BMC Bioinformatics

RESEARCH Open Access

A probabilistic approach for predicting indole-3-acetic acid synthesis in bacteria using genomic data

Zheng-Xiang Ye[1] and Steven H. Wu[1*]

*Correspondence:
Steven H. Wu
stevenwu@ntu.edu.tw
[1]Department of Agronomy, National Taiwan University, Taipei, Taiwan

Abstract

Background Indole-3-acetic acid (IAA) is a crucial plant hormone for growth and development. Among bacteria capable of IAA synthesis, tryptophan-dependent pathways are the primary pathway for IAA production. Identifying IAA-producing bacteria is essential for understanding plant–microbe interactions and developing sustainable agricultural practices. With the advancements in sequencing technologies which enable the acquisition of genomic data in a short amount of time at a reduced cost. This study developed a probabilistic framework to predict the likelihood of IAA synthesis in bacteria using their genomes, which offers a cost-effective screening approach compared to traditional wet-lab methods.

Results The prediction framework incorporates four tryptophan-dependent pathways and consists of two models. The Binary Model provides a binary prediction of IAA synthesis, and the Logistic Model ranks bacteria based on their synthesis potential. The models were validated against a comprehensive dataset of more than 15,000 bacterial genomes, and the results suggest that only 1.7% of the bacteria have the potential to synthesize IAA. The most promising phyla are Actinobacteria, Proteobacteria, and Firmicutes. The Logistic Model shows that the top-ranked genera include *Streptomyces*, *Paraburkholderia*, *Nonomuraea*, *Rhodococcus,* and *Nocardia*. The model also suggests that the potential for IAA synthesis is likely to be strain-specific, emphasizing the importance of analyzing individual strains within a genus.

Conclusion The proposed probabilistic prediction framework is a cost-effective *in-silico* approach for identifying bacteria with a high potential for IAA synthesis. This serves as a pre-screening tool for resource-intensive downstream wet-lab validation, significantly reducing the time and effort required to identify potential plant growth-promoting bacteria. Furthermore, the model can be customized for other pathways of interest, and making it a versatile framework to explore various bacterial metabolic processes. The source code of the model is freely available at https://github.com/ComputationalAgronomy/bcpip

Keywords Indole-3-acetic acid, Bacterial genomes, Biological pathway modeling, Prediction model, Plant growth-promoting rhizobacteria

Introduction

With the advancements in next-generation sequencing technology, a large amount of genomic data is available at a relatively low cost. Genome annotation and analysis of these sequences provide valuable information on the functional potential of organisms. By integrating this knowledge into the biological modeling framework, researchers are able to decode the information stored in the genome sequences.

Auxin, a type of plant hormone, plays an essential role in plant development. It governs cell growth, tropic response, tissue differentiation, and organ architecture and patterning [1, 2]. Among the various compounds with auxin activities, one important native and major form of auxin in plants is indole-3-acetic acid (IAA) [3]. IAA is synthesized using the amino acid tryptophan (Trp) as a precursor. Although Trp-independent pathways for IAA production have been suggested in several studies [4, 5], the underlying mechanisms have not yet been elucidated with solid evidence. The production of IAA is not exclusive to plants, some genera of bacteria are also capable of synthesizing IAA, including *Agrobacterium*, *Azospirillum*, *Pseudomonas*, *Xanthomonas*, and *Rhizobium* [6]. However, the production of IAA in bacteria does not guarantee a positive role in promoting plant growth. The phytopathogenic properties of *Agrobacterium* spp. and *Pseudomonas savastanoi* pv. *savastanoi* were well known and characterized [7, 8]. Their infection causes tumor and gall formation, which results from the disrupted IAA level near the infection sites.

Plant growth-promoting rhizobacteria (PGPR) are a group of bacteria that are found in the rhizosphere and possess the ability to promote plant growth [9]. One of the key mechanisms underlying the growth-promoting effects of PGPR is its capacity to synthesize and secrete IAA. The acquisition of additional IAA by plants from bacterial sources can lead to an increase in root surface area and root length, subsequently enhancing nutrient and water uptake [10, 11]. With the global population continuing to rise, improving crop yields and establishing sustainable agricultural practices have become primary objectives. Researchers have been utilizing IAA-producing PGPR strains as biostimulants to boost crop productivity and promote sustainable agricultural practices [12, 13].

There are four major Trp-dependent pathways for IAA synthesis have been described, along with a few minor pathways made up of rare enzymes that only exist in certain species [6, 14]. The pathways are often named after the key intermediate of each pathway. The first pathway is the indole-3-acetamide (IAM) pathway. The presence of the IAM pathway was found in plant pathogens such as *Pseudomonas syringae pv. savastanoi*, *Agrobacterium tumefaciens*, as well as nitrogen-fixing bacteria such as *Rhizobium* spp. and *Bradyrhizobium* spp. [15–17]. The second pathway, the indole-3-pyruvate (IPA) pathway, is a major pathway for plants to synthesize IAA, especially via a two-step conversion. This pathway was also detected in soil-dwelling bacteria *Azospirillum brasilense, Azospirillum lipoferum*, and *Enterobacter cloacae* [18–20]. The third pathway, the tryptamine (TAM) pathway, is a three-step conversion from Trp to IAA. Although this pathway was reported to be abundant in many plants and fungi, little research has been done on bacteria [21]. The fourth pathway is the indole-3-acetonitrile (IAN) pathway. Most studies on the IAN pathway have focused on higher plants, and only a few research studies in bacteria have been reported [21]. These include *Pseudomonas syringae* pv. *syringae* reported by Howden et al. [22] and the IAN-specific nitrilase identified

in *Alcaligenes faecalis* [23]. A schematic diagram for all four Trp-dependent pathways is shown in Fig. 1. The location of these genes varies depending on the species. Species such as *Azospirillum brasilense* and *Bacillus amyloliquefaciens* have these genes located in chromosomes [24, 25]. Genes located in plasmids are found in several species, including *Agrobacterium tumefaciens* and *Pseudomonas savastanoi* [26, 27]. These IAA genes located in plasmids might facilitate horizontal gene transfer (HGT) to different bacterial species throughout evolutionary history [28, 29].

