Machine Learning-based Prediction of Enzyme Substrate Scope:

Application to Bacterial Nitrilases

Running Title: Machine learning for predicting enzyme substrates

Zhongyu Mou,[†,#] Jason Eakes,[†,#] Connor J. Cooper,[‡] Carmen M. Foster,[†] Robert F. Standaert,[†] Mircea Podar,[†] Mitchel J. Doktycz[†,‡] and Jerry M. Parks[†,‡,*]

[†] Biosciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831-6309, United States

[‡] Graduate School of Genome Science and Technology, University of Tennessee, F225 Walters Life Science, Knoxville, Tennessee 37996, United States

[#] These authors contributed equally to this work.

* Corresponding author: UT/ORNL Center for Molecular Biophysics, Biosciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831-6309. Phone: (865) 574-9259. Email: parksjm@ornl.gov

Acknowledgement

This work was supported by Laboratory-Directed Research and Development funds from Oak Ridge National Laboratory (ORNL), which is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This work used resources of the Compute and Data Environment for Science (CADES) at ORNL. CJC was supported by a National Science Foundation Graduate Research Fellowship under Grant No. 2017219379.

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, although random forest offers some advantages. This approach is intended to be highly modular with respect to physicochemical property calculations and software used for structural modeling and docking.

Keywords

Functional annotation, substrate scope, enzyme specificity, machine learning, modular approach

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[1] 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.[2]

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.[3]

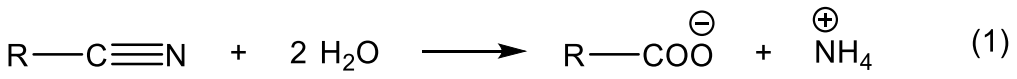
A combination of molecular docking of putative substrates to an available X-ray crystal structure, QM calculations of substrate reactivity, and experimental enzyme activity assays predicted substrate specificity of an enoyl-acyl carrier protein reductase (FabI).[4] In the absence of a crystal structure, homology modeling can be used in the context of ligand docking.[5–7] 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.[8] 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.[9]

Machine learning (ML) is widely applicable to a variety of problems from fields such as quantum mechanics, physical chemistry, biophysics, and physiology.[10] 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.[11] 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.[12] 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.[13] 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%.[14] 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.[15]

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.[16] 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.[17] In addition, nitriles are desirable for their use in efficient chemo- and enantioselective synthesis of carboxylic acids, making them attractive for drug design.[18,19] Typically, nitrilases are classified into three categories according to their substrate specificities: aliphatic, arylaceto-, and aromatic nitrilases.[17,20] 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.



(1)

Various nitrilase activity assays have been described and are based on either fluorogenic or chromogenic substrates or pH indicator methods.[21–23] Recently, a chromogenic method was developed as a convenient means to screen recombinantly produced nitrilases in crude cell extracts.[24] 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.[18] 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.[18] 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.[25] 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.[26]

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.

Materials and Methods

Chemicals and Reagents

All reagents and chemicals were purchased from Sigma Aldrich (St. Louis, MO), Bio-Rad (Hercules, CA), Pierce (Rockford, IL), ThermoFisher Scientific (Pittsburgh, PA), and New England Biolabs (NEB; Ispwich, MA).

Strains and Plasmids

The gene sequences (see Supporting Information) encoding each of the chosen nitrilases were codon-optimized for expression in *E. coli* and were cloned into the pet22(b) vector at the Nde-Sal site by GenScript (Piscataway, NJ; https://www.genscript.com, **Figure S10**). The resulting protein contained a C-terminal 6x histidine tag for protein detection. The plasmids were transformed into BL21(DE3) *E. coli* cells for crude extract preparation.

Crude Extract Preparation

To prepare crude extracts of the recombinant enzymes, nitrilase gene sequences were transformed and expressed using *E. coli* BL21(DE3) host cells as described previously.[24,27] Briefly, a 15-mL overnight culture of BL21(DE3) cells, with and without plasmids, grown in LB medium, was seeded into 250 mL baffled flasks containing 150 mL of 2x YPTG medium (yeast extract, 10 g/L; KH_2PO_4, 3 g/L; K_2HPO_4, 7 g/L; NaCL, 5 g/L; tryptone 16 g/L; and glucose 18 g/L) inoculated with carbenicillin (100 mg mL^{-1}). The 2x YPTG cultures were grown at 30°C and 250 RPM to an optical density (OD_{600}) = 0.6. Cells were induced with isopropyl-β-D-1-thiogalactopyranoside (IPTG, 1 M) to a concentration of 1 mM to express nitrilases. Induced cells were subsequently grown at 30°C and 250 RPM to an optical density of (OD_{600}) = 2.0. The cells were harvested by centrifugation (10 min, 4°C, 5000 RPM) and washed with potassium phosphate buffer (pH 7.2). Washed cells were centrifuged (15 min, 4°C, 6000 RPM), resuspended in potassium phosphate buffer (800 μL g^{-1} wet cell mass), and lysed by sonication (12 x {10 s on and 10 s off}, 50% amplitude with ice water cooling). The resulting cell-free extracts (CFEs) were cleared by centrifugation (10 min, 4°C, 12,000 x g), flash-frozen, and stored at -80°C. Successful nitrilase expression was confirmed by Western blot using His-tag antibodies. Total protein content per extract was estimated using a Bradford Assay.

