and 1A8 (88% sequence identity), as subtle changes in the amino acid composition of the active site may lead to substantial differences in activity (**Figure 26** and **Figure 29**). In contrast, distantly related sequences can have overlapping substrate scopes (i.e., 1A17 and 3WUY). Therefore, the substrate scope of an enzyme often cannot be accurately predicted based on inferences from phylogenic analysis alone.

To obtain more accurate predictions of substrate scope several machine learning models were generated (random forests (RF), support vector machines (SVM), gradient-boosted decision trees, and logistic regression). RF models performed as well as, or in some cases better than, the other three ML methods (**Figure 31**). Unlike kernel-based methods (i.e., SVM), decision tree-based methods (i.e., RF) allow for calculation of variable importance of each descriptor. In the RF model the top descriptors for accurately predicting substrate scope were often those that encode information from the structural models and docked poses (**Figure 33**).

As expected, small changes in sequence can cause large changes in specificity that would not be identified based on a phylogenetic analysis of the full sequence. In principle, our approach can capture these subtle effects if they lead to substantial changes in active site properties. Nitrilases are only one example where structure-based enzyme substrate scope prediction was applied. To adapt this approach to other types of enzymes the types of features used for machine learning would need to be carefully considered. Thus, this structure-based approach to predicting enzyme

substrate scope was designed to be highly modular. More specifically, physicochemical properties could be tailored to the enzymes of interest or calculated using different methods. For example, the types of properties that were used to describe the C≡N bond in nitriles are highly specific to this system and may not be widely applicable to other cases. In addition, QM geometries and properties, structural modeling, and ligand docking could be performed with other software packages and physicochemical properties/molecular fingerprints can be calculated with freely available software such as rdkit.[257]

To apply this approach to other types of enzymes, consideration also needs to be taken with respect to the type of machine learning methods used. The experimental assay used for nitrilase activity is semi-quantitative and so binary classification was used with a cutoff of 2 to allow for residual fluorescence in the assay results. For experimental data that is more quantitative (i.e., experimental binding affinities), regression may be a more suitable approach. It is also often beneficial to test multiple machine learning methods and evaluate their performance. Our system consisted of only 12 proteins and 20 substrates, resulting in 240 possible pairs to be used for machine learning. This dataset, while small, was large enough to produce accurate binary classification models using logistic regression, support vector machines, and two tree-based methods, random forest and gradient boosted decision trees. All four methods generally had similar performance and there are a variety of other machine learning algorithms available. However, using deep learning-based methods would require much larger data sets for training.

Overcoming antibiotic resistance

Bacteria are developing various resistance mechanisms to antibiotic treatments at an alarming rate and resistance has been detected against every approved antibiotic on the market. Antibiotic resistance can develop through (i) target modification, preventing drugs from binding and exhibiting activity, (ii) enzymatic inactivation, leading to degradation products that lack antibiotic activity, (iii) altering membrane composition to prevent antibiotics from entering bacterial cells, and (iv) using active efflux pumps that expel antibiotics out of cells. Understanding these various mechanisms of antibiotic resistance is important for developing strategies to overcome multi-drug resistance and restore antibiotic effectiveness.

Overcoming antibiotic resistance to Gram-negative bacteria is challenging due to synergistic interactions between two barriers: the low permeability of the outer membrane (OM) and active multidrug efflux pumps. These barriers can be separated by using strains of Gram-negative bacteria that have hyperporinated OMs, lack efflux pumps, or both. Using these various strains of *E. coli* and *P. aeruginosa* with controllable permeability and efflux barriers the activities of β-lactam and fluoroquinolone antibiotics were experimentally measured (**Figure 35** and **Figure 36**). This activity data was then used identify physicochemical descriptors that best classify their relative potencies in the different strains (Chapter IV).

One way to identify properties that trend with antibiotic activity is to make use of the variable importance calculations from tree-based methods such as RF, as other machine learning methods such as support vector machines are less interpretable. In addition, RF is appropriate for a dataset of this size, with experimental activities in the form of minimum inhibitory concentration data for ~50 antibiotics against each of the four strains in both *E. coli* and *P. aeruginosa*. Minimum inhibitor concentration data is discrete rather than continuous, making it suitable for binary classification rather than regression.

Using physicochemical properties of the antibiotics as features to build binary classification models using RF, we were able to identify which features were most important for generating accurate models of the activities in the different strains and species. The physicochemical properties selected by active efflux and the OM barriers were different for the two species. For *P. aeruginosa* antibiotic activity was better classified by electrostatic and surface area properties, whereas topology, physical properties, and atom or bond counts were important for *E. coli* (**Figure 38**). Interestingly, active antibiotics also suffered from significant barrier effects, highlighting the synergy between the two barriers where optimizing for one barrier promotes strengthening of the other barrier. Thus, optimizing molecules with favorable physicochemical properties to overcome both barriers should be considered. These properties provide a set of chemical guidelines that can be used for development and optimization of future antibiotics.

One way to restore the activity of existing antibiotics is to identify inhibitors of multidrug efflux pumps (EPIs, **Figure 34**). Similar to antibiotics, these molecules can be optimized to follow the

physicochemical property guidelines that promote OM permeability and minimize efflux. Using these existing filters in combination with ligand docking, a new class of inhibitors of *E. coli* AcrAB-TolC was identified (**Figure 39**). These six molecules had a shared scaffold and were found to potentiate antibiotic activity to varying degrees in different Gram-negative bacteria.

In addition to these six molecules identified as EPIs, we have docked all ~250,000 purchasable primary amines from the ZINC15 database to AcrA, as compounds that contain primary amines have been shown to permeate the OM. Docking was done to the 29 snapshots previously described and at all four sites on AcrA and ~50 compounds have been selected for experimental assays to measure antibiotic potentiation. We have also built coevolution-based models of AdeA and AdeI, which are AcrA analogs in the pathogen *Acinetobacter baumannii*. Future work will involve docking our primary amine library to conformations from MD simulations of these two proteins to prioritize compounds for experimental testing. These experiments will include antibiotic potentiation assays measuring minimum potentiating concentrations to assess the ability of the top predicted compounds to restore antibiotic activity of a known antibiotic that would otherwise be effluxed from the cell in the absence of an efflux pump inhibitor. The different hyperporinated and efflux-deficient strains of various Gram-negative bacteria could be used to identify specific barriers that limit efficacy of antibiotics or efflux pump inhibitors. This information could then be used to guide rational design of compounds with improved properties that enhance the desired activity. Surface plasmon resonance or a related technique could be used to verify binding of compounds to AdeA or AdeI predicted by the docking calculations.

Enzymatic inactivation of antibiotics is another primary mechanism of antibiotic resistance. The inactivation of β-lactam antibiotics, such as penicillin and amoxicillin, by β-lactamase enzymes is among the most extensively studied. β-lactam antibiotics fight bacterial infections by disrupting bacterial cell wall synthesis, resulting in cell death. A detailed understanding of this inactivation mechanism will inform the development of strategies for overcoming resistance to this commonly prescribed class of antibiotics. The first step of this inactivation reaction involves a proton transfer from Lys73 to Glu166 through Ser70 and an active site water molecule (**Figure 40**). By using a combination of neutron and X-ray crystallography, the protonation states of Lys73 and Glu166 in the active site of the precovalent (cefotaxime-bound) complex were determined (**Figure 42**). These

protonation states are consistent with the concerted base hypothesis[78] where substrate binding triggers this proton transfer and the resulting neutral Lys73 deprotonates Ser70 to then perform a nucleophilic attack on the β-lactam ring of cefotaxime. Cefotaxime binding was also found to initiate several conformational changes in the binding site.

To further investigate the role of the substrate on this initial proton transfer step and to quantify the role of the substrate in facilitating this reaction, QM/MM free energy simulations were performed on both the apo and cefotaxime-bound forms of this enzyme (**Figure 43**). The QM/MM simulations indicate that the proton transfers are concerted and synchronous in both the apoenzyme and the cefotaxime-bound system, as seen in the single free energy barrier in each case (**Figure 44**). In the apoenzyme the estimated proton transfer free energy barrier is ~5 kcal/mol, with the product state (i.e., neutral Lys73 and neutral Glu166) being 2.5 kcal/mol higher in energy than the reactant state (i.e., cationic Lys73 and anionic Glu166). These relative free energies are consistent with the neutron crystal structure of the apoenzyme, in which Lys73 is cationic and Glu166 is anionic. For the cefotaxime bound model the proton transfer barrier was computed to be 2.8 kcal/mol with a reaction free energy of -6.2 kcal/mol. Thus, this proton transfer was found to be more thermodynamically favorable when the substrate is present and the presence of the cefotaxime substrate alters the relative pK_a values of Lys73 and Glu166 to facilitate this reaction. Future work could involve using QM/MM simulations to investigate the remaining steps of this reaction in the presence of cefotaxime to provide additional insights into the effect of substrate binding on this reaction.