The publicly accessible databases on NCBI offer a huge amount of genomic data. Zhang et al. [30] conducted a large-scale genomic analysis on five phyla of bacteria. They predicted the genes in the genomes and determined whether these bacteria possess the enzymes in the IAA pathway. This study concluded that 82.2% of the bacteria have the potential to synthesize IAA from intermediates, 18.5% of the bacteria can de novo synthesize IAA from Trp, and around 45% of bacteria exhibit multiple coexisting pathways. Their results show that the existence of the enzymes in the IAA pathway is quite common in bacteria. However, the relative potential of the bacteria to synthesize IAA was not analyzed in their study. With the additional genomic data accumulated over the years, we are now able to screen a large number of bacterial genomes and estimate the potential for IAA synthesis in each bacterial strain.

In this study, we developed a probabilistic framework to predict the potential for IAA synthesis in bacteria based on the enzymes in the IAA pathway. The framework estimated a score that denotes the likelihood of the IAA synthesis activities. By ranking these scores in descending order, bacteria with high scores are more likely to have the potential to de novo synthesize IAA via the Trp-dependent pathways. The proposed probabilistic framework provides an efficient in-silico screening method for bacterial strains with potential for IAA synthesis.

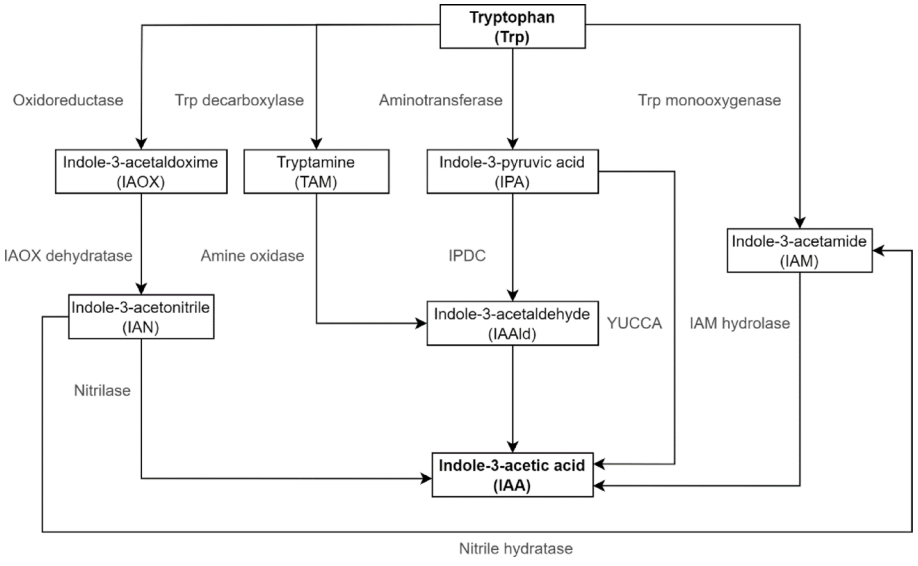


Fig. 1 Trp-dependent IAA pathways. Chemical compounds are represented by rectangle boxes (referred to as nodes). Abbreviations are given under the full name of each chemical compound. Enzymes are represented by arrows (referred to as edges). In the Trp-dependent pathways, indole-3-acetic acid (IAA) is synthesized from tryptophan (Trp). Four major pathways: 1) indole-3-acetamide (IAM) pathway, 2) indole-3-pyruvic acid (IPA) pathway, 3) tryptamine (TAM) pathway, and 4) indole-3-acetonitrile (IAN) pathway are included in this graph

Methods

Genome collection and functional annotation

This study included 15,377 bacterial genomes from 5 prevailing phyla in soil and rhizosphere environments. The selection of these five phyla (Pseudomonadota, Bacillota, Actinomycetota, Bacteroidota, and Acidobacteriota) was based on previous studies of their importance in agricultural systems and potential for IAA production [31–34]. Previous large-scale genomic analysis demonstrated that rhizobacteria from these phyla had high potential to synthesize IAA [30]. In addition, extensive experimental evidence shows IAA production rates ranging from 16.7% to 83% among isolates from these groups [35–37]. All available reference genomes from these five phyla were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/) at the time of analysis to ensure comprehensive coverage of agriculturally relevant bacteria with IAA production potential. Only reference genomes (RefSeq, the representative sequence of a species) were included in this study to maintain data quality and avoid redundancy [38]. The number of genomes from each phylum is summarized in Table 1. All the accession numbers of these reference genomes are available in Additional File 2.