Nitrilase Substrates

We generated a library of 20 commercially available nitriles consisting of 5 aliphatic, 9 aromatic, and 6 arylaliphatic nitriles (**Figure S1**). Nitriles were prepared as stock solutions in DMSO at a concentration of 100 mM and stored in the dark at room temperature.

Nitrilase Assays

Substrate-dependent enzyme activity was determined using a variant of the colorimetric assay optimized for crude extracts described previously.[24] This assay detects the ammonia produced from the nitrilase-catalyzed hydrolysis reaction using *o*-phthalaldehyde (OPA) under acidic conditions. OPA was dissolved into methanol (200 mg mL^{-1}) and subsequently diluted (1:100) into sodium tetraborate buffer (15 mM, pH 9.5) and stored until needed. Trichloroacetic acid was prepared to a 10% w v^{-1} solution and stored until needed. Nitrilase activity assays were prepared by diluting crude extracts into potassium phosphate buffer (1:2, 10 μL final volume) in 384-well microplates (Corning 3702) and adding the desired nitrile substrate (1 μL) into each well. Typically, each 384-well microplate contained two different crude extract preparations at two dilutions (50% and 25%, diluted (1:2) into potassium phosphate buffer). Each crude enzyme preparation was tested against a panel of 20 substrates and prepared in triplicate (**Figure S13**). The microplate also contained an array of ammonium chloride concentrations (0–10 mM), in triplicate, for comparison to a standard. In addition, wells containing potassium phosphate buffer were used as a blank. An Opentrons OT-2 liquid handling robot (Brooklyn, NY) was used to assist with plate filling. Details of the automated liquid handling procedure are described in the Supplemental Materials. After filling, the microplates were triple wrapped with parafilm and incubated overnight in the dark (18 hours, RT). Next, the Opentrons OT-2 liquid handling robot was used to add OPA reagent (36 μL, dissolved in DMSO (1:1.4)), and

trichloroacetic acid (TCA; (7.5 μL) to acidify the reaction for color development. In addition, DMSO (40.5 μL) was added to maintain solubility of OPA. Microplates were sealed with aluminum sealing tape and shaken (12 min, 40°C, 1500 RPM) to ensure solubility of the chromophores. Absorbance at 675 nm was measured in a Perkin Elmer 2300 spectrophotometer.

Nitrilase 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[28] in Geneious v9[29]. Nitrilase sequences from plants were also included as an outgroup. A phylogenetic tree was constructed using FastTree v. 2.1.12.[30]

Nitrilase 3D modeling

The amino acid sequences of 12 target nitrilases were aligned with Clustal Omega (**Figure S11**).[31,32] The GREMLIN web server[33,34] was used to search the UniProt20 database[35] 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.[36] 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[37] 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[38] 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[39] was greater than 50 $Å^3$.

Docking and docking descriptors

Three-dimensional structures of each nitrile were obtained from the ZINC database.[40] 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.[41–43] All quantum mechanical (QM) calculations were performed with Gaussian 16, revision A.03.[44] Restrained electrostatic potential (RESP) charges[45–47] 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.[48,49] The REF2015 score function[50] 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.[51] QM descriptors included atomic partial charges computed from natural population analysis[52] and Merz-Singh-Kollman

(MK) charges[53,54] on 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[55] was used to calculate active-site descriptors including thermodynamic indices of the folded and extended protein state, topographic indices, physicochemical and structural composition indices. A list of descriptors and definitions is provided in the Supporting Information.

Machine learning and statistical analysis

The *scikit-learn* package (version 0.22) was used to perform the binary classification analysis using four machine learning (ML) methods[56] including two decision tree-based ensemble methods: random forest (RF)[57] and gradient boosted decision trees (GBDT)[58], as well as a kernel-based method, support vector machines (SVM),[59] and logistic regression (LR). For this analysis, experimentally measured activities of < 2 mM ammonia (Figures S2) were considered inactive and descriptors with high correlation to other descriptors (≥ 0.9) were removed (**Figure S12**).

A GitHub repository with a jupyter notebook containing all code required to reproduce our analysis is available at https://github.com/ZhongyuMou/ML-substrate-scope. The jupyter notebook also provides additional information about the ML methods and analysis, including data pretreatment, oversampling, hyperparameter tuning, and the training and validation of the ML models. All statistical analyses and plotting were performed with Python 3.7,[60] Pandas,[61] Numpy,[62] and Matplotlib,[63] respectively.

