ACS Chemical Biology

Articles

Cite This: *ACS Chem. Biol.* 2019, 14, 2867−2875

pubs.acs.org/acschemicalbiology

Computationally Guided Discovery and Experimental Validation of Indole-3-acetic Acid Synthesis Pathways

David C. Garcia,[†,‡] Xiaolin Cheng,[§] Miriam L. Land,[∥] Robert F. Standaert,[†,⊥] Jennifer L. Morrell-Falvey,[†] and Mitchel J. Doktycz*[,†,‡]

[†] Biological and Nanoscale Systems Group, Biosciences Division Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

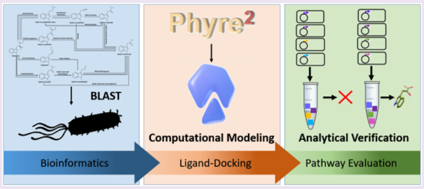
[‡] Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, Tennessee 37996-4519, United States

[§] College of Pharmacy, The Ohio State University, Columbus, Ohio 43210, United States

[∥] Computational Biology and Bioinformatics Group, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

[⊥] Department of Chemistry, East Tennessee State University, Johnson City, Tennessee 37604, United States

Supporting Information



ABSTRACT: Elucidating the interaction networks associated with secondary metabolite production in microorganisms is an ongoing challenge made all the more daunting by the rate at which DNA sequencing technology reveals new genes and potential pathways. Developing the culturing methods, expression conditions, and genetic systems needed for validating pathways in newly discovered microorganisms is often not possible. Therefore, new tools and techniques are needed for defining complex metabolic pathways. Here, we describe an *in vitro* computationally assisted pathway description approach that employs bioinformatic searches of genome databases, protein structural modeling, and protein−ligand-docking simulations to predict the gene products most likely to be involved in a particular secondary metabolite production pathway. This information is then used to direct *in vitro* reconstructions of the pathway and subsequent confirmation of pathway activity using crude enzyme preparations. As a test system, we elucidated the pathway for biosynthesis of indole-3-acetic acid (IAA) in the plant-associated microbe *Pantoea* sp. YR343. This organism is capable of metabolizing tryptophan into the plant phytohormone IAA. BLAST analyses identified a likely three-step pathway involving an amino transferase, an indole pyruvate decarboxylase, and a dehydrogenase. However, multiple candidate enzymes were identified at each step, resulting in a large number of potential pathway reconstructions (32 different enzyme combinations). Our approach shows the effectiveness of crude extracts to rapidly elucidate enzymes leading to functional pathways. Results are compared to affinity purified enzymes for select combinations and found to yield similar relative activities. Further, *in vitro* testing of the pathway reconstructions revealed the "underground" nature of IAA metabolism in *Pantoea* sp. YR343 and the various mechanisms used to produce IAA. Importantly, our experiments illustrate the scalable integration of computational tools and cell-free enzymatic reactions to identify and validate metabolic pathways in a broadly applicable manner.

Metabolism is a complex network of interconnected chemical pathways responsible for an organism's subsistence. In addition to macromolecules, small molecules (metabolites) are produced by cells. Primary metabolites (such as sugars, amino acids, lipids, nucleotides, and cofactors) provide energy, serve as building blocks for macromolecules, or otherwise support core cellular functions. These molecules are common to many organisms, and their metabolism (biosynthesis and subsequent transformations) is relatively well-understood. Secondary metabolites are small molecules made for other purposes, such as signaling and defense. They are highly diverse and produced by a limited number of organisms. While our understanding of primary metabolic networks has come a long way, the pathways by which the vast majority of secondary metabolites are created or utilized are poorly defined. For an overwhelming number of microorganisms, there is little in the way of analytical evidence that a given pathway is either present or active.[1,2] This problem has expanded with the increase in genome-scale sequencing efforts. Genetic information on millions of proteins provide hints as to the existence of metabolic pathways, but robust methods for deciphering and confirming the metabolic potential hidden in this information are lacking.[3] Large collections of genomic data, such as the National Center for Biotechnology Information (NCBI) database, house over 40 M protein sequences, the majority of which remain uncharacterized.[4] Automated annotations, while informative, are often incorrect and impede facile determination of metabolic capabilities.[5,6] The inability to confidently predict the function of an enzyme and its place in a metabolic pathway has become a consistent

Received: September 10, 2019
Accepted: November 6, 2019
Published: November 6, 2019