REFERENCES

1. Park, H.; Ovchinnikov, S.; Kim, D. E.; DiMaio, F.; Baker, D., Protein homology model refinement by large-scale energy optimization. *Proc Natl Acad Sci U S A* **2018,** *115* (12), 3054-3059.
2. Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C., Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **2011,** *6* (12), e28766.
3. Sulkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N., Genomics-aided structure prediction. *Proc Natl Acad Sci U S A* **2012,** *109* (26), 10340-5.
4. Nugent, T.; Jones, D. T., Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* **2012,** *109* (24), E1540-7.
5. Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S., Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **2012,** *149* (7), 1607-21.
6. Ovchinnikov, S.; Kinch, L.; Park, H.; Liao, Y.; Pei, J.; Kim, D. E.; Kamisetty, H.; Grishin, N. V.; Baker, D., Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **2015,** *4*, e09248.
7. Parks, J. M.; Johs, A.; Podar, M.; Bridou, R.; Hurt, R. A., Jr.; Smith, S. D.; Tomanicek, S. J.; Qian, Y.; Brown, S. D.; Brandt, C. C.; Palumbo, A. V.; Smith, J. C.; Wall, J. D.; Elias, D. A.; Liang, L., The genetic basis for bacterial mercury methylation. *Science* **2013,** *339* (6125), 1332-5.
8. Yu, R. Q.; Reinfelder, J. R.; Hines, M. E.; Barkay, T., Mercury methylation by the methanogen *Methanospirillum hungatei*. *Appl Environ Microbiol* **2013,** *79* (20), 6325-30.
9. Gilmour, C. C.; Podar, M.; Bullock, A. L.; Graham, A. M.; Brown, S. D.; Somenahally, A. C.; Johs, A.; Hurt, R. A., Jr.; Bailey, K. L.; Elias, D. A., Mercury methylation by novel microorganisms from new environments. *Environ Sci Technol* **2013,** *47* (20), 11810-20.
10. Gilmour, C. C.; Bullock, A. L.; McBurney, A.; Podar, M.; Elias, D. A., Robust mercury methylation across diverse methanogenic archaea. *MBio* **2018,** *9* (2).
11. Podar, M.; Gilmour, C. C.; Brandt, C. C.; Soren, A.; Brown, S. D.; Crable, B. R.; Palumbo, A. V.; Somenahally, A. C.; Elias, D. A., Global prevalence and distribution of genes and microorganisms involved in mercury methylation. *Sci Adv* **2015,** *1* (9), e1500675.
12. Gilmour, C. C.; Elias, D. A.; Kucken, A. M.; Brown, S. D.; Palumbo, A. V.; Schadt, C. W.; Wall, J. D., Sulfate-reducing bacterium *Desulfovibrio desulfuricans* ND132 as a model for understanding bacterial mercury methylation. *Appl Environ Microbiol* **2011,** *77* (12), 3938-51.
13. Svetlitchnaia, T.; Svetlitchnyi, V.; Meyer, O.; Dobbek, H., Structural insights into methyltransfer reactions of a corrinoid iron-sulfur protein involved in acetyl-CoA synthesis. *Proc Natl Acad Sci U S A* **2006,** *103* (39), 14331-6.
14. Kung, Y.; Ando, N.; Doukov, T. I.; Blasiak, L. C.; Bender, G.; Seravalli, J.; Ragsdale, S. W.; Drennan, C. L., Visualizing molecular juggling within a B_{12}-dependent methyltransferase complex. *Nature* **2012,** *484* (7393), 265-9.
15. Goetzl, S.; Jeoung, J. H.; Hennig, S. E.; Dobbek, H., Structural basis for electron and methyl-group transfer in a methyltransferase system operating in the reductive acetyl-CoA pathway. *J Mol Biol* **2011,** *411* (1), 96-109.

16. Smith, S. D.; Bridou, R.; Johs, A.; Parks, J. M.; Elias, D. A.; Hurt, R. A., Jr.; Brown, S. D.; Podar, M.; Wall, J. D., Site-directed mutagenesis of HgcA and HgcB reveals amino acid residues important for mercury methylation. *Appl Environ Microbiol* **2015,** *81* (9), 3205-17.
17. Qian, C.; Johs, A.; Chen, H.; Mann, B. F.; Lu, X.; Abraham, P. E.; Hettich, R. L.; Gu, B., Global proteome response to deletion of genes related to mercury methylation and dissimilatory metal reduction reveals changes in respiratory metabolism in *Geobacter sulfurreducens* PCA. *J Proteome Res* **2016,** *15* (10), 3540-3549.
18. Qian, C.; Chen, H.; Johs, A.; Lu, X.; An, J.; Pierce, E. M.; Parks, J. M.; Elias, D. A.; Hettich, R. L.; Gu, B., Quantitative proteomic analysis of biological processes and responses of the bacterium *Desulfovibrio desulfuricans* ND132 upon deletion of its mercury methylation genes. *Proteomics* **2018,** *18* (17), e1700479.
19. Date, S. S.; Parks, J. M.; Rush, K. W.; Wall, J. D.; Ragsdale, S. W.; Johs, A., Kinetics of enzymatic mercury methylation at nanomolar concentrations catalyzed by HgcAB. *Appl Environ Microbiol* **2019,** *85* (13).
20. Nou, X.; Kadner, R. J., Adenosylcobalamin inhibits ribosome binding to btuB RNA. *Proc Natl Acad Sci U S A* **2000,** *97* (13), 7190-5.
21. Nakamura, M.; Saeki, K.; Takahashi, Y., Hyperproduction of recombinant ferredoxins in *Escherichia coli* by coexpression of the ORF1-ORF2-*iscS-iscU-iscA-hscB-hscA-fdx*-ORF3 gene cluster. *J Biochem* **1999,** *126* (1), 10-8.
22. Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P. S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D., Protein structure determination using metagenome sequence data. *Science* **2017,** *355* (6322), 294-298.
23. Pál, C.; Papp, B.; Lercher, M. J., Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics* **2005,** *37* (12), 1372-1375.
24. Nielsen, J.; Keasling, Jay D., Engineering Cellular Metabolism. *Cell* **2016,** *164* (6), 1185-1197.
25. Porse, A.; Schou, T. S.; Munck, C.; Ellabaan, M. M. H.; Sommer, M. O. A., Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nature Communications* **2018,** *9* (1), 522.
26. Michener, J. K.; Camargo Neves, A. A.; Vuilleumier, S.; Bringel, F.; Marx, C. J., Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post–transfer refinement. *eLife* **2014,** *3*, e04279.
27. Clark, I. C.; Melnyk, R. A.; Youngblut, M. D.; Carlson, H. K.; Iavarone, A. T.; Coates, J. D., Synthetic and Evolutionary Construction of a Chlorate-Reducing Shewanella oneidensis MR-1. *mBio* **2015,** *6* (3), e00282-15.
28. Bugg, T. D. H.; Ahmad, M.; Hardiman, E. M.; Rahmanpour, R., Pathways for degradation of lignin in bacteria and fungi. *Natural Product Reports* **2011,** *28* (12), 1883.
29. Chain, P. S. G.; Denef, V. J.; Konstantinidis, K. T.; Vergez, L. M.; Agullo, L.; Reyes, V. L.; Hauser, L.; Cordova, M.; Gomez, L.; Gonzalez, M.; Land, M.; Lao, V.; Larimer, F.; LiPuma, J. J.; Mahenthiralingam, E.; Malfatti, S. A.; Marx, C. J.; Parnell, J. J.; Ramette, A.; Richardson, P.; Seeger, M.; Smith, D.; Spilker, T.; Sul, W. J.; Tsoi, T. V.; Ulrich, L. E.; Zhulin, I. B.; Tiedje, J. M., Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. *Proceedings of the National Academy of Sciences* **2006,** *103* (42), 15280-15287.