Results

Strategy. 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 (Eq. 1). To create an effective training set, we selected a set of 12 nitrilase sequences (**Figure 1**) 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.[18] 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),[36] which is a member of sequence clade 1B (**Figure 1**). 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 (PDB entry 3WUY).[36] These 11 sequences (**Table S1**) have varying degrees of sequence identity to Nit6803 and range from 32-71% (**Table S2**) with a sequence coverage of at least 81%. We used the structure of Nit6803 as a template to generate homology models of each enzyme. We generated a selected set of 20 nitriles (**Figure S1**) from among these substrate categories based on previous data sets[18,24] and docked them to each enzyme model and also to the Nit6803 crystal structure (**Figure 2**). 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. These enzyme-containing extracts were added to solutions containing a selected nitrile and enzymatic activity was measured using a semi-quantitative colorimetric assay (**Figure 3**) optimized for crude extracts based on a previously described method.[24]

All 12 enzymes were active toward at least one nitrile (**Figure 3)**. 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 S3**). 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 (**Figure 4**). 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, **Figure S4**) 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)[64] 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 (**Figure S4**), oversampling did not significantly improve model performance (**Figure S5**). All four ML methods perform similarly as evaluated by performing tenfold cross-validation (**Figure 4A**). The average areas under the ROC curve (ROC_AUC) were all ~0.90 and the models had average accuracies of 79-83%. These findings are consistent with the previous observation that featurization, i.e., encoding of physically meaningful molecular information, can be more important than the choice of a particular machine learning algorithm.[10] 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 4B**). 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 4C**). 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 (**Figure S6**). 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 (**Figure 1**) 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 (**Figure S7**).

Discussion

We have developed an approach for predicting substrate scope for enzymes by combining structural modeling, docking, physicochemical properties and various machine learning methods. 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. The complete workflow is shown in **Figure 5**. 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 3**). 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 drastic 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,[14] in which the training data are augmented iteratively to optimize model performance, are expected to be particularly useful.

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,[65] 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 (**Figure S8**). 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.[66] 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.[67] 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.

Although 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, RF has slightly better sensitivity and precision. In addition, unlike kernel-based methods, decision tree-based methods provide variable importance as a useful output that aids in the interpretability of the models. Thus, we recommend RF as a robust, generally applicable approach for future studies.

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 (**Table S3).** Descriptors from all four categories were present in the top 10 most important descriptors over the 20 runs (**Figure 6, Table S3** and **Figure S9**). Thus, including complementary information from each category indeed contributes to the predictive value of the model. Among the most important features (**Table S3**) are the attractive Lennard-Jones term from protein-ligand docking (*fa_atr*) and features describing the hydrophilicity of both the ligand and binding pocket including the amphiphilic moment (*vsurf_A*) and the water-

accessible surface area of all polar atoms (*ASA_P*) for ligands. Intuitively, these features capture key aspects of protein-ligand binding interactions and the hydrophilicity/hydrophobicity of both the ligand and the active site. However, the QM descriptors included here appear infrequently in the top 10, suggesting that descriptors intended to account for electronic properties and 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 are all between 0.25 and 0.30. 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 6**). 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,[68–70] MODELLER,[71–74] SWISS-MODEL,[75] PHYRE2,[76] and others. Similarly, ligand docking could be performed with software such as GOLD,[77] Glide,[78] DOCK,[79] AutoDock Vina,[80] and many others. Alternatives for calculating physicochemical descriptors include Rcpi,[81] Dragon,[82] PaDEL,[83] Mordred[84] 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,[85] 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.

Accession Codes

The NCBI accession numbers for the nitrilase sequences used in this study are: AAR97463 (3A2), AAR97509 (2A6), AAR97476 (1B15), AAR97423 (1B16), AAR97447 (1A27), AAR97388 (1A17), AAR97501 (1A8), AAR97472 (1A2), AAR97500 (1A1), EJM49671 (PMI26_00172), EJM57398 (PMI28_02655).

Associated Content

The Supporting Information is available free of charge on the Wiley Publications website at DOI: XXX

> Nitrilase models, optimized geometries of nitriles, all data files, Rosetta inputs, and scripts for operating the liquid handling robot are provided in a GitHub repository at https://github.com/ZhongyuMou/ML-substrate-scope.