DOI: 10.1021/acschembio.9b00725

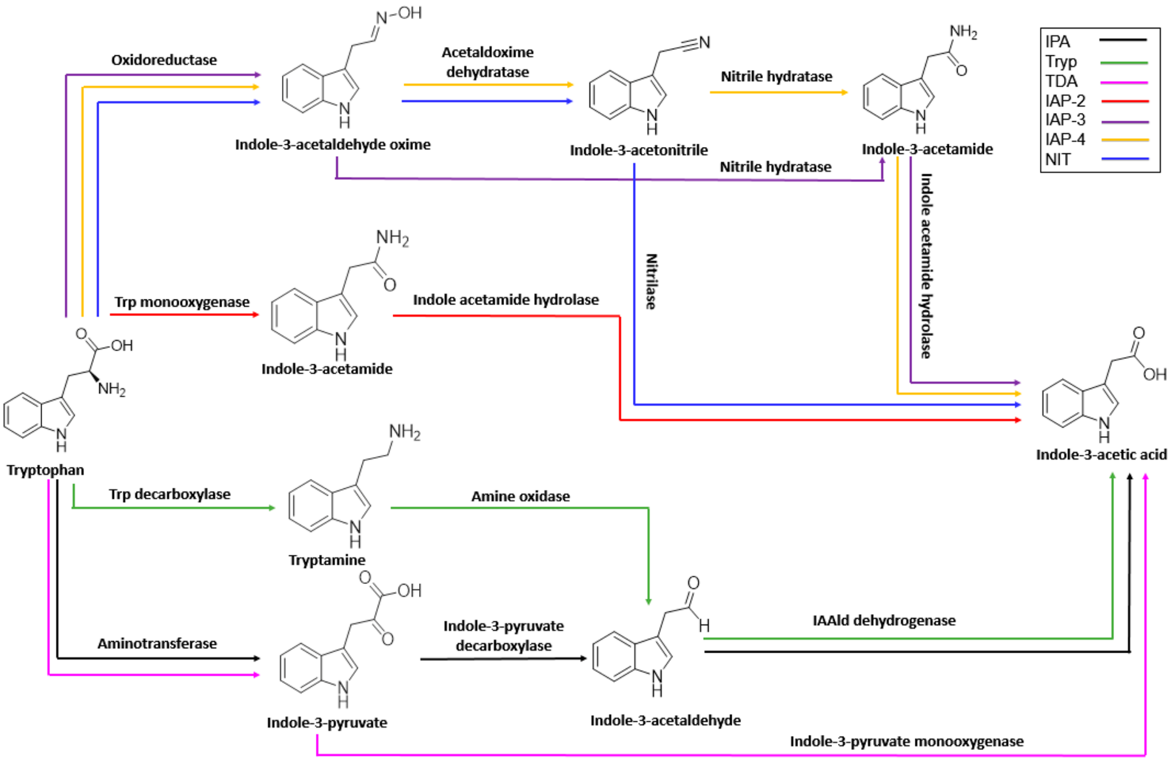


Figure 1. General metabolic model of the conversion of tryptophan to IAA. Enzymes that catalyze the individual steps are written above the colored arrows and names of ligands and products underneath their molecular structure. The seven pathways found to produce IAA from tryptophan are color coded as follows: black, indole pyruvate pathway (IPA); green, tryptamine pathway (Tryp); magenta, tryptophan-dependent auxin biosynthesis pathway (TDA); red, 2-step indole-3-acetamide pathway (IAP-2); purple, 3-step indole-3-acetamide pathway (IAP-3); yellow, four-step indole-3-acetamide pathway (IAP-4); blue, indole-acetaldoxime/indole-3-acetonitrile pathway (NIT). Note: pathways IAP-3 and IAP-4 share the use of nitrile hydratase, and pathways IAP-2, IAP-3, and IAP-4 share the use of indole acetamide hydrolase.

problem in the study and application of metabolic processes. Therefore, new approaches are needed for quickly and accurately elaborating the metabolic capabilities of microorganisms.

Current methods for defining metabolic networks can be slow and labor intensive. Given that methods for cultivation, pathway expression, and genetic manipulation are not always possible, conventional experimental approaches to the study of metabolic processes in newly identified microorganisms can be a challenging endeavor.[7−10] Previous work has sought to remedy this problem by leveraging computational tools to reveal the presence and potential function of a protein.[11] The combined use of bioinformatic software and structural data has resulted in significant strides in predicting enzymatic functions from uncharacterized proteins.[12−14] For example, the addition of complementary data, such as genomic context and ligand-docking analysis, has furthered this pursuit by demonstrating the functionality of multi-input predictions to generate testable hypotheses.[15] Such efforts have continued with recent advancements such as integrative pathway mapping wherein the function of a candidate enzyme and its potential metabolic pathways are predicted by combining information such as ligand docking, chemoinformatic analysis, genomic context, and chemical screening in a single analysis.[16] While these efforts have advanced significantly in recent years, determining the efficacy of these computationally based approaches for defining protein function and pathway connectivity still requires experimental verification, and determination of protein activity remains a major bottleneck for effective gene annotation and pathway analysis.

To facilitate the definition and confirmation of metabolic pathways, we sought to develop a simple stepwise method that initially culls a large subset of proteins related to a pathway, using *in silico* methods, and tests the remaining potential pathways through scalable, *in vitro* biochemical experiments. Our combined computational and empirical approach toward pathway description consists of three steps. First, bioinformatic analyses of the genome or genome database of interest are performed using query enzymes described in the literature in order to find homologous enzymes and identify potentially complete pathways. In addition to removing the overwhelming majority of the genome, this bioinformatic step has the benefit of providing a loosely culled set of enzymes with potential activity based on homology. Second, ligand-docking simulations are performed with protein crystal structures or computationally modeled structures to further cull the listed enzymes to those most likely to interact with their predicted substrates/intermediates. Third, small-scale, heterologous expression and *in vitro* reactions in the crude extract are performed to examine the contribution of individual enzymes to the predicted pathways, thus verifying pathway activity without the need for lengthy purification efforts. The resulting new understanding of the pathway and its component enzymes can then be used to refine gene annotation and make high-quality predictions for the presence of the same pathway in other organisms by clustering potential enzymes with the verified subset.

To test the effectiveness of using computationally guided discovery coupled with experimental validation in crude extracts, we examined the production of the phytohormone

DOI: 10.1021/acschembio.9b00725
ACS Chem. Biol. 2019, 14, 2867−2875

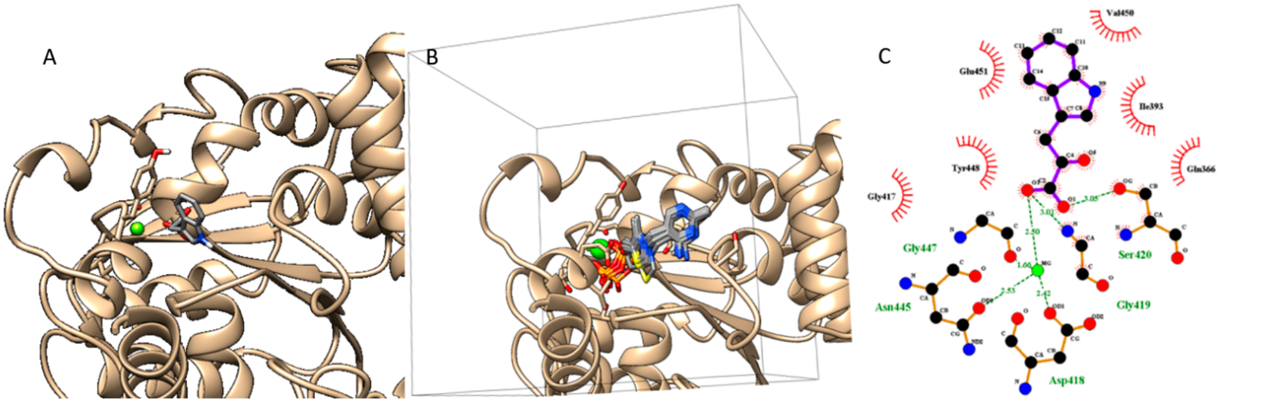


Figure 2. Homology modeling and ligand docking to pmi39_00059 IPDC. (A) Representative example of ligand docked to predicted protein in this case, docked indole-3-pyruvate in IPDC PMI39_00059. (B) Representative example of docking search space (black box) and region predicted to be binding pocket by Phyre2. (C) Schematic representations of docked ligands. Residues involved in hydrogen bonding interactions are shown as green dotted lines with the corresponding donor−acceptor distance shown as a ball and stick model. Residues involved in van der Waals interactions with the ligand are shown with spikes.