The protein-coding genes for each genome were predicted by Prodigal v2.6.3 [39] with default parameters, and the predicted coding sequences were converted to amino acid sequences using genetic code described by NCBI (Code 11: The Bacterial, Archaeal and Plant Plastid Code). The putative amino acid sequences were then annotated using Diamond v2.0.15 with the blastp mode [40], which uses an amino acid sequence as input to query against a custom curated IAA database and default parameters. The custom curated IAA database contained protein entries retrieved from UniProtKB/Swiss-Prot [41], which comprises manually annotated protein sequences. The protein entries related to the IAA pathway were collected by searching the corresponding EC numbers provided in the tryptophan metabolism pathway (map00380, Tryptophan metabolism) in the KEGG pathway database [42]. The EC numbers and keywords used for searching UniProtKB/Swiss-Prot and the collected entries are listed in Table S1 in Additional File 1 and Additional File 3. The alignment results from Diamond were filtered using query coverage $\geq$ 50% which is based on thresholds established from previous IAA pathway genomics studies [30]. This coverage criterion was specifically chosen to ensure substantial overlap with functional domains and reduce false positive annotations from short, high-identity matches that may not encompass critical catalytic regions. If a protein matches to multiple sequences in the database, only the highest-scoring hit is retained for subsequent pathway analysis. The BLAST-based functional annotations support transfer of the first three EC digits (general reaction type) but not the fourth digit (substrate specificity), which is consistent with established sequence similarity thresholds for functional annotation transfer. While this approach may not be the most precise, it

Table 1 The number of genomes from five phyla

Phylum (synonym)	Number of genomes
Acidobacteriota (Acidobacteria)	50
Actinomycetota (Actinobacteria)	3430
Bacteroidota (Bacteroidetes)	1938
Bacillota (Firmicutes)	3020
Pseudomonadota (Proteobacteria)	6937

provides a reasonable balance between accuracy and computational efficiency for pathway-level predictions.

Prediction framework using the biological pathway

A directed acyclic graph was constructed to represent the IAA pathway. The graph defines each chemical compound as a node, enzymes that catalyze the same conversion of the compounds as an edge, and the reaction direction towards the product as the direction of the edge. The overall structure is equivalent to the Trp-dependent IAA pathways in Fig. 1. The best hits retained in the last section were mapped to the edges of the IAA pathway graph and scores were assigned to the edges to denote the likelihood of the corresponding enzyme to be able to catalyze the conversion.

The score determination is hierarchical and takes three steps, which are summarized in Box 1. In the first step, a score is calculated from each hit and is referred to as "hit score". The second step calculates the edge scores from hit scores. The third step calculates the node scores for the prediction framework.

Step 1: Hit score calculation

This prediction framework has two different models of hit scores, the Binary Model and the Logistic Model. Both models use the percentage of identity from the Diamond alignment output to calculate hit scores, which range from 0 to 100. The hit score H_i for an alignment hit i are calculated as follows for these two models:

Hit scores–binary model

$$H_i = \begin{cases} 1, & \text{if percentage of identity} \geq 40 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where x_i is the percentage of identity of hit i, and t is the threshold for the percentage of identity which is set to 40 because of the biological reasoning that 40% identity is a rule of thumb to determine whether two enzymes have similar functions [43].

Hit scores–logistic model

$$H_i = \frac{1}{1 + exp\left(-\frac{1}{k}\left(x_i - x_0\right)\right)} \tag{2}$$

where x_i is the percentage of identity of hit i, x_0 is the percentage of identity to be set as the midpoint, and k is the steepness of the sigmoid curve. This is equivalent to a logistic function with two parameters k and x_0 to adjust the steepness and translation of the function. The hit score produced is between 0 and 1 as a characteristic of the logistic function. The value of x_0 is set to 40 using the same reasoning as the Binary model.

In order to investigate the effect of the steepness parameter *k*, a range of values from 1 to 100 was tested, with $k = 1$ representing the steepest curve and $k = 100$ representing the flattest curve. As k approaches zero, the Logistic Model is reduced to the Binary Model. The resulting logistic curves for different k values are shown in Fig. 2. Our analysis showed that if k is chosen to be 10, the hit score is mainly contributed by the high-identity alignment hits. This conveys the biological concept that a single high-identity alignment hit is more likely to perform the expected function in close association with the known enzyme than multiple low-identity alignment hits. Furthermore, the sensitivity analysis of k suggested that the prediction results remain consistent for