30. Clarkson, S. M.; Kridelbaugh, D. M.; Elkins, J. G.; Guss, A. M.; Michener, J. K., Construction and optimization of a heterologous pathway for protocatechuate catabolism in Escherichia coli enables rapid bioconversion of model lignin monomers. *bioRxiv* **2017**.
31. Standaert, R. F.; Giannone, R. J.; Michener, J. K., Identification of parallel and divergent optimization solutions for homologous metabolic enzymes. *Metabolic Engineering Communications* **2018,** *6*, 56-62.
32. Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D., BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Research* **2019,** *47* (D1), D542-D549.
33. Freund, G. S.; O'Brien, T. E.; Vinson, L.; Carlin, D. A.; Yao, A.; Mak, W. S.; Tagkopoulos, I.; Facciotti, M. T.; Tantillo, D. J.; Siegel, J. B., Elucidating Substrate Promiscuity within the FabI Enzyme Family. *ACS Chemical Biology* **2017,** *12* (9), 2465-2473.
34. Combs, S. A.; DeLuca, S. L.; DeLuca, S. H.; Lemmon, G. H.; Nannemann, D. P.; Nguyen, E. D.; Willis, J. R.; Sheehan, J. H.; Meiler, J., Small-molecule ligand docking into comparative models with Rosetta. *Nature Protocols* **2013,** *8* (7), 1277-1298.
35. Skolnick, J.; Zhou, H.; Gao, M., Are predicted protein structures of any value for binding site prediction and virtual ligand screening? *Current Opinion in Structural Biology* **2013,** *23* (2), 191-197.
36. Pierri, C. L.; Parisi, G.; Porcelli, V., Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2010,** *1804* (9), 1695-1712.
37. Jacobson, M. P.; Kalyanaraman, C.; Zhao, S.; Tian, B., Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends in Biochemical Sciences* **2014,** *39* (8), 363-371.
38. Bedbrook, C. N.; Yang, K. K.; Robinson, J. E.; Mackey, E. D.; Gradinaru, V.; Arnold, F. H., Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature Methods* **2019,** *16* (11), 1176-1184.
39. Chevrette, M. G.; Aicheler, F.; Kohlbacher, O.; Currie, C. R.; Medema, M. H., SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics* **2017,** *33* (20), 3202-3210.
40. Röttig, M.; Rausch, C.; Kohlbacher, O., Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families. *PLoS Computational Biology* **2010,** *6* (1), e1000636.
41. Yang, M.; Fehl, C.; Lees, K. V.; Lim, E.-K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G., Functional and informatics analysis enables glycosyltransferase activity prediction. *Nature Chemical Biology* **2018,** *14* (12), 1109-1117.
42. Chhiba-Govindjee, V. P.; van der Westhuyzen, C. W.; Bode, M. L.; Brady, D., Bacterial nitrilases and their regulation. *Applied Microbiology and Biotechnology* **2019,** *103* (12), 4679-4692.
43. Howden, A. J. M.; Preston, G. M., Nitrilase enzymes and their role in plant-microbe interactions. *Microbial Biotechnology* **2009,** *2* (4), 441-451.
44. Robertson, D. E.; Chaplin, J. A.; DeSantis, G.; Podar, M.; Madden, M.; Chi, E.; Richardson, T.; Milan, A.; Miller, M.; Weiner, D. P.; Wong, K.; McQuaid, J.; Farwell, B.; Preston, L. A.; Tan, X.; Snead, M. A.; Keller, M.; Mathur, E.; Kretz, P. L.; Burk, M. J.; Short, J. M., Exploring Nitrilase Sequence Space for Enantioselective Catalysis. *Applied and Environmental Microbiology* **2004,** *70* (4), 2429-2436.

45. Banerjee, A.; Kaul, P.; Sharma, R.; Banerjee, U. C., A High-Throughput Amenable Colorimetric Assay for Enantioselective Screening of Nitrilase-Producing Microorganisms Using pH Sensitive Indicators. *Journal of Biomolecular Screening* **2003,** *8* (5), 559-565.
46. Kobayashi, M.; Shimizu, S., Versatile nitrilases: Nitrile-hydrolysing enzymes. *FEMS Microbiology Letters* **1994,** *120* (3), 217-223.
47. He, Y.-C.; Ma, C.-L.; Xu, J.-H.; Zhou, L., A high-throughput screening strategy for nitrile-hydrolyzing enzymes based on ferric hydroxamate spectrophotometry. *Applied Microbiology and Biotechnology* **2011,** *89* (3), 817-823.
48. Santoshkumar, M.; Nayak, A. S.; Anjaneya, O.; Karegoudar, T. B., A plate method for screening of bacteria capable of degrading aliphatic nitriles. *Journal of Industrial Microbiology & Biotechnology* **2010,** *37* (1), 111-115.
49. Black, G. W.; Brown, N. L.; Perry, J. J. B.; Randall, P. D.; Turnbull, G.; Zhang, M., A high-throughput screening method for determining the substrate scope of nitrilases. *Chemical Communications* **2015,** *51* (13), 2660-2662.
50. Current Epidemiology and Growing Resistance of Gram-Negative Pathogens FAU - Livermore, David M. *Korean J Intern Med* **2012,** *27* (2), 128-142.
51. Tamber, S.; Hancock, R. E., The outer membranes of Pseudomonads. In *Pseudomonas*, Ramos, J.-L., Ed. Kluwer Academic/Plenum Publishers: New York, 2004; Vol. 1, pp 575-601.
52. Pages, J. M.; James, C. E.; Winterhalter, M., The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria. *Nat Rev Microbiol* **2008,** *6* (12), 893-903.
53. Zgurskaya, H. I.; Krishnamoorthy, G.; Tikhonova, E. B.; Lau, S. Y.; Stratton, K. L., Mechanism of antibiotic efflux in Gram-negative bacteria. *Front Biosci* **2003,** *8*, s862-73.
54. Trusts, P. C., A Scientific Roadmap for Antibiotic Discovery: a sustainable and robust pipeline of new antibacterial drugs and therapies is critical to preserve public health. project, A. r., Ed. http://www.pewtrusts.org/en/projects/antibiotic-resistance-project, 2016.
55. Zgurskaya, H. I.; Lopez, C. A.; Gnanakaran, S., Permeability Barrier of Gram-Negative Cell Envelopes and Approaches To Bypass It. *ACS Infect Dis* **2015,** *1* (11), 512-522.
56. Lewis, K., Antibiotics: Recover the lost art of drug discovery. *Nature* **2012,** *485* (7399), 439-440.
57. Krishnamoorthy, G.; Leus, I. V.; Weeks, J. W.; Wolloscheck, D.; Rybenkov, V. V.; Zgurskaya, H. I., Synergy between Active Efflux and Outer Membrane Diffusion Defines Rules of Antibiotic Permeation into Gram-Negative Bacteria. *MBio* **2017,** *8* (5).
58. Krishnamoorthy, G.; Wolloscheck, D.; Weeks, J. W.; Croft, C.; Rybenkov, V. V.; Zgurskaya, H. I., Breaking the Permeability Barrier of Escherichia coli by Controlled Hyperporination of the Outer Membrane. *Antimicrob Agents Chemother* **2016,** *60* (12), 7372-7381.
59. Bystrova, O. V.; Lindner, B.; Moll, H.; Kocharova, N. A.; Knirel, Y. A.; Zahringer, U.; Pier, G. B., Full structure of the lipopolysaccharide of Pseudomonas aeruginosa immunotype 5. *Biochemistry (Mosc)* **2004,** *69* (2), 170-5.
60. Nikaido, H., Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* **2003,** *67* (4), 593-656.
61. Nikaido, H., Porins and specific diffusion channels in bacterial outer membranes. *The Journal of biological chemistry* **1994,** *269* (6), 3905-8.
62. Chevalier, S.; Bouffartigues, E.; Bodilis, J.; Maillot, O.; Lesouhaitier, O.; Feuilloley, M. G. J.; Orange, N.; Dufour, A.; Cornelis, P., Structure, function and regulation of Pseudomonas aeruginosa porins. *FEMS Microbiol Rev* **2017,** *41* (5), 698-722.