ORCID

Zhongyu Mou: 0000-0003-2240-3129

Jason Eakes: 0000-0002-6356-0772

Connor J. Cooper: 0000-0002-5527-9948

Carmen Foster: 0000-0002-0927-9859

Robert F. Standaert: 0000-0002-5684-1322

Mircea Podar: 0000-0003-2776-0205

Mitchel J. Doktycz: 0000-0003-4856-8343

Jerry M. Parks 0000-0002-3103-9333

References

1. Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res*. 2019;47(D1):D351-D360. doi:10.1093/nar/gky1100
2. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res*. 2019;47(D1):D542-D549. doi:10.1093/nar/gky1048
3. Park H, Ovchinnikov S, Kim DE, DiMaio F, Baker D. Protein homology model refinement by large-scale energy optimization. *Proc Natl Acad Sci*. 2018;115(12):3054-3059. doi:10.1073/pnas.1719115115
4. Freund GS, O'Brien TE, Vinson L, et al. Elucidating Substrate Promiscuity within the FabI Enzyme Family. *ACS Chem Biol*. 2017;12(9):2465-2473. doi:10.1021/acschembio.7b00400
5. Combs SA, DeLuca SL, DeLuca SH, et al. Small-molecule ligand docking into comparative models with Rosetta. *Nat Protoc*. 2013;8(7):1277-1298. doi:10.1038/nprot.2013.074
6. Skolnick J, Zhou H, Gao M. Are predicted protein structures of any value for binding site prediction and virtual ligand screening? *Curr Opin Struct Biol*. 2013;23(2):191-197. doi:10.1016/j.sbi.2013.01.009
7. Pierri CL, Parisi G, Porcelli V. Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening. *Biochim Biophys Acta - Proteins Proteomics*. 2010;1804(9):1695-1712. doi:10.1016/j.bbapap.2010.04.008
8. Jacobson MP, Kalyanaraman C, Zhao S, Tian B. Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends Biochem Sci*. 2014;39(8):363-371.

doi:10.1016/j.tibs.2014.05.006

9. Zhao S, Kumar R, Sakai A, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*. 2013;502(7473):698-702. doi:10.1038/nature12576

10. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*. 2018;9(2):513-530. doi:10.1039/C7SC02664A

11. Bedbrook CN, Yang KK, Robinson JE, Mackey ED, Gradinaru V, Arnold FH. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat Methods*. 2019;16(11):1176-1184. doi:10.1038/s41592-019-0583-8

12. Chevrette MG, Aicheler F, Kohlbacher O, Currie CR, Medema MH. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Birol I, ed. *Bioinformatics*. 2017;33(20):3202-3210. doi:10.1093/bioinformatics/btx400

13. Röttig M, Rausch C, Kohlbacher O. Combining Structure and Sequence Information Allows Automated Prediction of Substrate Specificities within Enzyme Families. Ponting CP, ed. *PLoS Comput Biol*. 2010;6(1):e1000636. doi:10.1371/journal.pcbi.1000636

14. Pertusi DA, Moura ME, Jeffryes JG, Prabhu S, Walters Biggs B, Tyo KEJ. Predicting novel substrates for enzymes with minimal experimental effort with active learning. *Metab Eng*. 2017;44:171-181. doi:10.1016/j.ymben.2017.09.016

15. Yang M, Fehl C, Lees K V., et al. Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat Chem Biol*. 2018;14(12):1109-1117. doi:10.1038/s41589-018-0154-9

16. Chhiba-Govindjee VP, van der Westhuyzen CW, Bode ML, Brady D. Bacterial nitrilases and their regulation. *Appl Microbiol Biotechnol*. 2019;103(12):4679-4692. doi:10.1007/s00253-019-09776-1

17. Howden AJM, Preston GM. Nitrilase enzymes and their role in plant-microbe interactions.

Microb Biotechnol. 2009;2(4):441-451. doi:10.1111/j.1751-7915.2009.00111.x

18. Robertson DE, Chaplin JA, DeSantis G, et al. Exploring Nitrilase Sequence Space for Enantioselective Catalysis. *Appl Environ Microbiol*. 2004;70(4):2429-2436. doi:10.1128/AEM.70.4.2429-2436.2004
19. Kaul P, Banerjee A, Banerjee UC. Nitrile Hydrolases. In: *Industrial Enzymes*. Dordrecht: Springer Netherlands; 2007:531-547. doi:10.1007/1-4020-5377-0_30
20. Kobayashi M, Shimizu S. Versatile nitrilases: Nitrile-hydrolysing enzymes. *FEMS Microbiol Lett*. 1994;120(3):217-223. doi:10.1111/j.1574-6968.1994.tb07036.x
21. Banerjee A, Kaul P, Sharma R, Banerjee UC. A High-Throughput Amenable Colorimetric Assay for Enantioselective Screening of Nitrilase-Producing Microorganisms Using pH Sensitive Indicators. *J Biomol Screen*. 2003;8(5):559-565. doi:10.1177/1087057103256910
22. He YC, Ma CL, Xu JH, Zhou L. A high-throughput screening strategy for nitrile-hydrolyzing enzymes based on ferric hydroxamate spectrophotometry. *Appl Microbiol Biotechnol*. 2011;89(3):817-823. doi:10.1007/s00253-010-2977-5
23. Santoshkumar M, Nayak AS, Anjaneya O, Karegoudar TB. A plate method for screening of bacteria capable of degrading aliphatic nitriles. *J Ind Microbiol Biotechnol*. 2010;37(1):111-115. doi:10.1007/s10295-009-0663-3
24. Black GW, Brown NL, Perry JJB, Randall PD, Turnbull G, Zhang M. A high-throughput screening method for determining the substrate scope of nitrilases. *Chem Commun*. 2015;51(13):2660-2662. doi:10.1039/C4CC06021K
25. Podar M, Eads JR, Richardson TH. Evolution of a microbial nitrilase gene family: A comparative and environmental genomics study. *BMC Evol Biol*. 2005;5:1-13. doi:10.1186/1471-2148-5-42
26. Timm CM, Campbell AG, Utturkar SM, et al. Metabolic functions of Pseudomonas fluorescens strains from Populus deltoides depend on rhizosphere or endosphere isolation compartment.