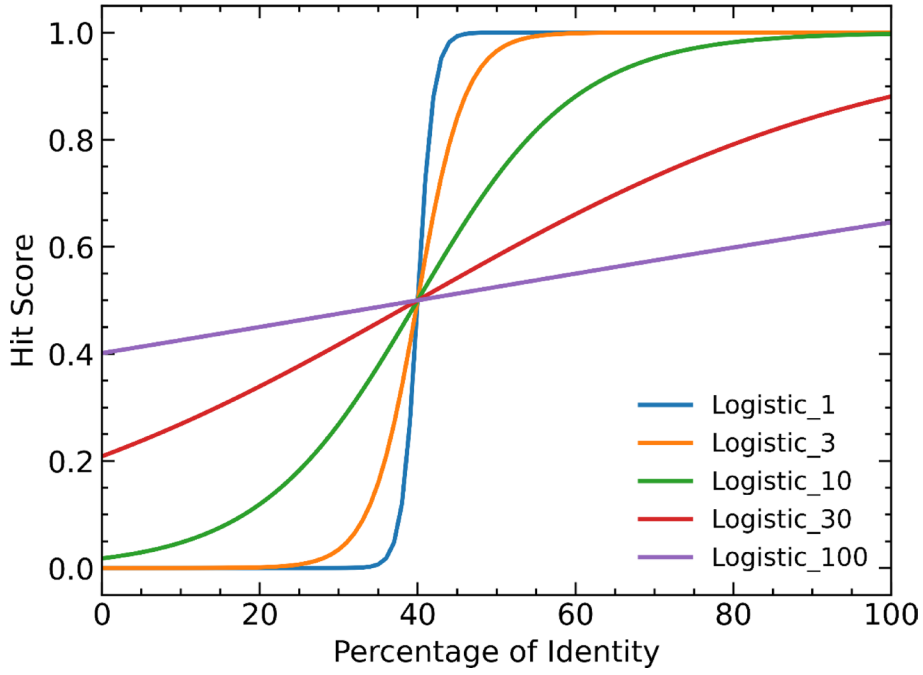


Fig. 2 The Logistic Model with different values of k. The number following the Logistic label is k. The model converts the percentage of identity of an alignment hit to a hit score

values approximately equal to 10. These findings support the decision to use $k = 10$ in the subsequent analysis (See Supplementary Materials in Additional File 1 for detailed information).

Step 2: Edge score calculation

The catalytic efficiency of an enzyme may be influenced by multiple factors, including variations in the copy numbers or multiple independent enzymes that perform the same function. Therefore, this model aggregated hit scores on the same edge by the following formula:

$$E_i = 1 - \prod_{j \in D_i} (1 - H_j) \tag{3}$$

where i is an edge, j is an alignment hit, D_i is the set of all hits mapped to edge i, and H is calculated from Eq. 1 or Eq. 2 depending on which model is used in the previous step. This formula is analogous to calculating the probability that at least one hit is functional by assuming hit scores are independent. The derived edge score is between 0 and 1. Due to the binary property, Eq. 3 can be reduced to the Binary Model using the following equation:

$$E_i = \begin{cases} 1, & \text{if at least one } H_{j \in D_i} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Step 3: Prediction model–node score calculation

A node in the pathway graph represents a chemical compound, and the score for node i can then be determined as follows:

$$N_i = 1 - \prod_{j \in N^-(i), k = e_{ji}} (1 - N_j \times E_k) \tag{5}$$

where e_{ji} is a directed edge between node j and node i, E is the edge score calculated from Eq. 3, and $N^-(i)$ is the set of neighboring nodes connected to node i, where each neighboring node has a directed edge e_{ji}. This approach creates a network structure to represent the biological pathway, and the information propagates from the upstream to downstream nodes. This also requires that all node scores of j to be defined before the node score of i can be calculated. The order for calculating all node scores is determined by the depth-first search (DFS) algorithm. This algorithm traverses the graph by exploring the first child of each node until it encounters a node without a child, at which point it backtracks to explore the next child of the previous node. A modified DFS algorithm is implemented where the child will be visited only when all the upstream nodes are visited. This algorithm ensures the node score of a node is calculated right after the node score of its last upstream node has been calculated.

In this research, tryptophan serves as the initial compound with a predefined node score of 1. All other node scores reflect the relative potential for that bacterium to synthesize the compound from tryptophan. The complete analytical workflow, from curating bacterial genomes to estimating the capabilities of synthesizing IAA, is summarized in Fig. 3.

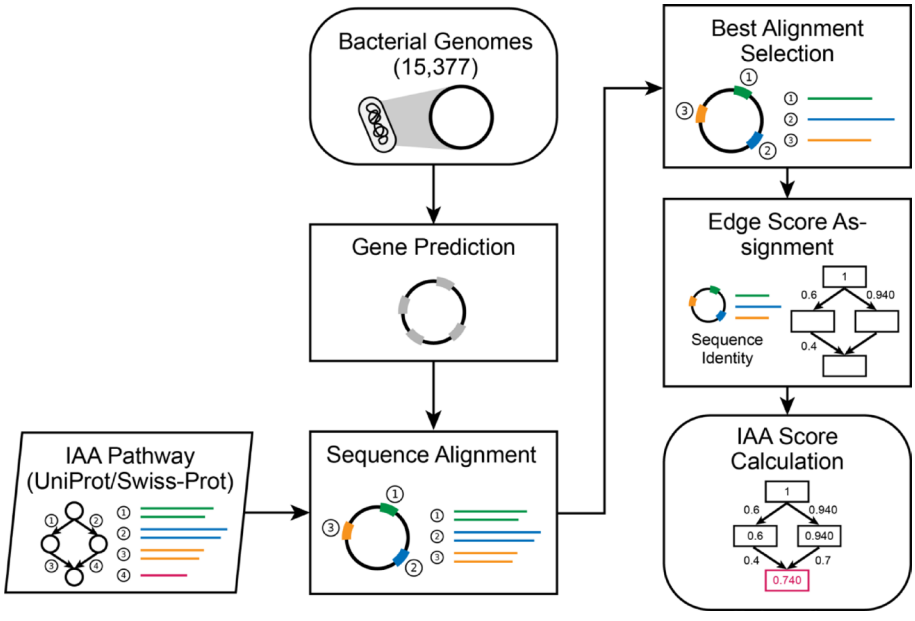


Fig. 3 The complete analytic workflow for predicting the potential of bacteria to synthesize IAA from tryptophan using the Binary and Logistic Models

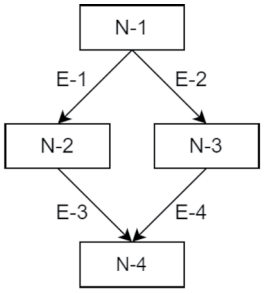
Box 1: An example of the prediction model

To demonstrate how the model works, an imaginary pathway is built as an example with 4 nodes and 4 edges. The calculation is walked through step by step and the Logistic Model is used in hit score conversion.