63. Zgurskaya, H. I.; Nikaido, H., Bypassing the periplasm: reconstitution of the AcrAB multidrug efflux pump of *Escherichia coli*. *Proc Natl Acad Sci U S A* **1999,** *96* (13), 7190-5.
64. Murakami, S.; Nakashima, R.; Yamashita, E.; Matsumoto, T.; Yamaguchi, A., Crystal structures of a multidrug transporter reveal a functionally rotating mechanism. *Nature* **2006,** *443* (7108), 173-9.
65. Murakami, S.; Nakashima, R.; Yamashita, E.; Yamaguchi, A., Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature* **2002,** *419* (6907), 587-93.
66. Koronakis, V.; Sharff, A.; Koronakis, E.; Luisi, B.; Hughes, C., Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **2000,** *405* (6789), 914-9.
67. Richter, M. F.; Drown, B. S.; Riley, A. P.; Garcia, A.; Shirai, T.; Svec, R. L.; Hergenrother, P. J., Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **2017,** *545* (7654), 299-304.
68. Fleming, A., On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae. *Br. J. Exp. Pathol.* **1929,** *10*, 226-236.
69. Chain, E.; Florey, H. W.; Adelaide, M. B.; Gardner, A. D.; Oxfd, D. M.; Heatley, N. G.; Jennings, M. A.; Orr-Ewing, J.; Sanders, A. G., PENICILLIN AS A CHEMOTHERAPEUTIC AGENT. *The Lancet* **1940,** *236* (6104), 226-228.
70. Ambler, R. P.; Coulson, A. F. W.; Frere, J. M.; Ghuysen, J. M.; Joris, B.; Forsman, M.; Levesque, R. C.; Tiraby, G.; Waley, S. G., A standard numbering scheme for the class A β-lactamases [1]. *Biochemical Journal* **1991,** *276* (1), 269-270.
71. Matagne, A.; Lamotte-Brasseur, J.; Frère, J. M., Catalytic properties of class A β-lactamases: Efficiency and diversity. *Biochemical Journal* **1998,** *330* (2), 581-598.
72. Tzouvelekis, L. S.; Tzelepi, E.; Tassios, P. T.; Legakis, N. J., CTX-M-type β-lactamases: An emerging group of extended-spectrum enzymes. *International Journal of Antimicrobial Agents* **2000,** *14* (2), 137-142.
73. Ishii, Y.; Ohno, A.; Taguchi, H.; Imajo, S.; Ishiguro, M.; Matsuzawa, H., Cloning and sequence of the gene encoding a cefotaxime-hydrolyzing class A β-lactamase isolated from Escherichia coli. *Antimicrobial Agents and Chemotherapy* **1995,** *39* (10), 2269-2275.
74. Minasov, G.; Wang, X.; Shoichet, B. K., An ultrahigh resolution structure of TEM-1 β-lactamase suggests a role for Glu166 as the general base in acylation. *Journal of the American Chemical Society* **2002,** *124* (19), 5333-5340.
75. Escobar, W. A.; Tan, A. K.; Lewis, E. R.; Fink, A. L., Site-Directed Mutagenesis of Glutamate-166 in β-Lactamase Leads to a Branched Path Mechanism. *Biochemistry* **1994,** *33* (24), 7619-7626.
76. Golemi-Kotra, D.; Meroueh, S. O.; Kim, C.; Vakulenko, S. B.; Bulychev, A.; Stemmler, A. J.; Stemmler, T. L.; Mobashery, S., The importance of a critical protonation state and the fate of the catalytic steps in class A β-lactamases and penicillin-binding proteins. *Journal of Biological Chemistry* **2004,** *279* (33), 34665-34673.
77. Lietz, E. J.; Truher, H.; Kahn, D.; Hokenson, M. J.; Fink, A. L., Lysine-73 is involved in the acylation and deacylation of β-lactamase. *Biochemistry* **2000,** *39* (17), 4971-4981.
78. Meroueh, S. O.; Fisher, J. F.; Schlegel, H. B.; Mobashery, S., Ab initio QM/MM study of class A β-lactamase acylation: Dual participation of Glu166 and Lys73 in a concerted base promotion of Ser70. *Journal of the American Chemical Society* **2005,** *127* (44), 15397-15407.

79. Hermann, J. C.; Pradon, J.; Harvey, J. N.; Mulholland, A. J., High level QM/MM modeling of the formation of the tetrahedral intermediate in the acylation of wild type and K73A mutant TEM-I class A β-lactamase. *Journal of Physical Chemistry A* **2009,** *113* (43), 11984-11994.
80. Hermann, J. C.; Hensen, C.; Ridder, L.; Mulholland, A. J.; Höltje, H. D., Mechanisms of antibiotic resistance: QM/MM modeling of the acylation reaction of a class A β-lactamase with benzylpenicillin. *Journal of the American Chemical Society* **2005,** *127* (12), 4454-4465.
81. Hermann, J. C.; Ridder, L.; Mulholland, A. J.; Höltje, H. D., Identification of Glu166 as the general base in the acylation reaction of class A, β-lactamases through QM/MM modeling. *Journal of the American Chemical Society* **2003,** *125* (32), 9590-9591.
82. Hermann, J. C.; Ridder, L.; Höltje, H. D.; Mulholland, A. J., Molecular mechanisms of antibiotic resistance: QM/MM modelling of deacylation in a class A β-lactamase. *Organic and Biomolecular Chemistry* **2006,** *4* (2), 206-210.
83. Chudyk, E. I.; Limb, M. A. L.; Jones, C.; Spencer, J.; Van Der Kamp, M. W.; Mulholland, A. J., QM/MM simulations as an assay for carbapenemase activity in class A β-lactamases. *Chemical Communications* **2014,** *50* (94), 14736-14739.
84. Zhou, J.; Riccardi, D.; Beste, A.; Smith, J. C.; Parks, J. M., Mercury methylation by HgcA: theory supports carbanion transfer to Hg(II). *Inorg Chem* **2014,** *53* (2), 772-7.
85. Rempel, S.; Colucci, E.; de Gier, J. W.; Guskov, A.; Slotboom, D. J., Cysteine-mediated decyanation of vitamin B_{12} by the predicted membrane transporter BtuM. *Nat Commun* **2018,** *9* (1), 3038.
86. Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; UniProt, C., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015,** *31* (6), 926-32.
87. Chen, I. A.; Chu, K.; Palaniappan, K.; Pillay, M.; Ratner, A.; Huang, J.; Huntemann, M.; Varghese, N.; White, J. R.; Seshadri, R.; Smirnova, T.; Kirton, E.; Jungbluth, S. P.; Woyke, T.; Eloe-Fadrosh, E. A.; Ivanova, N. N.; Kyrpides, N. C., IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **2019,** *47* (D1), D666-D677.
88. Brown, S. D.; Gilmour, C. C.; Kucken, A. M.; Wall, J. D.; Elias, D. A.; Brandt, C. C.; Podar, M.; Chertkov, O.; Held, B.; Bruce, D. C.; Detter, J. C.; Tapia, R.; Han, C. S.; Goodwin, L. A.; Cheng, J. F.; Pitluck, S.; Woyke, T.; Mikhailova, N.; Ivanova, N. N.; Han, J.; Lucas, S.; Lapidus, A. L.; Land, M. L.; Hauser, L. J.; Palumbo, A. V., Genome sequence of the mercury-methylating strain *Desulfovibrio desulfuricans* ND132. *J Bacteriol* **2011,** *193* (8), 2078-9.
89. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **2011,** *9* (2), 173-5.
90. Soding, J., Protein homology detection by HMM-HMM comparison. *Bioinformatics* **2005,** *21* (7), 951-60.
91. Kamisetty, H.; Ovchinnikov, S.; Baker, D., Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **2013,** *110* (39), 15674-9.
92. Balakrishnan, S.; Kamisetty, H.; Carbonell, J. G.; Lee, S. I.; Langmead, C. J., Learning generative models for protein fold families. *Proteins* **2011,** *79* (4), 1061-78.
93. Wang, G.; Dunbrack, R. L., Jr., PISCES: a protein sequence culling server. *Bioinformatics* **2003,** *19* (12), 1589-91.
94. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M.