indole-3-acetic acid (IAA) from tryptophan in *Pantoea* sp. YR343 (YR343), a root colonizing member of the *Populus deltoides* plant-root microbiome.[17] This plant regulatory metabolite can be the product of a complex set of interconnected metabolic reactions. Notably, *Pantoea* sp. YR343 was shown previously to produce IAA in the presence of tryptophan, but the expected enzymes do not exist in a common operon.[18] This lack of common genomic context coupled with the annotation of many of the potential candidate enzymes confounds pathway determination. Further, as with many products of secondary metabolism, the number of potential pathways, as well as their interconnected nature, complicates understanding of flux through the pathway and its potential control mechanisms. As described below, the combined computational culling and *in vitro* verification approach effectively defines the functional capabilities of the component enzymes and the tryptophan to IAA pathway in *Pantoea* sp. YR343.

RESULTS AND DISCUSSION

Bioinformatic Analyses Indicate Complex IAA Metabolism in *Pantoea* sp. YR343. Plants experience complex interactions with the microbial communities immediately surrounding and within their root systems.[19] The microbes in these systems are composed of bacteria, fungi, and archaea capable of influencing plant behavior through chemical signals.[20] *Pantoea* sp. YR343, a previously sequenced member of the *P. deltoides* plant-root community, was found to produce the plant phytohormone indole-3-acetic acid from tryptophan through an uncharacterized pathway.[17,21] Because some intermediates are shared by these pathways, there are seven interconnected but discrete potential pathways for IAA production from tryptophan (Figure 1).[18,22] Prospective IAA biosynthetic pathways in YR343 were identified following a BLAST search using 11 different enzymes associated with these pathways acquired from bacterial and plant genomes (Table S1). *Pantoea* sp. YR343 contains at least one enzyme homologue for each of the known IAA pathways, but the only complete route from tryptophan to IAA is the IPA pathway, in which tryptophan is first converted to indole-3-pyruvate, then indole-3-acetaldehyde, and finally IAA by an aminotransferase (Am-Trf), an indole pyruvate decarboxylase (IPDC), and an indole acetaldehyde dehydrogenase (IALDh), respectively (Figure 1, Figure S1). An *E*-value of 1 × 10^{-25} was used as a cutoff value as it removed all enzymes from the known non-IAA producer, *Escherichia coli*. Two of the three nodes in *Pantoea* sp. YR343's potential IAA pathway had multiple possible enzymes, specifically one aminotransferase, two IPDCs, and 16 dehydrogenases. Therefore, 32 distinct enzyme combinations could potentially complete a pathway for tryptophan to IAA conversion.

Computational Screening of Ligand-Docking Interactions. To cull the list of potential gene products and resulting enzyme combinations possibly responsible for IAA production in *Pantoea* sp. YR343, ligand docking was used as a means of testing metabolic interactions and removing enzymes that are unlikely to bind their respective substrates/products. Given the unavailability of a crystal structure for any of the 19 enzymes, Phyre2 was used to generate homology models and predict ligand binding sites (Figure 2A).[23] Phyre2's top scoring model and its predicted ligand-binding site were used as the basis for analyzing how the enzyme interacts with its potential substrate/product (Figure 2B,C).

The binding of a ligand in the binding pocket was visually inspected in $n \geq 5$ independent simulations using Vina. A simple binary designation of binding or nonbinding was then used to cull enzymes from being part of the IAA conversion pathway (Figure S2). Phyre2's 3DLigandSite was used to determine the position of the binding pocket based on 3DLigandSite's structural library of homologous peptides with bound ligands (Figure S3).[24] Proteins capable of binding the expected ligands were kept in the pool of potential enzymes while those unable to bind were removed (Table S2). Consequently, as the aminotransferase successfully accommodated tryptophan in its binding pocket, it was retained. Only one of the two IPDCs was capable of binding indole-3-pyruvate, and 9 of the 16 potential IALDh enzymes were predicted to bind indole-3-acetaldehyde. Removing those enzymes that failed the docking step lowered the possible pathways to nine enzyme combinations.

Pathway Verification. Our first set of *in vitro* experiments tested each of the nine predicted IAA enzyme combinations using heterologously expressed enzymes in clarified crude extracts (Am-Trf + IPDC + IALDh). Testing each combination for IAA production revealed that five of the potential enzyme sets produced IAA in relatively large amounts