Step 1: Calculate the hit score for each alignment hit

Consider the following alignment hits H-1 and H-2 from the Diamond result. Only necessary information is retained here.

Hit ID	Edge Matched	Percentage of Identity
H-1	E-2	40
H-2	E-2	60



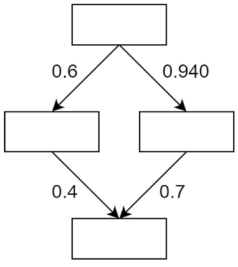
Edge Matched is the enzyme that each predicted protein sequence aligned to. Using the Logistic Model (Eq. 3) with $k = 10$ and $x_0 = 40$, the calculated hit score for H-1 is 0.5 and H-2 is 0.881.

Step 2: Calculate the edge score from hit scores

Since H-1 and H-2 are matched to the same edge, the edge score for edge E-2 is determined by these two hits and is derived by using Eq. 4.

$$E_2 = 1 - \prod_{j\in\{H-1,H-2\}} \left(1 - H_j\right) = 1 - (1 - 0.5) \times (1 - 0.881) = 0.940$$

The edge scores for the other edges are derived in the same way (arbitrarily assigned here).

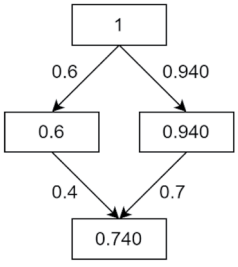


Step 3: Calculate the node score from edge scores

Give the starting material a score (N-1 is assigned with a score of 1). Traverse the graph and calculate the node scores by using Eq. 5. For instance, the node score for N-4 is calculated as follows:

$$N_4 = 1 - \prod_{(j,k)\in\{(N-2,E-3),(N-3,E-4)\}} \left(1 - N_j \times E_k\right)$$
$$= 1 - (1 - 0.6 \times 0.4) \times (1 - 0.940 \times 0.7) = 0.740$$

The node score for each compound is the likelihood the compound can be synthesized from the starting material under the model assumption.



Prediction result analysis and validation

The likelihood of being able to synthesize IAA from Trp was calculated with both models for all bacterial genomes described above. The Binary Model was used to predict whether a bacterium has at least one complete pathway (IAM, IPA, TAM, or IAN pathway) to synthesize IAA from Trp. The Logistic Model estimates a score between 0 and 1, and can be used to rank and assess the relative potential to synthesize IAA from Trp for each bacterium. The top 100 ranked bacteria were extracted and classified by phylum or genus to determine which phylum or genus contains bacteria that are likely to synthesize

IAA from Trp. This number serves as an example for demonstration purposes; the actual selection should be based on experimental requirements as mentioned in the Discussion section.

In order to evaluate the performance of both models, two validation datasets were curated. The first dataset contains a list of bacteria that are capable of producing IAA, which was curated from a recent review for validation [34]. Bacteria recorded at the strain level were considered to be the positive validation set of species. Unfortunately, the genomes of the reported strains were unavailable. Therefore, the whole taxonomy group was collected instead of specific strains. For those bacteria only reported at the genus level, all the species in the genus were analyzed to see whether the ability to synthesize IAA varies from species to species within a genus. In total, 22 species and seven genera were selected for validation. In addition, a sub-dataset was extracted to investigate the strain specificity for the prediction result. This sub-dataset contains six species with at least ten genomes each, and four genera with at least 50 species each in the dataset (See Additional File 4). In this validation study, the performance was only assessed on complete genomes. While the method can technically be applied to incomplete genomes, this would artificially inflate the false negative rate of the prediction model.

The IAA validation dataset only includes bacteria that utilize one of these four major Trp-dependent pathways, which are the major pathways utilized by most bacteria. This excluded one of the minor pathways which converted indole to IAA in one step via a single enzyme, IAA acetyltransferase (IAAT) (Figure S1 in Additional File 1). This is not a Trp-dependent pathway validation since indole is one of the precursors of tryptophan, and this bypassed all four pathways mentioned above. There is a small group of bacteria that utilize this minor pathway to synthesize IAA, such as *Bacillus amyloliquefaciens*. A study demonstrated that *B. amyloliquefaciens FZB42* has diminished IAA synthesis when the IAAT gene, *ysnE*, was mutated [44]. One additional validation dataset was curated with 168 genomes of *Bacillus amyloliquefaciens*. The presence of IAAT was confirmed in all strains by gene prediction and sequence alignment.

Results

We introduced this prediction framework with two models, the Binary Model, and the Logistic Model, to evaluate the likelihood of bacteria synthesizing IAA from Trp. The validation dataset is comprised of 15,377 bacterial genomes spanning five phyla. Both models use an IAA pathway graph with four major pathways to estimate each bacterium's potential to synthesize IAA.

Potential of bacteria to synthesize IAA

For the Binary Model, an alignment hit requires a minimum of 40 percent identity to an enzyme in the IAA pathway graph to be considered in the model. An edge score of one represents successful compound conversion, whereas a node score of one indicates that the corresponding chemical compound can be synthesized from Trp.