S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput* **2017,** *13* (6), 3031-3048.
95. Zhang, Y.; Skolnick, J., Scoring function for automated assessment of protein structure template quality. *Proteins* **2004,** *57* (4), 702-10.
96. Fleishman, S. J.; Leaver-Fay, A.; Corn, J. E.; Strauch, E. M.; Khare, S. D.; Koga, N.; Ashworth, J.; Murphy, P.; Richter, F.; Lemmon, G.; Meiler, J.; Baker, D., RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **2011,** *6* (6), e20161.
97. Dauter, Z.; Wilson, K. S.; Sieker, L. C.; Meyer, J.; Moulis, J. M., Atomic resolution (0.94 A) structure of Clostridium acidurici ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein. *Biochemistry* **1997,** *36* (51), 16065-73.
98. Conway, P.; Tyka, M. D.; DiMaio, F.; Konerding, D. E.; Baker, D., Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci* **2014,** *23* (1), 47-55.
99. Holm, L.; Laakso, L. M., Dali server update. *Nucleic Acids Res* **2016,** *44* (W1), W351-5.
100. Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 2.0. 2015.
101. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **2004,** *32* (5), 1792-7.
102. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; Thierer, T.; Ashton, B.; Meintjes, P.; Drummond, A., Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012,** *28* (12), 1647-9.
103. Price, M. N.; Dehal, P. S.; Arkin, A. P., FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **2010,** *5* (3), e9490.
104. Letunic, I.; Bork, P., Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **2019,** *47* (W1), W256-W259.
105. Cooper, C. J., Ovchinnikov, S., Zheng, K., Rush, K.W., Podar, M., Pavlopoulos, G., Kyrpides, N.C., Johs, A., Ragsdale, S.W., and Parks, J.M. , Structure determination of the HgcAB complex using metagenome sequence data: insights into the mechanism of mercury methylation. *Commun. Biol.* **2020,** *In press*.
106. Firth, R. A.; Hill, H. A. O.; Mann, B. E.; Pratt, J. M.; Thorp, R. G.; Williams, R. J. P., The chemistry of vitamin B_{12}. Part IX. Evidence for five-coordinate cobalt(III) complexes. *J Chem Soc A Inorg Phys Theor* **1968**.
107. Giannotti, C., Electronic spectra of B_{12} and related systems. In *B_{12}*, Dolphin, D., Ed. John Wiley & Sons, Inc.: USA, 1982; Vol. 1, pp 393-430.
108. Gionfriddo, C. M.; Tate, M. T.; Wick, R. R.; Schultz, M. B.; Zemla, A.; Thelen, M. P.; Schofield, R.; Krabbenhoft, D. P.; Holt, K. E.; Moreau, J. W., Microbial mercury methylation in Antarctic sea ice. *Nat Microbiol* **2016,** *1* (10), 16127.
109. Nicolet, Y.; Piras, C.; Legrand, P.; Hatchikian, C. E.; Fontecilla-Camps, J. C., *Desulfovibrio desulfuricans* iron hydrogenase: the structure shows unusual coordination to an active site Fe binuclear center. *Structure* **1999,** *7* (1), 13-23.
110. Hattori, M.; Tanaka, Y.; Fukai, S.; Ishitani, R.; Nureki, O., Crystal structure of the MgtE Mg^{2+} transporter. *Nature* **2007,** *448* (7157), 1072-5.
111. Adman, E. T.; Sieker, L. C.; Jensen, L. H., Structure of a bacterial ferredoxin. *J Biol Chem* **1973,** *248* (11), 3987-96.

112. Unciuleac, M.; Boll, M.; Warkentin, E.; Ermler, U., Crystallization of 4-hydroxybenzoyl-CoA reductase and the structure of its electron donor ferredoxin. *Acta Crystallogr D Biol Crystallogr* **2004,** *60* (Pt 2), 388-91.
113. DiMaio, F.; Leaver-Fay, A.; Bradley, P.; Baker, D.; Andre, I., Modeling symmetric macromolecular structures in Rosetta3. *PLoS One* **2011,** *6* (6), e20450.
114. Morcos, F.; Jana, B.; Hwa, T.; Onuchic, J. N., Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc Natl Acad Sci U S A* **2013,** *110* (51), 20533-8.
115. Dowling, D. P.; Croft, A. K.; Drennan, C. L., Radical use of Rossmann and TIM barrel architectures for controlling coenzyme B_{12} chemistry. *Annu Rev Biophys* **2012,** *41*, 403-27.
116. Barkay, T.; Miller, S. M.; Summers, A. O., Bacterial mercury resistance from atoms to ecosystems. *FEMS Microbiology Reviews* **2003,** *27* (2-3), 355-384.
117. Moore, M. J.; Miller, S. M.; Walsh, C. T., C-terminal cysteines of Tn501 mercuric ion reductase. *Biochemistry* **1992,** *31* (6), 1677-85.
118. Menon, S.; Ragsdale, S. W., Role of the [4Fe-4S] cluster in reductive activation of the cobalt center of the corrinoid iron-sulfur protein from *Clostridium thermoaceticum* during acetate biosynthesis. *Biochemistry* **1998,** *37* (16), 5689-98.
119. Menon, S.; Ragsdale, S. W., The role of an iron-sulfur cluster in an enzymatic methylation reaction. Methylation of CO dehydrogenase/acetyl-CoA synthase by the methylated corrinoid iron-sulfur protein. *J Biol Chem* **1999,** *274* (17), 11513-8.
120. Demissie, T. B.; Garabato, B. D.; Ruud, K.; Kozlowski, P. M., Mercury methylation by cobalt corrinoids: relativistic effects dictate the reaction mechanism. *Angew Chem Int Ed* **2016,** *55* (38), 11503-6.
121. Ragauskas, A. J.; Beckham, G. T.; Biddy, M. J.; Chandra, R.; Chen, F.; Davis, M. F.; Davison, B. H.; Dixon, R. A.; Gilna, P.; Keller, M.; Langan, P.; Naskar, A. K.; Saddler, J. N.; Tschaplinski, T. J.; Tuskan, G. A.; Wyman, C. E., Lignin valorization: improving lignin processing in the biorefinery. *Science* **2014,** *344* (6185), 1246843.
122. Masai, E.; Katayama, Y.; Fukuda, M., Genetic and Biochemical Investigations on Bacterial Catabolic Pathways for Lignin-Derived Aromatic Compounds. *Bioscience, Biotechnology, and Biochemistry* **2007,** *71* (1), 1-15.
123. Parke, D.; Ornston, L. N., Hydroxycinnamate (hca) Catabolic Genes from Acinetobacter sp. Strain ADP1 Are Repressed by HcaR and Are Induced by Hydroxycinnamoyl-Coenzyme A Thioesters. *Applied and Environmental Microbiology* **2003,** *69* (9), 5398-5409.
124. Otani, H.; Lee, Y.-E.; Casabon, I.; Eltis, L. D., Characterization of p-Hydroxycinnamate Catabolism in a Soil Actinobacterium. *Journal of Bacteriology* **2014,** *196* (24), 4293-4303.
125. Kim, J.; Copley, S. D., Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. *Proceedings of the National Academy of Sciences* **2012,** *109* (42), E2856-E2864.
126. Zimmermann, L.; Stephens, A.; Nam, S. Z.; Rau, D.; Kubler, J.; Lozajic, M.; Gabler, F.; Soding, J.; Lupas, A. N.; Alva, V., A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **2018,** *430* (15), 2237-2243.
127. Katoh, K.; Standley, D. M., MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013,** *30* (4), 772-780.

128. Song, Y.; DiMaio, F.; Wang, Ray Y.-R.; Kim, D.; Miles, C.; Brunette, T.; Thompson, J.; Baker, D., High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013,** *21* (10), 1735-1742.
129. Kim, D. E.; Chivian, D.; Baker, D., Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research* **2004,** *32* (Web Server), W526-W531.
130. Gront, D.; Kulp, D. W.; Vernon, R. M.; Strauss, C. E. M.; Baker, D., Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. *PLoS ONE* **2011,** *6* (8), e23294.
131. Makowska-Grzyska, M.; Kim, Y.; Maltseva, N.; Osipiuk, J.; Gu, M.; Zhang, M.; Mandapati, K.; Gollapalli, D. R.; Gorla, S. K.; Hedstrom, L.; Joachimiak, A., A Novel Cofactor-binding Mode in Bacterial IMP Dehydrogenases Explains Inhibitor Selectivity. *Journal of Biological Chemistry* **2015,** *290* (9), 5893-5911.
132. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling* **2012,** *52* (7), 1757-1768.
133. Meiler, J.; Baker, D., ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65* (3), 538-548.
134. Humphrey, W.; Dalke, A.; Schulten, K., VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996,** *14* (1), 33-38.
135. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. A.03*, Wallingford, CT, 2016.
136. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K., Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005,** *26* (16), 1781-1802.
137. Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *Journal of Chemical Theory and Computation* **2012,** *8* (9), 3257-3273.
138. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983,** *79* (2), 926-935.
139. Jo, S.; Kim, T.; Iyer, V. G.; Im, W., CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **2008,** *29* (11), 1859-1865.
140. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology* **2017,** *13* (7), e1005659.