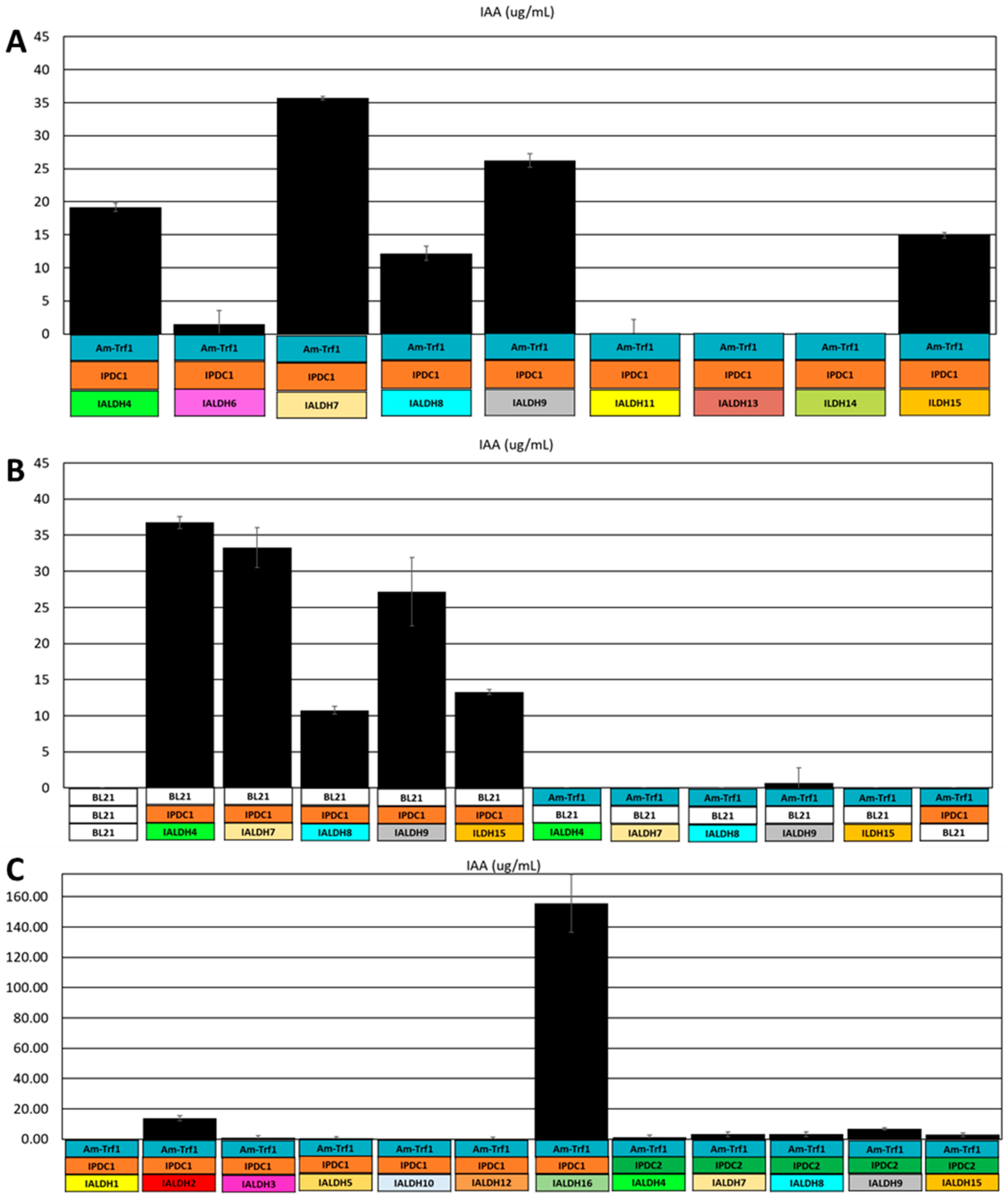


Figure 3. Complete pathways created using single proteins from each enzymatic step to produce IAA. Gene loci are listed in Table S2. (A) Only enzymes found to be bioinformatically related to a verified IAA enzyme and capable of docking their respective ligand were used to measure the initial concentration of IAA. Each reaction included a single aminotransferase, a single IPDC, and a single IALDh. (B) Control reactions were performed to show the importance of individual metabolic steps. BL21 corresponds to an extract replaced with a BL21 Star DE3 extract containing no heterologously expressed protein. (C) Culled enzymes were tested for activity to determine the rate of false negatives from Vina.

(>12 μg mL^{-1}) while two sets produced relatively low amounts (<3 μg mL^{-1}), and two other combinations showed no IAA production (Figure 3A). As expected, the control reactions containing just host extract derived from either BL21 Star(DE3) or BL21(DE3)pLysS did not produce IAA. Each node in the IAA pathway was subsequently tested by replacing it with blank extract from BL21 Star(DE3) (Figure 3B). Control reactions without a heterologously expressed Am-Trf were performed using the most active IALDh enzymes as their production had previously been verified. In all cases, each pathway showed IAA production in the absence of Am-Trf. The IAA levels produced in these control reactions are relatively high and are likely due to the presence of aminotransferase activity in the *E. coli* extract. A similar experiment was performed, but without the presence of the heterologously produced IPDC. In all cases the absence of

IPDC prevented IAA production, showing that IPDC is essential for IAA production in the bacterial extract. Similarly, the presence of a heterologously produced IALDh was essential for IAA production as in its absence no detectable IAA was produced. Of the nine IALDhs that passed the ligand-docking test, five were able to complete the IAA pathway when combined with an Am-Trf and IPDC.

It was important to determine if the *in silico* culling had eliminated viable enzymes. Accordingly, the remaining IPDC and seven of the IALDhs removed from the pool of enzymes during the ligand-docking analysis were similarly cloned, expressed, and tested for a potential role in IAA production. When combined with verified IALDhs, the culled IPDC, IPDC2, showed low IAA production, while two of the seven potential IALDhs were found to produce IAA at levels comparable to those of the nonculled enzymes (Figure 3C). The IALDh reactions showed that ligand docking had a false negative rate of 28.6%, substantially lower than the false positive rate of 44.4%.

In order to determine the advantages and effectiveness of using crude extract preparations for pathway validation, enzymes with positive activity results in crude extracts were affinity purified using hexa-histidine tags. From the potential 10 enzymes, 5 were successfully affinity purified and used for IAA production experiments. Enzymatic combinations were prepared using AMTRF1, IPDC1, IPDC2, IALDH7, and IALDH16 (Figure S4). The purified protein reactions showed the same activity trends as those that employed using crude extract combinations. Specifically, the presence of IPDC1 was shown to be necessary for IAA production as its absence halted IAA production. IALDH16 maintained its relatively high rate of activity as reactions containing IALDH16 produced 5.66 μg mL^{-1} IAA compared to the IALDH7's 2.04 μg mL^{-1}. The aminotransferase (AMTRF1) was able to carry out the expected transamination, though only ~1 μg mL^{-1} IAA was produced (Figure S5).

The plant microbiome consists of a complex network of chemical communications between a host and its microbial colonizers. IAA metabolism is highly relevant to plant-microbe interactions and serves as an important test case in developing a deeper understanding of both cellular metabolism and microbial communities. Even an apparently simple, three-step secondary metabolic pathway can lead to a large number of potential enzyme candidates and combinations that traditional methods cannot easily decipher. This work aims to refine an approach that can define relevant metabolic pathways, thus decreasing the resources necessary to establish secondary metabolic pathways.

An initial model for potential IAA metabolic pathways was created from the literature in order to generate BLAST hits. A relatively small set of enzymes involved in the known IAA pathways generated a large list of potential IAA pathway components in *Pantoea* sp. YR343. However, comparison of potential pathway models and BLAST results provided a valuable culling step as it effectively removed enzymes from the querying pool and left the IPA pathway as the most likely generator of IAA. Further, BLAST analyses were key for IAA pathway description due to the lack of genomic context. Notably, of the nine enzymes that were eventually verified to be active in *Pantoea* sp. YR343's IAA metabolism, only three were found to exist within potential operons, all of which were unrelated to IAA metabolism (Figure S2).