Front Microbiol. 2015;6:1118. doi:10.3389/fmicb.2015.01118

27. Garcia DC, Mohr BP, Dovgan JT, Hurst GB, Standaert RF, Doktycz MJ. Elucidating the potential of crude cell extracts for producing pyruvate from glucose. *Synth Biol*. 2018;3(1):ysy006. doi:10.1093/synbio/ysy006
28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340
29. Kearse M, Moir R, Wilson A, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28(12):1647-1649. doi:10.1093/bioinformatics/bts199
30. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol Evol*. 2009;26(7):1641-1650. doi:10.1093/molbev/msp077
31. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7(1):539. doi:10.1038/msb.2011.75
32. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci*. 2018;27(1):135-145. doi:10.1002/pro.3290
33. Balakrishnan S, Kamisetty H, Carbonell JG, Lee S-I, Langmead CJ. Learning generative models for protein fold families. *Proteins Struct Funct Bioinforma*. 2011;79(4):1061-1078. doi:10.1002/prot.22934
34. 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*. 2013;110(39):15674-15679. doi:10.1073/pnas.1314045110
35. Bateman A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019;47(D1):D506-D515. doi:10.1093/nar/gky1049

36. Zhang L, Yin B, Wang C, et al. Structural insights into enzymatic activity and substrate specificity determination by a single amino acid in nitrilase from Synechocystis sp. PCC6803. *J Struct Biol*. 2014;188(2):93-101. doi:10.1016/j.jsb.2014.10.003

37. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science*. 2017;355(6322):294-298. doi:10.1126/science.aah4043

38. Song Y, DiMaio F, Wang RY-R, et al. High-Resolution Comparative Modeling with RosettaCM. *Structure*. 2013;21(10):1735-1742. doi:10.1016/j.str.2013.08.005

39. Durrant JD, Votapka L, Sørensen J, Amaro RE. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput*. 2014;10(11):5047-5056. doi:10.1021/ct500381c

40. Sterling T, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. *J Chem Inf Model*. 2015;55(11):2324-2337. doi:10.1021/acs.jcim.5b00559

41. Stephens PJ, Devlin FJ, Chabalowski CF, Frisch MJ. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J Phys Chem*. 1994;98(45):11623-11627. doi:10.1021/j100096a001

42. Petersson GA, Bennett A, Tensfeldt TG, Al-Laham MA, Shirley WA, Mantzaris J. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *J Chem Phys*. 1988;89(4):2193-2218. doi:10.1063/1.455064

43. Petersson GA, Al-Laham MA. A complete basis set model chemistry. II. Open-shell systems and the total energies of the first-row atoms. *J Chem Phys*. 1991;94(9):6081-6090. doi:10.1063/1.460447

44. Frisch MJ, Trucks GW, Schlegel HB, et al. Gaussian 16 Revision 16.A.03 Inc., Wallingford CT. Inc., Wallingford CT.

45. Bayly CI, Cieplak P, Cornell WD, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem*.

1993;97(40):10269-10280. doi:10.1021/j100142a004

46. Cornell WD, Cieplak P, Bayly CI, Kollman PA. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J Am Chem Soc*. 1993;115(21):9620-9631. doi:10.1021/ja00074a030

47. Cieplak P, Cornell WD, Bayly C, Kollman PA. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J Comput Chem*. 1995;16(11):1357-1377. doi:10.1002/jcc.540161106

48. Meiler J, Baker D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. *Proteins Struct Funct Bioinforma*. 2006;65(3):538-548. doi:10.1002/prot.21086

49. Davis IW, Baker D. RosettaLigand Docking with Full Ligand and Receptor Flexibility. *J Mol Biol*. 2009;385(2):381-392. doi:10.1016/j.jmb.2008.11.010

50. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput*. 2017;13(6):3031-3048. doi:10.1021/acs.jctc.7b00125

51. *Molecular Operating Environment (MOE)*. (Chemical Computing Group Inc., 2015).; 2016.