141. Close, D. M.; Cooper, C. J.; Wang, X.; Chirania, P.; Gupta, M.; Ossyra, J. R.; Giannone, R. J.; Engle, N.; Tschaplinski, T. J.; Smith, J. C.; Hedstrom, L.; Parks, J. M.; Michener, J. K., Horizontal transfer of a pathway for coumarate catabolism unexpectedly inhibits purine nucleotide biosynthesis. *Mol Microbiol* **2019,** *112* (6), 1784-1797.
142. Saxer, G.; Krepps, M. D.; Merkley, E. D.; Ansong, C.; Deatherage Kaiser, B. L.; Valovska, M.-T.; Ristic, N.; Yeh, P. T.; Prakash, V. P.; Leiser, O. P.; Nakhleh, L.; Gibbons, H. S.; Kreuzer, H. W.; Shamoo, Y., Mutations in Global Regulators Lead to Metabolic Selection during Adaptation to Complex Environments. *PLoS Genetics* **2014,** *10* (12), e1004872.
143. Kurnasov, O. V.; Polanuyer, B. M.; Ananta, S.; Sloutsky, R.; Tam, A.; Gerdes, S. Y.; Osterman, A. L., Ribosylnicotinamide Kinase Domain of NadR Protein: Identification and Implications in NAD Biosynthesis. *Journal of Bacteriology* **2002,** *184* (24), 6906-6917.
144. Hedstrom, L., IMP Dehydrogenase: Structure, Mechanism, and Inhibition. *Chemical Reviews* **2009,** *109* (7), 2903-2928.
145. Pisithkul, T.; Jacobson, T. B.; O'Brien, T. J.; Stevenson, D. M.; Amador-Noguez, D., Phenolic Amides Are Potent Inhibitors of *De Novo* Nucleotide Biosynthesis. *Applied and Environmental Microbiology* **2015,** *81* (17), 5761-5772.
146. Kizer, L.; Pitera, D. J.; Pfleger, B. F.; Keasling, J. D., Application of Functional Genomics to Pathway Optimization for Increased Isoprenoid Production. *Applied and Environmental Microbiology* **2008,** *74* (10), 3229-3241.
147. Michener, J. K.; Nielsen, J.; Smolke, C. D., Identification and treatment of heme depletion attributed to overexpression of a lineage of evolved P450 monooxygenases. *Proceedings of the National Academy of Sciences* **2012,** *109* (47), 19504-19509.
148. Mills, T. Y.; Sandoval, N. R.; Gill, R. T., Cellulosic hydrolysate toxicity and tolerance mechanisms in Escherichia coli. *Biotechnology for Biofuels* **2009,** *2* (1), 26.
149. Clarkson, S. M.; Hamilton-Brehm, S. D.; Giannone, R. J.; Engle, N. L.; Tschaplinski, T. J.; Hettich, R. L.; Elkins, J. G., A comparative multidimensional LC-MS proteomic analysis reveals mechanisms for furan aldehyde detoxification in Thermoanaerobacter pseudethanolicus 39E. *Biotechnology for Biofuels* **2014,** *7* (1), 165.
150. Yi, X.; Gu, H.; Gao, Q.; Liu, Z. L.; Bao, J., Transcriptome analysis of Zymomonas mobilis ZM4 reveals mechanisms of tolerance and detoxification of phenolic aldehyde inhibitors from lignocellulose pretreatment. *Biotechnology for Biofuels* **2015,** *8* (1), 153.
151. Huang, X.; Holden, H. M.; Raushel, F. M., Channeling of Substrates and Intermediates in Enzyme-Catalyzed Reactions. *Annual Review of Biochemistry* **2001,** *70* (1), 149-180.
152. Mukhopadhyay, A., Tolerance engineering in bacteria for the production of advanced biofuels and chemicals. *Trends in Microbiology* **2015,** *23* (8), 498-508.
153. Zhao, S.; Kumar, R.; Sakai, A.; Vetting, M. W.; Wood, B. M.; Brown, S.; Bonanno, J. B.; Hillerich, B. S.; Seidel, R. D.; Babbitt, P. C.; Almo, S. C.; Sweedler, J. V.; Gerlt, J. A.; Cronan, J. E.; Jacobson, M. P., Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **2013,** *502* (7473), 698-702.
154. Pertusi, D. A.; Moura, M. E.; Jeffryes, J. G.; Prabhu, S.; Walters Biggs, B.; Tyo, K. E. J., Predicting novel substrates for enzymes with minimal experimental effort with active learning. *Metabolic Engineering* **2017,** *44*, 171-181.
155. Podar, M.; Eads, J. R.; Richardson, T. H., Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. *BMC Evolutionary Biology* **2005,** *5* (1), 42.
156. Timm, C. M.; Campbell, A. G.; Utturkar, S. M.; Jun, S.-R.; Parales, R. E.; Tan, W. A.; Robeson, M. S.; Lu, T.-Y. S.; Jawdy, S.; Brown, S. D.; Ussery, D. W.; Schadt, C. W.; Tuskan,

G. A.; Doktycz, M. J.; Weston, D. J.; Pelletier, D. A., Metabolic functions of Pseudomonas fluorescens strains from Populus deltoides depend on rhizosphere or endosphere isolation compartment. *Frontiers in Microbiology* **2015,** *6*.
157. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **2011,** *7* (1), 539.
158. Zhang, L.; Yin, B.; Wang, C.; Jiang, S.; Wang, H.; Yuan, Y. A.; Wei, D., Structural insights into enzymatic activity and substrate specificity determination by a single amino acid in nitrilase from Syechocystis sp. PCC6803. *Journal of Structural Biology* **2014,** *188* (2), 93-101.
159. Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E., POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *Journal of Chemical Theory and Computation* **2014,** *10* (11), 5047-5056.
160. Sterling, T.; Irwin, J. J., ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015,** *55* (11), 2324-2337.
161. Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J., Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *The Journal of Physical Chemistry* **1994,** *98* (45), 11623-11627.
162. Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Al-Laham, M. A.; Shirley, W. A.; Mantzaris, J., A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *The Journal of Chemical Physics* **1988,** *89* (4), 2193-2218.
163. 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. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.
164. Davis, I. W.; Baker, D., RosettaLigand Docking with Full Ligand and Receptor Flexibility. *Journal of Molecular Biology* **2009,** *385* (2), 381-392.
165. Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J., The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017,** *13* (6), 3031-3048.
166. *Molecular Operating Environment (MOE)* 2016; Chemical Computing Group Inc: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2013.08.
167. Reed, A. E.; Weinstock, R. B.; Weinhold, F., Natural population analysis. *The Journal of Chemical Physics* **1985,** *83* (2), 735-746.
168. Besler, B. H.; Merz, K. M.; Kollman, P. A., Atomic charges derived from semiempirical methods. *Journal of Computational Chemistry* **1990,** *11* (4), 431-439.
169. Singh, U. C.; Kollman, P. A., An approach to computing electrostatic charges for molecules. *Journal of Computational Chemistry* **1984,** *5* (2), 129-145.
170. Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y., ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics* **2015,** *16* (1), 162.
171. Breiman, L., Random Forests. *Machine Learning* **2001,** *45* (1), 5-32.
172. Friedman, J. H., Stochastic gradient boosting. *Computational Statistics & Data Analysis* **2002,** *38* (4), 367-378.
173. Cortes, C.; Vapnik, V., Support-vector networks. *Machine Learning* **1995,** *20* (3), 273-297.

174. Mou, Z.; Eakes, J.; Cooper, C.; Foster, C.; Standaert, R.; Podar, M.; Doktycz, M.; Parks, J., Machine Learning-based Prediction of Enzyme Substrate Scope: Application to Bacterial Nitrilases. **2020**.
175. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P., SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **2002,** *16*, 321-357.
176. Fernandes, B. C. M.; Mateo, C.; Kiziak, C.; Chmura, A.; Wacker, J.; van Rantwijk, F.; Stolz, A.; Sheldon, R. A., Nitrile Hydratase Activity of a Recombinant Nitrilase. *Advanced Synthesis & Catalysis* **2006,** *348* (18), 2597-2603.
177. Bertolani, S. J.; Siegel, J. B., A new benchmark illustrates that integration of geometric constraints inferred from enzyme reaction chemistry can increase enzyme active site modeling accuracy. *PLOS ONE* **2019,** *14* (4), e0214126.
178. Hermann, J. C.; Marti-Arbona, R.; Fedorov, A. A.; Fedorov, E.; Almo, S. C.; Shoichet, B. K.; Raushel, F. M., Structure-based activity prediction for an enzyme of unknown function. *Nature* **2007,** *448* (7155), 775-779.
179. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **2010,** *5* (4), 725-738.
180. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **2015,** *12* (1), 7-8.
181. Yang, J.; Zhang, Y., I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Research* **2015,** *43* (W1), W174-W181.
182. Webb, B.; Sali, A., Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* **2016,** *54* (1).
183. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **2018,** *46* (W1), W296-W303.
184. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *Journal of Medicinal Chemistry* **2006,** *49* (21), 6177-6196.
185. Trott, O.; Olson, A. J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2009**, NA-NA.
186. Cao, D.-S.; Xiao, N.; Xu, Q.-S.; Chen, A. F., Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* **2015,** *31* (2), 279-281.
187. Yap, C. W., PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011,** *32* (7), 1466-1474.
188. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T., Mordred: a molecular descriptor calculator. *Journal of Cheminformatics* **2018,** *10* (1), 4.
189. Sharma, N.; Verma, R.; Savitri; Bhalla, T. C., Classifying nitrilases as aliphatic and aromatic using machine learning technique. *3 Biotech* **2018,** *8* (1), 68.
190. Nikaido, H., Multidrug resistance in bacteria. *Annu Rev Biochem* **2009,** *78*, 119-46.
191. Silver, L. L., A Gestalt approach to Gram-negative entry. *Bioorg Med Chem* **2016,** *24* (24), 6379-6389.