In an effort to further cull the list of potential pathway components, homology modeling and ligand docking were employed. These emerging computational tools have the ability to test a large set of potential candidates by virtue of the candidate's homology to the query set or ability to bind the predicted ligand. This study leverages previous efforts showing that the combined use of computational methods such as BLAST, homology modeling, and ligand docking, while generating false positives and false negatives, is capable of identifying active proteins.[25,26] In this work, we used a readily accessible program, Vina, to dock ligands to the binding pockets of the protein homology models predicted by Phyre2 and 3DLigandSite. This process can be applied easily to a large number of candidate proteins and lends itself to further optimization and large-scale automation. As with many types of predictive analysis, binding affinity calculations are only estimates and may yield errors. Of the 19 proteins that passed the ligand-docking test, 7 were verified as having relevant substrate activity. Our analysis of each protein removed from the pool of enzymes during the ligand-docking step showed that, of eight culled enzymes tested using Am-Trf1 and IPDC1, two of the culled IALDhs were found to produce IAA at levels comparable to those of the enzymes originally selected based on the ligand docking (Figure 3C). Overall, the presence of false negatives and positives is expected in this form of analysis, and our results favor well when compared to similar studies.[27] Notably, less stringent conditions are capable of removing many false negatives by using higher-energy confirmations, but this practice may exacerbate the acquisition of false positives.[28,29] While these accessible structural modeling and ligand-docking tools can reduce search space and simplify downstream *in vitro* experiments, further refinements are needed before confidently employing these tools on more complex pathways.

Future efforts to improve the performance of docking and homology modeling could lean more on an ensemble-based docking strategy akin to the similarity ensemble approach (SEA), in which the binding ability is evaluated by a set of ligands (e.g., ligands similar to the target substrate/product and transition state structures), in order to obtain a more robust prediction of enzyme function.[30] Moreover, improvements can also be made by accounting for protein flexibility and by the use of methods such as molecular dynamics and quantum chemistry calculations that provide more accurate descriptions of the protein–ligand interactions.[31,32] At the moment, Vina's empirical scoring function relies on energetic factors to assign fitness. The use of automated computational tools will be essential for evaluation of complex pathways and for keeping pace with genomic data.

A critical, complementary step in computationally based pathway determination is experimental validation. Traditional methods employing genetic manipulation to verify activity are slow and not practical when analyzing more than one or a few enzymatic steps. Therefore, we employed a method of expressing and testing each of the predicted enzymes in a crude cell extract as a means to increase throughput. Traditional methods for enzyme characterization employ heterologous expression followed by affinity purification. Though affinity purification can lead to more definitive information, such as in terms of understanding reaction kinetics, it suffers due to poor success rates and low throughput.[33−35] Even high-throughput, automated systems result in success rates as low as 20%, thus making affinity purification impractical for keeping up with expanding gene discovery.[36,37] In this work, reactions with purified enzymes were performed to evaluate the efficacy of using crude extracts. However, of the 9 different enzyme isolations attempted, only AMTRF1, IPDC1, IPDC2, IALDH7, and IALDH16 could be successfully purified (Figure S4). With these purified enzymes, activity trends were similar to those observed when using crude extracts of the proteins; the absence of IPDC1 still quenched IAA production, and IALDH16 was still significantly more active than the other purified aldehyde dehydrogenase IALDH7. The time-consuming nature of purification is in contrast to the simplicity and effectiveness of using crude extracts. Here, the enzymes involved in IAA synthesis could be identified, expressed in soluble form, and tested without the need for expansive troubleshooting (Figure S5).

Having defined the IAA pathway components employed by *Pantoea* sp. YR343, insights into the origin and function of the pathway can be gleaned. In the case of the initial aminotransferase step, the presence of an additional Am-Trf1 was not required for the reaction to progress. This is indicative of the well-known promiscuity of aminotransferases and suggests that the initial step in IAA metabolism is generally preserved among many organisms and used for other metabolic reactions.[38] Further testing of an affinity purified version of the *Pantoea* sp. YR343 aminotransferase showed that the enzyme was capable of effectively catalyzing the reaction but only with high concentrations of the enzyme (Figure S5). Additionally, we found that the IAA pathway in *Pantoea* sp. YR343 has a dependence on the IPDC PMI39_00059 (IPDC1). The importance of IPDC is made evident by the high concentrations of IAA produced by each *in vitro* reaction containing IPDC1. This suggests that the predominant IAA pathway in YR343 utilizes PMI39_00059 and is a good target for further exploration of *Pantoea* sp. YR343 IAA metabolism. This result is substantiated by previous work wherein a full deletion of the PMI39_00059 gene generated an 80% drop in IAA.[18]

The significant number of enzymes capable of catalyzing the dehydrogenation of indole-3-pyruvate to IAA emphasizes the need for rapid and effective *in vitro* tools. Sorting out the 16 different dehydrogenases by traditional genomic deletion approaches would be prohibitively difficult. While the lynchpin step in *Pantoea* sp. YR343's IAA metabolism is catalyzed by IPDC, the organism maintains several IALDh capable enzymes in its genome. Previous proteomic analyses of *Pantoea* sp. YR343 found that up to eight IALDhs could be detected at one time.[18] The redundancy of expression may indicate that *Pantoea* sp. YR343 maintains multiple pathways for IAA production. Interestingly, some of the most active enzymes found in this work were not found to be expressed in previous studies of *Pantoea* sp. YR343.[18] This may indicate that environmental conditions can control expression of particular IALDh enzymes and consequently IAA. Regardless of other potential substrates for these IALDhs, these multiple reaction paths can be defined by use of computational predictions when combined with validation using mixtures of crude enzyme extracts.

The definition of *Pantoea* sp. YR343's IAA pathway components uncovers its potential evolutionary development. The ability to find active enzymes, independent of genomic context, allows identification of those enzymes explicitly designed to carry out a function as well as those that exhibit flexibility in regard to substrate recognition. Underground metabolic functions describe potential side reactions of an enzyme and can serve as the basis for the emergence of new metabolic pathways without the need for significant evolutionary jumps.[39] In the case of *Pantoea* sp. YR343, the prevalence of specific enzymatic components related to IAA production and simultaneous lack of others show the potential for underground metabolic reactions to generate novel metabolic functions. For *Pantoea* sp. YR343, the genes responsible for the three key conversion steps exist outside a common operon. IAA metabolism may therefore be an opportunistic phenotype generated through the crossover of a single gene in the form of IPDC or the mutation of a homologous IPDC gene. Using a crude extract approach to explore the metabolic potential of YR343, we were able to discern these promiscuous functions and provide evolutionary context for the development of IAA metabolism in *Pantoea* sp. YR343.