52. Reed AE, Weinstock RB, Weinhold F. Natural population analysis. *J Chem Phys*. 1985;83(2):735-746. doi:10.1063/1.449486

53. Besler BH, Merz KM, Kollman PA. Atomic charges derived from semiempirical methods. *J Comput Chem*. 1990;11(4):431-439. doi:10.1002/jcc.540110404

54. Singh UC, Kollman PA. An approach to computing electrostatic charges for molecules. *J Comput Chem*. 1984;5(2):129-145. doi:10.1002/jcc.540050204

55. Ruiz-Blanco YB, 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. doi:10.1186/s12859-015-0586-0

56. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach*

Learn Res. 2012;12(10):2825-2830.

57. Breiman L. *Machine Learning, 45(1), 5–32.*; 2001. doi:10.1023/A:1010933404324

58. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

59. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

60. Van Rossum, G.; Drake FL. Python 3 Reference Manual. *Scotts Val CA*. 2009;CreateSpac.

61. McKinney W. Data Structures for Statistical Computing in Python. *Proc 9th Python Sci Conf*. 2010;1697900(Scipy):51-56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

62. Oliphant TE. Guide to NumPy. *Methods*. 2010;1:378. doi:10.1016/j.jmoldx.2015.02.001

63. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng*. 2007;9(3):99-104. doi:10.1109/MCSE.2007.55

64. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953

65. Fernandes BCM, Mateo C, Kiziak C, et al. Nitrile Hydratase Activity of a Recombinant Nitrilase. *Adv Synth Catal*. 2006;348(18):2597-2603. doi:10.1002/adsc.200600269

66. Bertolani SJ, Siegel JB. 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):1-15. doi:10.1371/journal.pone.0214126

67. Hermann JC, Marti-Arbona R, Fedorov AA, et al. Structure-based activity prediction for an enzyme of unknown function. *Nature*. 2007;448(7155):775-779. doi:10.1038/nature05981

68. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 2010;5(4):725-738. doi:10.1038/nprot.2010.5

69. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: Protein structure and function prediction. *Nat Methods*. 2014;12(1):7-8. doi:10.1038/nmeth.3213

70. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43(W1):W174-W181. doi:10.1093/nar/gkv342

71. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinforma*. 2016;(1):5.6.1-5.6.37. doi:10.1002/cpbi.3

72. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu Rev Biophys Biomol Struct*. 2000;29(1):291-325. doi:10.1146/annurev.biophys.29.1.291

73. Šali A. Comparative protein modeling by satisfaction of spatial restraints. *Mol Med Today*. 1995;1(6):270-277. doi:10.1016/S1357-4310(95)91170-7

74. Fiser A, Do RKG, Šali A. Modeling of loops in protein structures. *Protein Sci*. 2000;9(9):1753-1773. doi:10.1110/ps.9.9.1753

75. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303. doi:10.1093/nar/gky427

76. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*. 2015;10(6):845-858. doi:10.1038/nprot.2015.053

77. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997;267(3):727-748. doi:10.1006/jmbi.1996.0897

78. Friesner RA, Murphy RB, Repasky MP, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J Med Chem*. 2006;49(21):6177-6196. doi:10.1021/jm051256o

79. Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK. Ligand Pose and Orientational Sampling in Molecular Docking. Romesberg F, ed. *PLoS One*. 2013;8(10):e75992. doi:10.1371/journal.pone.0075992

80. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new

scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2009;31(2):NA-NA. doi:10.1002/jcc.21334

81. Cao D-S, Xiao N, Xu Q-S, Chen AF. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*. 2015;31(2):279-281. doi:10.1093/bioinformatics/btu624

82. Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: An easy approach to molecular descriptor calculations. *MATCH Commun Math Comput Chem*. 2006;56(2):237-248.

83. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32(7):1466-1474. doi:10.1002/jcc.21707

84. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform*. 2018;10(1):4. doi:10.1186/s13321-018-0258-y

85. Sharma N, Verma R, Savitri, Bhalla TC. Classifying nitrilases as aliphatic and aromatic using machine learning technique. *3 Biotech*. 2018;8(1):1-8. doi:10.1007/s13205-018-1102-9

Figure Legends

Figure 1. Phylogenetic tree of a family of nitrilases that encompass the enzymes used in this study (grey) and characterized in refs 15 and 22 as well as identified in rhizosphere bacteria (ref 23). 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 2. (**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.

Figure 3. 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. A sequence distance tree generated with Clustal Omega[31,32] is shown on the left. Similar activity patterns were observed with cell extracts at 25% dilution (**Figure S2**).

Figure 4. 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 5. Graphical overview of the structure-based approach to predict the substrate scope of enzymes. After target selection, homology 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 6. (**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.