192. Westfall, D. A.; Krishnamoorthy, G.; Wolloscheck, D.; Sarkar, R.; Zgurskaya, H. I.; Rybenkov, V. V., Bifurcation kinetics of drug uptake by Gram-negative bacteria. *PLoS One* **2017,** *12* (9), e0184671.
193. Lomovskaya, O.; Lee, A.; Hoshino, K.; Ishida, H.; Mistry, A.; Warren, M. S.; Boyer, E.; Chamberland, S.; Lee, V. J., Use of a genetic approach to evaluate the consequences of inhibition of efflux pumps in Pseudomonas aeruginosa. *Antimicrob Agents Chemother* **1999,** *43* (6), 1340-6.
194. Cao, L.; Srikumar, R.; Poole, K., MexAB-OprM hyperexpression in NalC-type multidrug-resistant Pseudomonas aeruginosa: identification and characterization of the nalC gene encoding a repressor of PA3720-PA3719. *Mol Microbiol* **2004,** *53* (5), 1423-36.
195. De Kievit, T. R.; Parkins, M. D.; Gillis, R. J.; Srikumar, R.; Ceri, H.; Poole, K.; Iglewski, B. H.; Storey, D. G., Multidrug Efflux Pumps: Expression Patterns and Contribution to Antibiotic Resistance in Pseudomonas aeruginosa Biofilms. *Antimicrob Agents Chemother* **2001,** *45* (6), 1761-70.
196. Du, D.; Wang, Z.; James, N. R.; Voss, J. E.; Klimont, E.; Ohene-Agyei, T.; Venter, H.; Chiu, W.; Luisi, B. F., Structure of the AcrAB-TolC multidrug efflux pump. *Nature* **2014,** *509* (7501), 512-5.
197. Fralick, J. A., Evidence that TolC is required for functioning of the Mar/AcrAB efflux pump of Escherichia coli. *J Bacteriol* **1996,** *178* (19), 5803-5.
198. Morona, R.; Manning, P. A.; Reeves, P., Identification and characterization of the TolC protein, an outer membrane protein from Escherichia coli. *J Bacteriol* **1983,** *153* (2), 693-9.
199. Zgurskaya, H. I.; Krishnamoorthy, G.; Ntreh, A.; Lu, S., Mechanism and Function of the Outer Membrane Channel TolC in Multidrug Resistance and Physiology of Enterobacteria. *Front Microbiol* **2011,** *2*, 189.
200. Poole, K.; Srikumar, R., Multidrug efflux in Pseudomonas aeruginosa: components, mechanisms and clinical significance. *Curr Top Med Chem* **2001,** *1* (1), 59-71.
201. Cooper, S. J.; Krishnamoorthy, G.; Wolloscheck, D.; Walker, J. K.; Rybenkov, V. V.; Parks, J. M.; Zgurskaya, H. I., Molecular Properties That Define the Activities of Antibiotics in Escherichia coli and Pseudomonas aeruginosa. *ACS Infect Dis* **2018,** *4* (8), 1223-1234.
202. Acosta-Gutierrez, S.; Ferrara, L.; Pathania, M.; Masi, M.; Wang, J.; Bodrenko, I.; Zahn, M.; Winterhalter, M.; Stavenger, R. A.; Pages, J. M.; Naismith, J. H.; van den Berg, B.; Page, M. G. P.; Ceccarelli, M., Getting drugs into Gram-negative bacteria: Rational rules for permeation through general porins. *ACS infectious diseases* **2018,** *4* (10), 1487-1498.
203. Irwin, J. J.; Shoichet, B. K., ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **2005,** *45* (1), 177-182.
204. Marvin 17.9.0, ChemAxon: Cambridge, MA, 2017; http://www.chemaxon.com. .
205. P.R., G.; K., M., MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. *J Comput Aided Mol Des.* **1995,** *9* (251-268).
206. Molecular Operating Environment (MOE), 2012.10. Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2015.
207. Leach, A. R.; Gillet, V. J., *An Introduction to Chemoinformatics.* Springer Publishing Company, Inc.: 2007.
208. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B. *caret: Classification and Regression Training. R package version 6.0-73.* , https://CRAN.R-project.org/package=caret, 2016.

209. Team, R. C. R: A language and environment for statistical computing.
210. Darzynkiewicz, Z. M.; Green, A. T.; Abdali, N.; Hazel, A.; Fulton, R. L.; Kimball, J.; Gryczynski, Z.; Gumbart, J. C.; Parks, J. M.; Smith, J. C.; Zgurskaya, H. I., Identification of Binding Sites for Efflux Pump Inhibitors of the AcrAB-TolC Component AcrA. *Biophys J* **2019,** *116* (4), 648-658.
211. Ellingson, S. R.; Miao, Y.; Baudry, J.; Smith, J. C., Multi-conformer ensemble docking to difficult protein targets. *J Phys Chem B* **2015,** *119* (3), 1026-34.
212. Ellingson, S. R.; Smith, J. C.; Baudry, J., VinaMPI: facilitating multiple receptor high-throughput virtual docking on high-performance computers. *J Comput Chem* **2013,** *34* (25), 2212-21.
213. Wolloscheck, D.; Krishnamoorthy, G.; Nguyen, J.; Zgurskaya, H. I., Kinetic control of quorum sensing in Pseudomonas aeruginosa by multidrug efflux pumps. *ACS Infect Dis* **2017**.
214. Zhang, Y.; Bao, Q.; Gagnon, L. A.; Huletsky, A.; Oliver, A.; Jin, S.; Langaee, T., ampG Gene of Pseudomonas aeruginosa and Its Role in β-Lactamase Expression. *Antimicrobial Agents and Chemotherapy* **2010,** *54* (11), 4772-4779.
215. Juan, C.; Maciá, M. D.; Gutiérrez, O.; Vidal, C.; Pérez, J. L.; Oliver, A., Molecular Mechanisms of β-Lactam Resistance Mediated by AmpC Hyperproduction in Pseudomonas aeruginosa Clinical Strains. *Antimicrobial Agents and Chemotherapy* **2005,** *49* (11), 4733-4738.
216. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980,** *36* (22), 3219-3228.
217. Richter, M. F.; Drown, B. S.; Riley, A. P.; Garcia, A.; Shirai, T.; Svec, R. L.; Hergenrother, P. J., Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **2017**.
218. Ochs, M. M.; McCusker, M. P.; Bains, M.; Hancock, R. E. W., Negative Regulation of the Pseudomonas aeruginosa Outer Membrane Porin OprD Selective for Imipenem and Basic Amino Acids. *Antimicrobial Agents and Chemotherapy* **1999,** *43* (5), 1085-1090.
219. Abdali, N.; Parks, J. M.; Haynes, K. M.; Chaney, J. L.; Green, A. T.; Wolloscheck, D.; Walker, J. K.; Rybenkov, V. V.; Baudry, J.; Smith, J. C.; Zgurskaya, H. I., Reviving antibiotics: Efflux pump inhibitors that interact with AcrA, a membrane fusion protein of the AcrAB-TolC multidrug efflux pump. *ACS infectious diseases* **2017,** *3* (1), 89-98.
220. Lomovskaya, O.; Warren, M. S.; Lee, A.; Galazzo, J.; Fronko, R.; Lee, M.; Blais, J.; Cho, D.; Chamberland, S.; Renau, T.; Leger, R.; Hecker, S.; Watkins, W.; Hoshino, K.; Ishida, H.; Lee, V. J., Identification and characterization of inhibitors of multidrug resistance efflux pumps in *Pseudomonas aeruginosa*: novel agents for combination therapy. *Antimicrob Agents Chemother* **2001,** *45* (1), 105-16.
221. Davies, J.; Davies, D., Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews* **2010,** *74* (3), 417-433.
222. Babic, M.; Hujer, A. M.; Bonomo, R. A., What's new in antibiotic resistance? Focus on beta-lactamases. *Drug Resistance Updates* **2006,** *9* (3), 142-156.
223. Bush, K.; Jacoby, G. A.; Medeiros, A. A., A functional classification scheme for β-lactamases and its correlation with molecular structure. *Antimicrobial Agents and Chemotherapy* **1995,** *39* (6), 1211-1233.
224. Walsh, T. R.; Toleman, M. A.; Poirel, L.; Nordmann, P., Metallo-β-lactamases: The quiet before the storm? *Clinical Microbiology Reviews* **2005,** *18* (2), 306-325.
225. Ibuka, A.; Taguchi, A.; Ishiguro, M.; Fushinobu, S.; Ishii, Y.; Kamitori, S.; Okuyama, K.; Yamaguchi, K.; Konno, M.; Matsuzawa, H., Crystal structure of the E166A mutant of