This work has shown a rapid and scalable approach allowing for the identification and verification of active metabolic pathways in an organism by empirically generating functional annotations. The use of crude cell extracts accurately predicted metabolic pathways despite the functional redundancy and lack of genomic context for the enzymes involved in IAA metabolism in the organism *Pantoea* sp. YR343. We expect that improvements to computational tools will enhance the use of predictive analysis and the throughput of enzymatic discovery. As the breadth of genomic data continues to expand so too will the need to study such data without genetically tractable or even culturable organisms. The work presented demonstrates the effectiveness of crude extracts as a discovery and validation tool for both well-defined and underground metabolisms. Further, the development of high-throughput DNA synthesis and cell-free expression from eukaryotic cells could allow similar rapid explorations of metabolic pathways found in eukaryotic genomes and metagenomes. *In vitro* enabled analysis such as that presented in this work can help facilitate such studies by providing an accurate and rapid method of testing large search spaces in short amounts of time.

■ METHODS

Pathway and Gene Identification. Following a literature search, a general model of IAA production from L-tryptophan was created from a set of query proteins in order to perform a BLASTP search against the *Pantoea* sp. YR343 translated genome (Table S1, Figure 2).[40] At least one BLAST hit from the literature-derived query proteins was used to designate a protein as being related to IAA biosynthesis in *Pantoea* sp. YR343. Multiple queries were used when possible. As the goal of the BLAST analysis was to create a subset of proteins from the genome of interest for further analyses, a single verified enzyme was deemed sufficient for the initial annotation. The BLAST analysis was performed by compiling a database of query proteins to BLAST against the *Pantoea* sp. YR343 genome. *Pantoea* sp. YR343's genome was downloaded from GenBank. The output was parsed by searching for potentially complete pathways in the genome using the general model created from the query sequences. The final set of pathways was obtained after setting the *E*-value cutoff at 1×10^{-25} in order to eliminate known non-IAA producers, in this case, *Escherichia coli*.

Ligand-Docking Simulations. Protein homology models were created using the Protein Homology/analogy Recognition Engine V 2.0 (Phyre2) by providing the target sequence from *Pantoea* sp. YR343 (Table S2).[41] The top rated models were used, except in cases of low template identity, in 3DLigandSite to predict the substrate binding sites if this information was not immediately available from the template protein structures.[24] AutoDock Vina version 1.1.2 was employed to predict the interactions of each protein and its potential binding partners.[42] The search volume was set to 30 × 30 × 30 Å^3 centered around the binding site predicted by 3DLigandSite; the exhaustiveness was set to 8, and maximum energy difference was set to 3 kcal mol^{-1} for each protein−ligand combination (Figure 2A). Successful binding partners were determined based on both the docking poses and docking score rankings (i.e., binding to the ligand-docking site predicted by 3DLigandSite); a set of binding and nonbinding representative examples can be seen in (Figure S3). The lowest-energy pose was used determine binding to the predicted site. Proteins predicted to bind their putative substrates were subsequently used for biochemical experiments and culled if no docked poses were identified by Vina.

Enriched Extract Preparation. One Shot TOP10 Chemically Competent *E. coli* (ThermoFisher) was used as the cloning strain for plasmid preparation. Each potentially IAA-related enzyme was amplified from *Pantoea* sp. YR343 and inserted into the NdeI/SalI site of pET-30a(+) growing on kanamycin (50 μg mL^{-1}) (Table S3). All primers were designed using NEBuilder (New England Biolabs) and purchased from IDT. Crude cell-free extracts were prepared by culturing *E. coli* BL21 Star (DE3) in 2×YPTG (16 g L^{-1} tryptone, 10 g L^{-1} yeast extract, 5 g L^{-1} NaCl, 7 g L^{-1} KH_2PO_4, 3 g L^{-1} K_2HPO_4, 18 g L^{-1} glucose). Cultures of 50 mL volumes were grown using 250 mL baffled flasks at 37 °C shaking at 250 rpm. Induction was performed using 0.1 mM IPTG at OD600 = 0.6−0.8 and harvested after growing for 4 h at 30 °C. No antibiotics were used during growth. Each culture was harvested by centrifugation at 5000*g* for 10 min and washed twice with S30 buffer (2 g L^{-1} magnesium acetate, 14.05 g L^{-1} potassium glutamate, 0.154 g L^{-1} dithiothreitol (DTT), and 1.81 g L^{-1} Tris-acetate, pH 8.2). After the final wash, the cell pellets were weighed, flash-frozen in liquid nitrogen, and stored at −80 °C. Cell extracts were made by thawing and resuspending the pellet in 0.8 mL of S30 buffer per gram of wet cell weight before sonicating with 530 J mL^{-1} of suspension at 50% amplitude while in ice water. After sonication, the lysed cells were centrifuged twice at 4 °C for 10 min at 21 100*g*. The clarified lysate was flash-frozen, and stored at −80 °C. The BL21 Star (DE3) strain was used as the base expression strain. Plasmids that did not produce soluble protein in BL21 Star (DE3) were moved to BL21(DE3)/pLysS but were otherwise prepared in the same manner. Both pmi39_00977 and pmi39_04201 successfully expressed in BL21(DE3)/pLysS. After preparing each extract, 8 ng of the soluble crude extract was loaded onto an SDS-PAGE gel in order to verify expression of each enzyme. His-tag purified enzyme production required varying growth conditions outlined in the Supporting Information (Figure S4).

***In Vitro* Reactions.** *In vitro* IAA synthesis reactions were prepared by combining 50 mM L-tryptophan and ∼5 mg mL^{-1} of each enriched extract in a 30 μL volume. Control reactions that omitted putative reaction steps had volume shortage made up using cell extract from *E. coli*. The reactions were then placed in a 28 °C incubator shaking at 250 rpm for 24 h. Following incubation, each reaction was placed on ice and deactivated by adjusting the pH to 2.0 with HCl. IAA was extracted by the addition of 500 μL of ethyl acetate and vigorously vortexed. Each sample was then incubated on ice for 5 min and briefly centrifuged to separate aqueous and organic layers. A 400 μL portion of the organic layer was removed and dried in an analytical vial with argon gas. Vials were stored at 4 °C before analysis. Stored samples were resuspended in 500 μL of water before injection. Reaction components and conditions for purified enzyme reactions are described in detail in Figure S5.