Figures

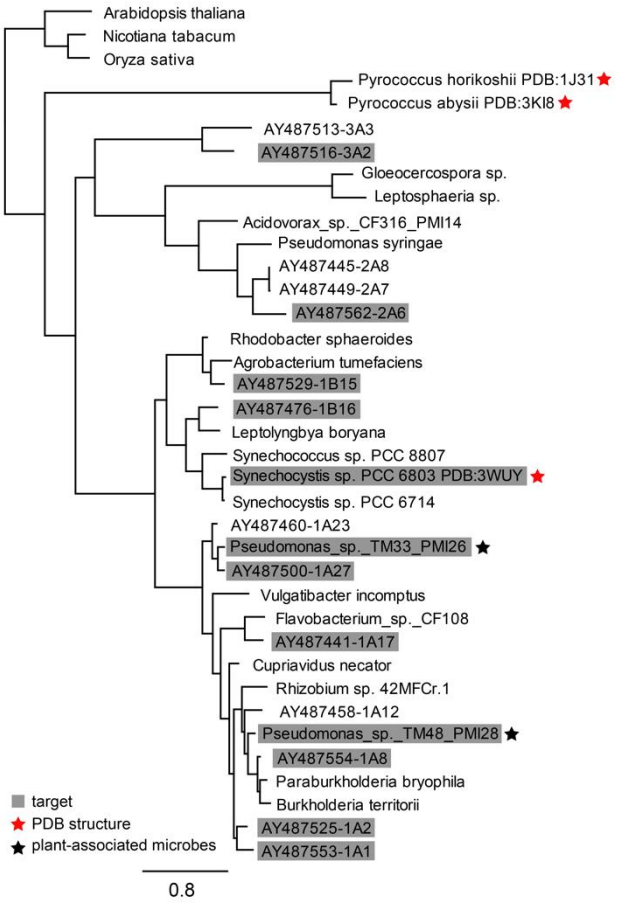


Figure 1

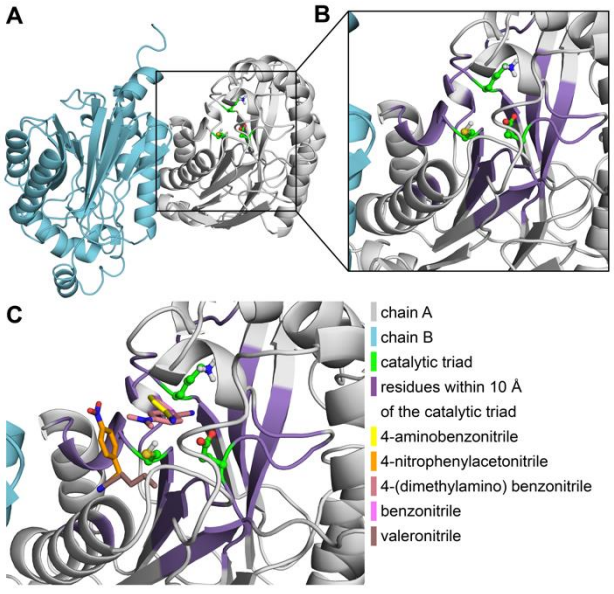
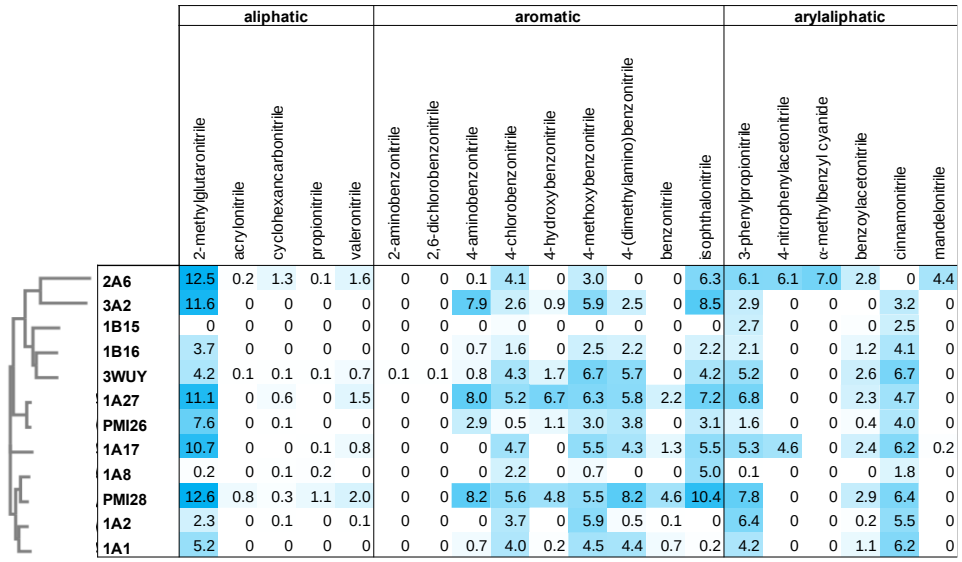