extended-spectrum β-lactamase Toho-1 at 1.8 Å resolution. *Journal of Molecular Biology* **1999,** *285* (5), 2079-2087.
226. Ibuka, A. S.; Ishii, Y.; Galleni, M.; Ishiguro, M.; Yamaguchi, K.; Frère, J. M.; Matsuzawa, H.; Sakai, H., Crystal structure of extended-spectrum β-lactamase Toho-1: Insights into the molecular mechanism for catalytic reaction and substrate specificity expansion. *Biochemistry* **2003,** *42* (36), 10634-10643.
227. Shimamura, T.; Ibuka, A.; Fushinobu, S.; Wakagi, T.; Ishiguro, M.; Ishii, Y.; Matsuzaw, H., Acyl-intermediate structures of the extended-spectrum class A β-lactamase, Toho-1, in complex with cefotaxime, cephalothin, and benzylpenicillin. *Journal of Biological Chemistry* **2002,** *277* (48), 46601-46608.
228. Tomanicek, S. J.; Standaert, R. F.; Weiss, K. L.; Ostermann, A.; Schrader, T. E.; Ng, J. D.; Coates, L., Neutron and X-ray crystal structures of a perdeuterated enzyme inhibitor complex reveal the catalytic proton network of the Toho-1 β-lactamase for the acylation reaction. *Journal of Biological Chemistry* **2013,** *288* (7), 4715-4722.
229. Vandavasi, V. G.; Langan, P. S.; Weiss, K. L.; Parks, J. M.; Cooper, J. B.; Ginell, S. L.; Coates, L., Active-site protonation states in an acyl-enzyme intermediate of a class A β-lactamase with a monobactam substrate. *Antimicrobial Agents and Chemotherapy* **2017,** *61* (1).
230. Vandavasi, V. G.; Weiss, K. L.; Cooper, J. B.; Erskine, P. T.; Tomanicek, S. J.; Ostermann, A.; Schrader, T. E.; Ginell, S. L.; Coates, L., Exploring the Mechanism of β-Lactam Ring Protonation in the Class A β-lactamase Acylation Mechanism Using Neutron and X-ray Crystallography. *Journal of Medicinal Chemistry* **2016,** *59* (1), 474-479.
231. York, D. M.; Yang, W.; Lee, H.; Darden, T.; Pedersen, L. G., Toward the Accurate Modeling of DNA: The Importance of Long-Range Electrostatics. *Journal of the American Chemical Society* **1995,** *117* (17), 5001-5002.
232. Vasilevskaya, T.; Thiel, W., Periodic Boundary Conditions in QM/MM Calculations: Implementation and Tests. *Journal of Chemical Theory and Computation* **2016,** *12* (8), 3561-3570.
233. Lewandowski, E. M.; Lethbridge, K. G.; Sanishvili, R.; Skiba, J.; Kowalski, K.; Chen, Y., Mechanisms of proton relay and product release by Class A β-lactamase at ultrahigh resolution. *FEBS Journal* **2018,** *285* (1), 87-100.
234. Nichols, D. A.; Hargis, J. C.; Sanishvili, R.; Jaishankar, P.; Defrees, K.; Smith, E. W.; Wang, K. K.; Prati, F.; Renslo, A. R.; Woodcock, H. L.; Chen, Y., Ligand-Induced Proton Transfer and Low-Barrier Hydrogen Bond Revealed by X-ray Crystallography. *Journal of the American Chemical Society* **2015,** *137* (25), 8086-8095.
235. Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; Legrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A., *AMBER 14* **2014**.
236. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general Amber force field. *Journal of Computational Chemistry* **2004,** *25* (9), 1157-1174.
237. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **2002,** *23* (16), 1623-1641.

238. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *Journal of Computational Chemistry* **2000,** *21* (2), 132-146.
239. Walker, R. C.; Crowley, I. F.; Case, D. A., The implementation of a fast and accurate QM/MM potential method in Amber. *Journal of Computational Chemistry* **2008,** *29* (7), 1019-1031.
240. Tomanicek, S. J.; Wang, K. K.; Weiss, K. L.; Blakeley, M. P.; Cooper, J.; Chen, Y.; Coates, L., The active site protonation states of perdeuterated Toho-1 β-lactamase determined by neutron diffraction support a role for Glu166 as the general base in acylation. *FEBS Letters* **2011,** *585* (2), 364-368.
241. Langan, P. S.; Vandavasi, V. G.; Cooper, S. J.; Weiss, K. L.; Ginell, S. L.; Parks, J. M.; Coates, L., Substrate Binding Induces Conformational Changes in a Class A β-lactamase That Prime It for Catalysis. *ACS Catalysis* **2018,** *8* (3), 2428-2437.
242. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015,** *11* (8), 3696-3713.
243. Miyamoto, S.; Kollman, P. A., Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **1992,** *13* (8), 952-962.
244. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993,** *98* (12), 10089-10092.
245. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **1995,** *103* (19), 8577-8593.
246. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts* **2008,** *120* (1-3), 215-241.
247. Brás, N. F.; Perez, M. A. S.; Fernandes, P. A.; Silva, P. J.; Ramos, M. J., Accuracy of density functionals in the prediction of electronic proton affinities of amino acid side chains. *Journal of Chemical Theory and Computation* **2011,** *7* (12), 3898-3908.
248. Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M., Enzyme mechanisms with hybrid quantum and molecular mechanical potentials. I. Theoretical considerations. *International Journal of Quantum Chemistry* **1996,** *60* (6), 1189-1200.
249. Stewart, J. J. P., Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *Journal of Molecular Modeling* **2007,** *13* (12), 1173-1213.
250. Amorimmadeira, P. J.; Vaz, P. D.; Bettencourtdasilva, R. J. N.; Florêncio, M. H., Can semi-empirical calculations help solve mass spectrometry problems? Protonation sites and proton affinities of amino acids. *ChemPlusChem* **2013,** *78* (9), 1149-1156.
251. Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **1992,** *13* (8), 1011-1021.
252. Langan, P. S.; Vandavasi, V. G.; Weiss, K. L.; Cooper, J. B.; Ginell, S. L.; Coates, L., The structure of Toho1 β-lactamase in complex with penicillin reveals the role of Tyr105 in substrate recognition. *FEBS Open Bio* **2016,** *6* (12), 1170-1177.
253. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J., Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput Biol* **2017,** *13* (1), e1005324.

254. Xu, J., Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A* **2019,** *116* (34), 16856-16865.
255. Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D., Improved protein structure prediction using potentials from deep learning. *Nature* **2020,** *577* (7792), 706-710.
256. Greener, J. G.; Kandathil, S. M.; Jones, D. T., Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun* **2019,** *10* (1), 3977.
257. Landrum, G. *RDKit: Open-source cheminformatics*.

VITA

Connor Cooper was born in Derry, NH. He graduated from Raymond High School in 2011 and continued his education at the University of New England (UNE) in Biddeford, ME, where he pursued Bachelor of Science degree in Biochemistry and Neuroscience. At UNE's Westbrook College of Health Professions he obtained research experience in a motion analysis laboratory working with Dr. Katherine Rudolph. He also spent the majority of his undergraduate career working with Dr. John Stubbs in the Department of Chemistry and Physics at UNE where he conducted computational chemistry research. This work involved the use of Monte Carlo simulations to characterize the effect of dangling end length on DNA microarray hybridization. In 2015 Connor entered the Genome Science & Technology program at the University of Tennessee working with Dr. Jerry Parks at the UT/ORNL Center for Molecular Biophysics. His interdisciplinary graduate work involves using various computational methods often coupled with experimental collaborations to answer a variety of biological questions ranging from antibiotic resistance to protein structure prediction, to catalytic pulping of wood. During his time at the University of Tennessee he received a Program for Excellence and Equity in Research Graduate Fellowship from the University of Tennessee as well as an NSF Graduate Research Fellowship and contributed to several scientific publications.