Analysis of IAA Production. Quantitative analysis was performed by injecting 100 μL of suspended reaction into an Agilent 1260 HPLC instrument equipped with an Agilent ZORBAX Eclipse Plus C18 column and a diode array detector set at 265 nm. The mobile phase comprised A, 0.08% trifluoroacetic acid in water; and B, acetonitrile. Analytes were eluted at a flow rate of 1 mL min^{-1} with the initial eluent composition of 5% B held for 3 min, followed by a step to 30% B, a 5 min linear gradient to 45% B followed by a 5 min hold, a step to 75% B followed by a 1 min hold, and a step to 5% B followed by a 6 min hold. A calibration curve was prepared using pure IAA. Representative spectra are shown in the Supporting Information (Figure S6). All chemicals were acquired from Sigma-Aldrich. Error bars were calculated using the standard deviation from $n \geq 3$ independent reactions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acschembio.9b00725.

> Protein gel of IAA related proteins; protein neighborhoods of IAA related proteins; query proteins used in this work for BLAST analysis; *Pantoea* sp. YR343 proteins used in ligand-docking experiments; compiled results of BLAST analysis, ligand docking, and IAA production; representative example of ligand-docking analysis; protein gel of purified IAA related proteins; IAA production of purified IAA related protein combinations; and HPLC analysis of IAA standard and output from IAA related protein combinations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: doktyczmj@ornl.gov.

ORCID

Xiaolin Cheng: 0000-0002-7396-3225
Jennifer L. Morrell-Falvey: 0000-0002-9362-7528
Mitchel J. Doktycz: 0000-0003-4856-8343

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Special thanks to J. Parks for providing valuable advice during the preparation of this manuscript. This research was sponsored by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research, as part of the Plant Microbe Interfaces Scientific Focus Area (http://pmi.ornl.gov). Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725. D.C.G. is an NSF Graduate Fellow. This manuscript has been authored by UT-Battelle, LLC under Contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

■ REFERENCES

(1) Martínez-del Campo, A., Bodea, S., Hamer, H. A., Marks, J. A., Haiser, H. J., Turnbaugh, P. J., and Balskus, E. P. (2015) Characterization and Detection of a Widely Distributed Gene Cluster That Predicts Anaerobic Choline Utilization by Human Gut Bacteria. *mBio 6*, 1−12.

(2) Tan, G.-Y., Deng, Z., and Liu, T. (2015) Recent advances in the elucidation of enzymatic function in natural product biosynthesis. *F1000Research 4*, 1399.

(3) Gerlt, J. A., Allen, K. N., Almo, S. C., Armstrong, R. N., Babbitt, P. C., Cronan, J. E., Dunaway-Mariano, D., Imker, H. J., Jacobson, M. P., Minor, W., Poulter, C. D., Raushel, F. M., Sali, A., Shoichet, B. K., and Sweedler, J. V. (2011) The Enzyme Function Initiative. *Biochemistry 50*, 9950−9962.

(4) Zaslavsky, L., Ciufo, S., Fedorov, B., and Tatusova, T. (2016) Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC Bioinf. 17*, 276.

(5) Koskinen, P., Törönen, P., Nokso-Koivisto, J., and Holm, L. (2015) PANNZER: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics 31*, 1544−1552.

(6) Lamontagne, J., Béland, M., Forest, A., Côté-Martin, A., Nassif, N., Tomaki, F., Moriyón, I., Moreno, E., and Paramithiotis, E. (2010) Proteomics-based confirmation of protein expression and correction of annotation errors in the Brucella abortus genome. *BMC Genomics 11*, 300.

(7) Lewis, K., Epstein, S., D'Onofrio, A., and Ling, L. L. (2010) Uncultured microorganisms as a source of secondary metabolites. *J. Antibiot. 63*, 468−476.

(8) Cobb, R. E., Wang, Y., and Zhao, H. (2015) High-Efficiency Multiplex Genome Editing of Streptomyces Species Using an Engineered CRISPR/Cas System. *ACS Synth. Biol. 4*, 723−728.

(9) Tripathi, S. A., Olson, D. G., Argyros, D. A., Miller, B. B., Barrett, T. F., Murphy, D. M., McCool, J. D., Warner, A. K., Rajgarhia, V. B., Lynd, L. R., Hogsett, D. A., and Caiazza, N. C. (2010) Development of pyrF-Based genetic system for targeted gene deletion in clostridium thermocellum and creation of a pta mutant. *Appl. Environ. Microbiol. 76*, 6591−6599.

(10) Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E. J., Short, J. M., and Keller, M. (2002) Cultivating the uncultured. *Proc. Natl. Acad. Sci. U. S. A. 99*, 15681−15686.

(11) McClerklin, S. A., Lee, S. G., Harper, C. P., Nwumeh, R., Jez, J. M., and Kunkel, B. N. (2018) Indole-3-acetaldehyde dehydrogenase-dependent auxin synthesis contributes to virulence of Pseudomonas syringae strain DC3000. *PLoS Pathog. 14*, 1−24.

(12) Hitchcock, D. S., Fan, H., Kim, J., Vetting, M., Hillerich, B., Seidel, R. D., Almo, S. C., Shoichet, B. K., Sali, A., and Raushel, F. M. (2013) Structure-guided discovery of new deaminase enzymes. *J. Am. Chem. Soc. 135*, 13927−13933.

(13) Dhoke, G. V., Ensari, Y., Davari, M. D., Ruff, A. J., Schwaneberg, U., and Bocola, M. (2016) What's My Substrate? Computational Function Assignment of Candida parapsilosis ADH5 by Genome Database Search, Virtual Screening, and QM/MM Calculations. *J. Chem. Inf. Model. 56*, 1313−1323.

(14) Hermann, J. C., Marti-Arbona, R., Fedorov, A. A., Fedorov, E., Almo, S. C., Shoichet, B. K., and Raushel, F. M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature 448*, 775−779.

(15) 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., and Jacobson, M. P. (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature 502*, 698−702.

(16) Calhoun, S., Korczynska, M., Wichelecki, D. J., Francisco, B. S., Zhao, S., Rodionov, D. A., Vetting, M. W., Al-Obaidi, N. F., Lin, H., Meara, M. J. O., Scott, D. A., Morris, J. H., Russel, D., Almo, S. C., Osterman, A. L., Gerlt, J. A., Jacobson, M. P., Shoichet, B. K., and Sali, A. (2018) Prediction of enzymatic pathways by integrative pathway mapping. *eLife 7*, 1−27.