Figure 2



	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 3

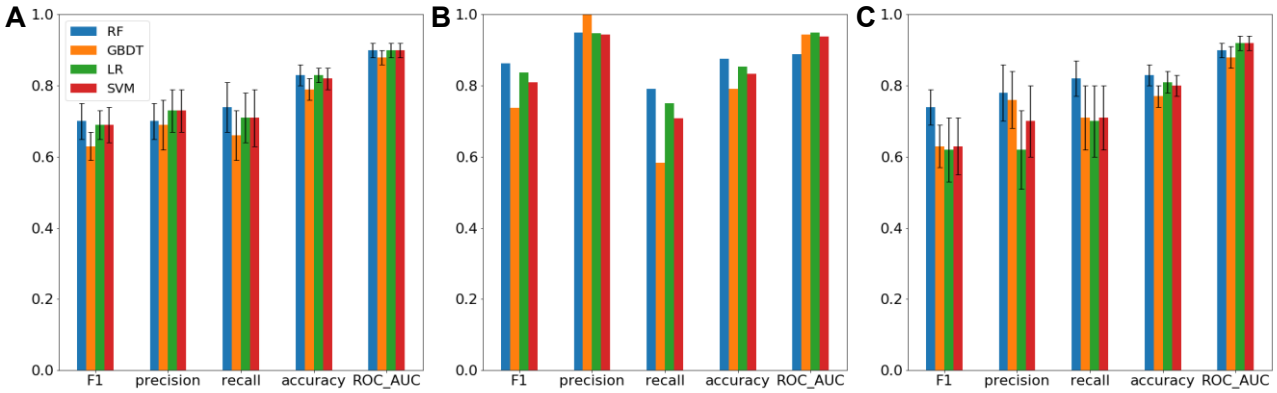


Figure 4

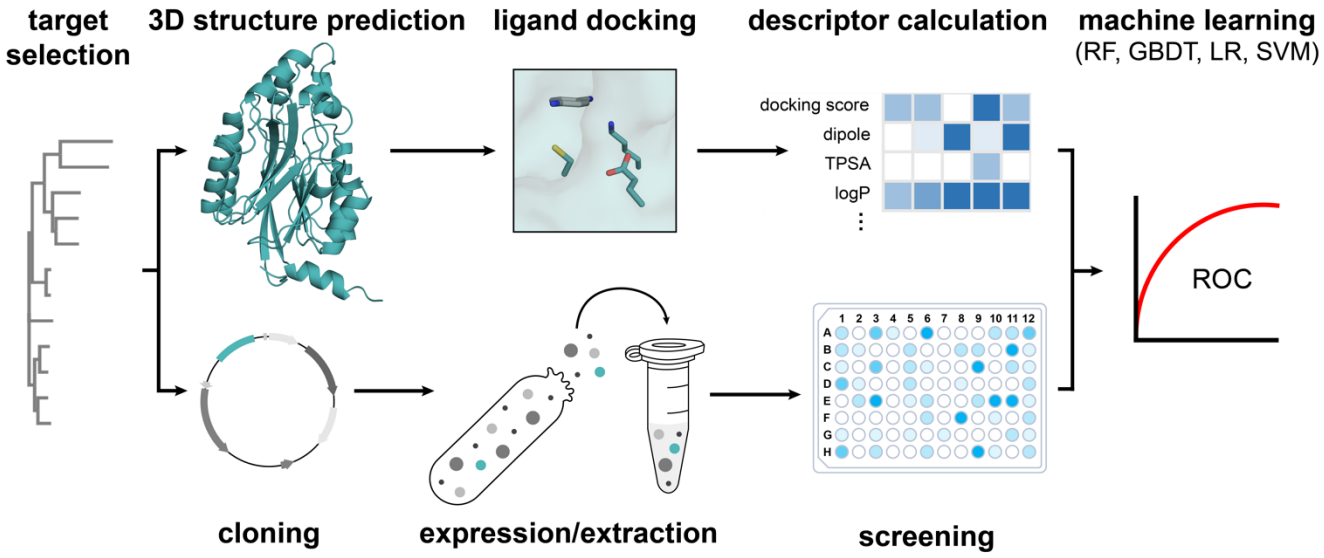


Figure 5

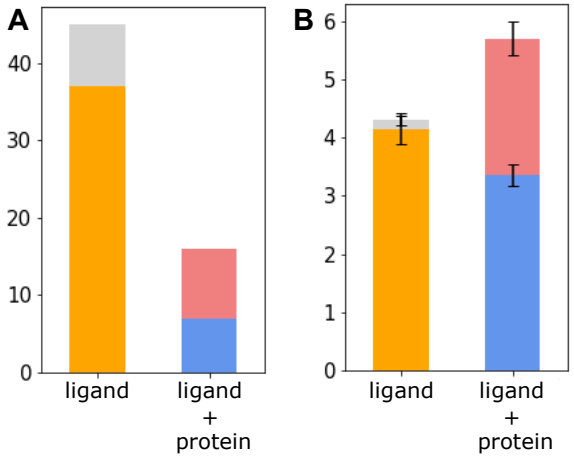


Figure 6