The IAA pathway graph is separated into three groups: the conversion from Trp to an intermediate compound (Fig. 4a), from an intermediate compound to IAA (Fig. 4c), and from an intermediate compound to another intermediate compound in the IAA pathway (Fig. 4b). The result shows that conversion from Trp to IAM and IAOX is lacking in most bacteria. Only 73 (0.5%) bacteria, mostly in Proteobacteria, had the potential to

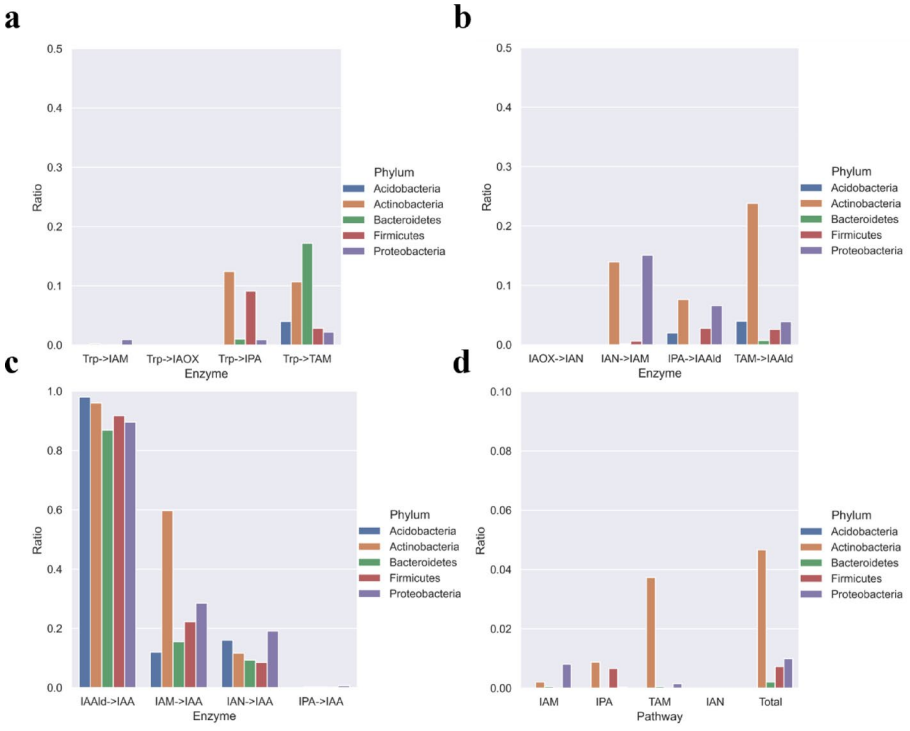


Fig. 4 Ratio of bacteria that possess a specific enzyme or pathway. **a** The conversion from Trp to an intermediate; **b** The conversion of intermediates; **c** The conversion from an intermediate to IAA; **d** The ability to convert Trp to IAA via a specific pathway. IAM: (Trp → IAM → IAA); IPA: (Trp → IPA → IAAld → IAA); TAM: (Trp → TAM → IAAld → IAA); IAN: (Trp → IAOX → IAN → IAA)

synthesize IAM from Trp, and no bacterium was likely to synthesize IAOX from Trp. On the other hand, 783 (5.1%) and 937 (6.1%) bacteria had the potential to synthesize IPA and TAM from Trp. Despite the lack of potential for bacteria to catalyze the conversion of IAM from Trp, a higher proportion of bacteria are able to convert IAM to IAA (15.5–59.7% for each phylum). By combining with the potential to synthesize IAM from Trp, 56 (0.8%) Proteobacteria, 7 (0.2%) Actinobacteria, and 1 (< 0.1%) Bacteroidetes were found to be able to synthesize IAA from Trp via the two-step IAM pathway (Fig. 4d). Though also a high proportion of bacteria (8.5–19.1%) could convert IAN to IAA, no bacteria were found to convert IAOX to IAN and therefore, in the IAN pathway, only the synthesis from the intermediate IAN to IAA was possible for the bacteria analyzed in this study. The ability of bacteria to catalyze the intermediate conversion in the IPA and the TAM pathway, IPA to IAAld and TAM to IAAld, ranged from 0 to 7.6% and 0.7 to 23.8% in each phylum. The enzyme IAAld dehydrogenase, shared between the IPA and the TAM pathway, was found to exist in 86.9 to 98% of bacteria in each phylum, and thus it was the existence of the initial and intermediate enzymes that differentiated the likelihood of bacteria to synthesize IAA from Trp. 30 (0.9%) Actinobacteria, 20 (0.7%) Firmicutes, and 2 (< 0.1%) Proteobacteria were able to synthesize IAA via the IPA pathway, whereas 128 (3.7%) Actinobacteria, 10 (0.1%) Proteobacteria, and 1 (< 0.1%) Bacteroidetes could synthesize IAA via the TAM pathway. The direct conversion of IPA to IAA through YUCCA, an enzyme being reported in the IAA pathway of plants, was absent in most genomes and only 69 (0.4%) bacteria were found to have enzymes similar to YUCCA. Overall, 160 (4.7%) Actinobacteria, 4 (0.2%) Bacteroidetes, 22 (0.7%)

Firmicutes, and 69 (1.0%) Proteobacteria were capable of synthesizing IAA from Trp if all possible pathways are considered.