(17) Bible, A. N., Fletcher, S. J., Pelletier, D. A., Schadt, C. W., Jawdy, S. S., Weston, D. J., Engle, N. L., Tschaplinski, T., Masyuko, R., Polisetti, S., Bohn, P. W., Coutinho, T. A., Doktycz, M. J., and Morrell-Falvey, J. L. (2016) A carotenoid-deficient mutant in Pantoea sp. YR343, a bacteria isolated from the Rhizosphere of Populus deltoides, is defective in root colonization. *Front. Microbiol. 7*, 1−15.

(18) Estenson, K., Hurst, G. B., Standaert, R. F., Bible, A. N., Garcia, D., Chourey, K., Doktycz, M. J., and Morrell-Falvey, J. L. (2018) Characterization of Indole-3-acetic Acid Biosynthesis and the Effects of This Phytohormone on the Proteome of the Plant-Associated Microbe *Pantoea* sp. YR343. *J. Proteome Res. 17*, 1361.

(19) Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Del Rio, T. G., Edgar, R. C., Eickhorst, T., Ley, R. E., Hugenholtz, P., Tringe, S. G., and Dangl, J. L. (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature 488*, 86−90.

(20) Utturkar, S. M., Cude, W. N., Robeson, M. S., Yang, Z. K., Klingeman, D. M., Land, M. L., Allman, S. L., Lu, T.-Y. S., Brown, S. D., Schadt, C. W., Podar, M., Doktycz, M. J., and Pelletier, D. A. (2016) Enrichment of root endophytic bacteria from Populus deltoides and single-cell genomics analysis. *Appl. Environ. Microbiol. 82*, 5698.

(21) Utturkar, S. M., Cude, W. N., Robeson, M. S., Yang, Z. K., Klingeman, D. M., and Land, M. L. (2016) Enrichment of Root Endophytic Bacteria from Populus deltoides and. *Appl. Environ. Microbiol. 82*, 5698−5708.

(22) Spaepen, S., Vanderleyden, J., and Remans, R. (2007) Indole-3-acetic acid in microbial and microorganism-plant signaling. *FEMS Microbiol. Rev. 31*, 425−448.

(23) Kelley, L. A., Mezulis, S., Yates, C., Wass, M., and Sternberg, M. (2015) The Phyre2 web portal for protein modelling, prediction, and analysis. *Nat. Protoc. 10*, 845−858.

(24) Wass, M. N., Kelley, L. A., and Sternberg, M. J. E. (2010) 3D Ligand Site: predicting ligand-binding sites using similar structures. *Nucleic Acids research 38*, 469−473.

(25) Deng, N., Forli, S., He, P., Perryman, A., Wickstrom, L., Vijayan, R. S. K., Tiefenbrunn, T., Stout, D., Gallicchio, E., Olson, A. J., and Levy, R. M. (2015) Distinguishing binders from false positives by free energy calculations: Fragment screening against the flap site of HIV protease. *J. Phys. Chem. B 119*, 976−988.

(26) Fujimoto, M. S., Suvorov, A., Jensen, N. O., Clement, M. J., and Bybee, S. M. (2016) Detecting false positive sequence homology: A machine learning approach. *BMC Bioinformatics 17*, 1−11.

(27) Houston, D. R., and Walkinshaw, M. D. (2013) Consensus docking: Improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model. 53*, 384−390.

(28) Stigliani, J. L., Bernardes-Génisson, V., Bernadou, J., and Pratviel, G. (2012) Cross-docking study on InhA inhibitors: A combination of Autodock Vina and PM6-DH2 simulations to retrieve bio-active conformations. *Org. Biomol. Chem. 10*, 6341−6349.

(29) Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., and Hou, T. (2016) Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: The prediction accuracy of sampling power and scoring power. *Phys. Chem. Chem. Phys. 18*, 12964−12975.

(30) Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol. 25*, 197−206.

(31) Durrant, J., and McCammon, J. A. (2011) Molecular dynamics simulations and drug discovery. *BMC Biol. 9*, 1−9.

(32) Spyrakis, F., Benedetti, P., Decherchi, S., Rocchia, W., Cavalli, A., Alcaro, S., Ortuso, F., Baroni, M., and Cruciani, G. (2015) A Pipeline to Enhance Ligand Virtual Screening: Integrating Molecular Dynamics and Fingerprints for Ligand and Proteins. *J. Chem. Inf. Model. 55*, 2256−2274.

(33) Stewart, E. J. (2012) Growing unculturable bacteria. *J. Bacteriol. 194*, 4151−4160.

(34) Lagier, J. C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M., and Raoult, D. (2015) Current and past strategies for bacterial culture in clinical microbiology. *Clin. Microbiol. Rev. 28*, 208−236.

(35) Jeon, W. B., Aceti, D. J., Bingman, C. A., Vojtik, F. C., Olson, A. C., Ellefson, J. M., McCombs, J. E., Sreenath, H. K., Blommel, P. G., Seder, K. D., Burns, B. T., Geetha, H. V., Harms, A. C., Sabat, G., Sussman, M. R., Fox, B. G., and Phillips, G. N. (2005) High-

throughput purification and quality assurance of Arabidopsis thaliana proteins for eukaryotic structural genomics. *J. Struct. Funct. Genomics 6*, 143−147.

(36) Lesley, S. A. (2001) High-throughput proteomics: Protein expression and purification in the postgenomic world. *Protein Expression Purif. 22*, 159−164.

(37) Kim, Y., Babnigg, G., Jedrzejczak, R., Eschenfeldt, W. H., Li, H., Maltseva, N., Hatzos-Skintges, C., Gu, M., Makowska-Grzyska, M., Wu, R., An, H., Chhor, G., and Joachimiak, A. (2011) High-throughput protein purification and quality assessment for crystallization. *Methods 55*, 12−28.

(38) Lal, P. B., Schneider, B. L., Vu, K., and Reitzer, L. (2014) The redundant aminotransferases in lysine and arginine synthesis and the extent of aminotransferase redundancy in. *Mol. Microbiol. 94*, 843−856.

(39) Notebaart, R. A., Szappanos, B., Kintses, B., Pal, F., Gyorkei, A., Bogos, B., Lazar, V., Spohn, R., Csorg, B., Wagner, A., Ruppin, E., Pal, C., and Papp, B. (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc. Natl. Acad. Sci. U. S. A. 111*, 11762−11767.

(40) Altschul, S. F., Gish, W., Miller, W., Myers, E. E. W. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403.

(41) Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. E. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc. 10*, 845−858.

(42) Trott, O., and Olson, A. J. (2009) Auto Dock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem. 31*, NA−NA.