Analysis of top-ranking genomes

The Logistic Model was used to estimate the IAA node score for all 15,377 bacterial genomes. The distribution of non-zero IAA node scores for 6132 bacterial genomes is plotted in Fig. 5a and resembles an exponentially decaying function. The IAA node score decreases rapidly from a peak of 0.999 to 0.821 at rank 100, reaching 0.591 at rank 500. The node score is less than 50% of the highest value by the rank of 1000. Therefore, the Logistic Model has a higher discriminating power for high-ranking genomes, subsequently increasing the confidence in selecting top candidates.

The genomes that received an IAA node score of one from the Binary Model were compared with their corresponding IAA node scores from the Logistic Model. The score distribution provides an empirical method to establish a threshold for the Logistic Model. Among 255 genomes with an IAA node score of one from the Binary Model, the median, mean, and standard deviation of the IAA node score are 0.731, 0.723, and 0.158, respectively (Fig. 5b). Forty-five percent of the genomes have scores between 0.6 and 0.8. The median score of 0.731 corresponds to rank 211 in the Logistic Model result (Fig. 5a).

The top 100 ranked bacteria all belong to the following three phyla: 61 Actinobacteria, 36 Proteobacteria, and 3 Firmicutes (See Additional File 5). This result was in congruence with the Binary Model, where the phylum with the highest potential was Actinobacteria, followed by Proteobacteria and Firmicutes. The top 100 bacteria in the Logistic model spread across 30 genera, and the top 5 most abundant genera were *Streptomyces*, *Paraburkholderia*, *Nonomuraea*, *Rhodococcus,* and *Nocardia*. All of the genera are Actinobacteria except *Paraburkholderia*, which is Proteobacteria.

Prediction model validation

Two validation datasets were curated to evaluate the performance of the model. The first dataset contains a list of experimentally validated IAA-producing bacteria based on previous literature. The second dataset, the *Bacillus amyloliquefaciens* dataset, was curated because *B. amyloliquefaciens* utilizes an alternative IAA pathway that is not incorporated in the models.

For the *B. amyloliquefaciens* dataset, the Binary Model assigned a score of 0 for all strains, and the highest score from the Logistic Model is only 0.167, which is relatively low in the score distribution figure (Fig. 5a, c). Notably, not all the strains have the same Logistic score, with a mean of 0.102 and a standard deviation of 0.033.

For the dataset with previously reported taxonomy groups, the Binary model shows only one out of 22 reported species and four out of seven reported genera that were predicted to possess the potential to synthesize IAA from Trp (Table S2 in Additional File 1). This discrepancy along with the score difference between the strains in *B. amyloliquefaciens* suggested further investigation of the specificity of IAA production within a species or genus.

The strain-specificity was investigated with the Logistic Model on the sub-dataset, which consists of six species with at least ten genomes each, and four genera with at least 50 species each in the dataset. The score distributions are shown in Figs. 6 and 7, with the summary statistics provided in Table 2. The Logistic Model revealed a wide range of

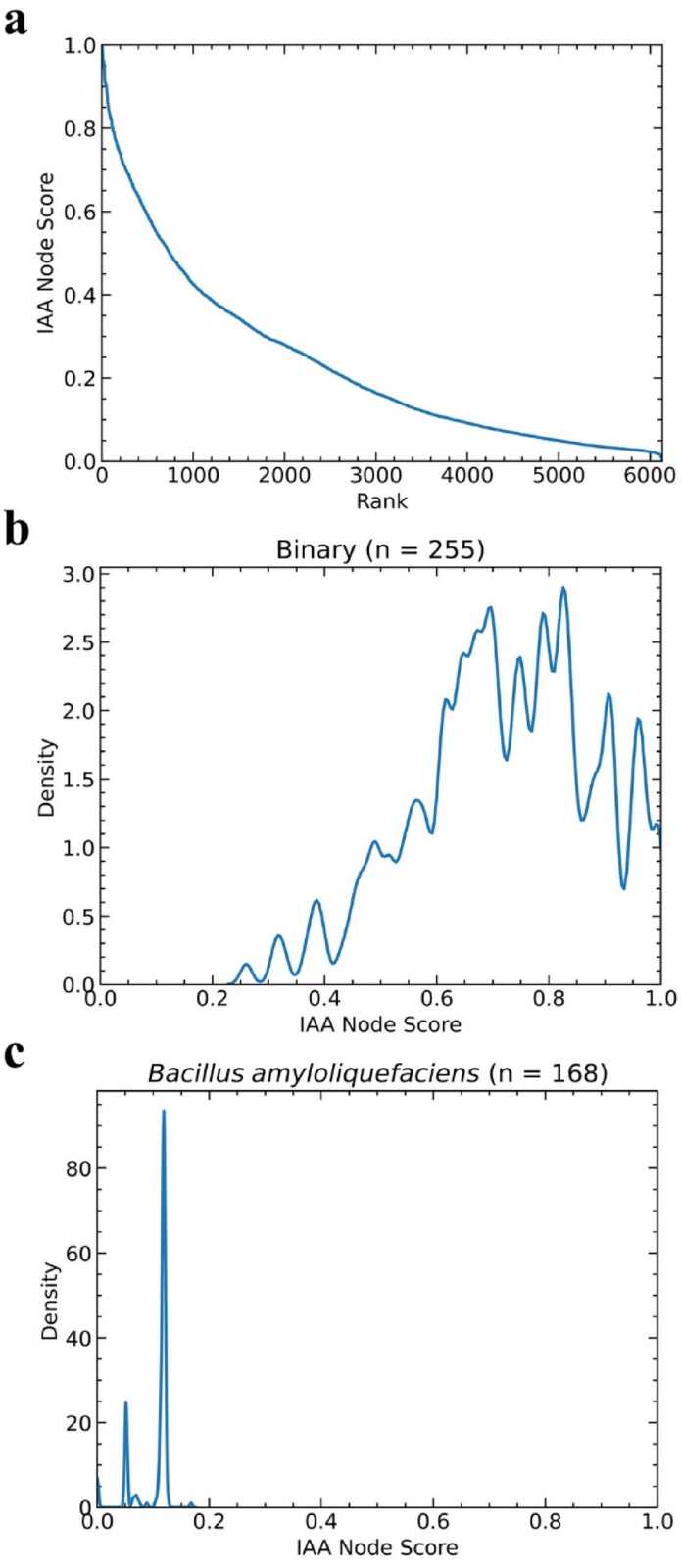


Fig. 5 IAA node distribution analysis using Logistic and Binary Models. **a** IAA node score distribution using the Logistic Model with $k = 1\ 0$. **b** Logistic IAA node score distribution of IAA-producing bacteria predicted by the Binary Model (mean ± sd = 0.723 ± 0.158, median = 0.731). **c** IAA node score distribution of 168 *Bacillus amyloliquefaciens* strains

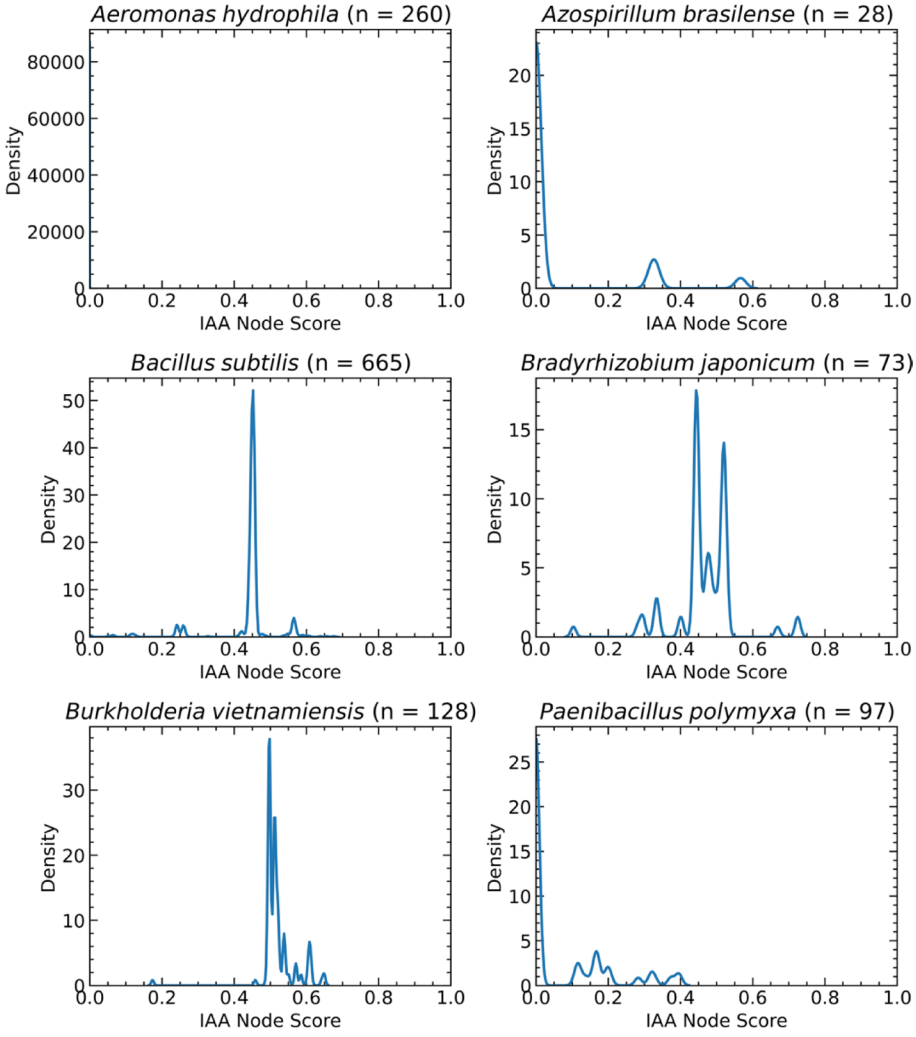


Fig. 6 IAA node score distribution of 6 species in the positive set

IAA production potential across bacterial species. Most of the *Aeromonas hydrophila*, *Azospirillum brasilense*, and *Paenibacillus polymyxa* strains showed little to no activity. On the other hand, *Bacillus subtilis*, *Bradyrhizobium japonicum*, and *Burkholderia vietnamiensis* showed moderate and consistent potential. *Acinetobacter* and *Rhizobium* generally had low activity and *Mesorhizobium* and *Streptomyces* had greater variability within the group. The pattern of the distribution observed within a species suggested that the production of IAA might be strain-specific.

Discussion

The genome sequences encapsulate the genetic information of an organism and the blueprints for the enzymes in the metabolic pathways. In this study, a probabilistic model framework was developed to connect the alignment hits (hit score), enzymes (edge score), and compounds (node score) in the Trp-dependent IAA pathways. The goal of this study is to estimate the likelihood of a bacterium producing IAA from four Trp-dependent pathways. Two models were constructed: the Binary Model aims to determine whether a bacterium has the potential to synthesize IAA from Trp with a binary output of 0 or 1. Additionally, the Logistic Model aims to rank and identify the bacteria

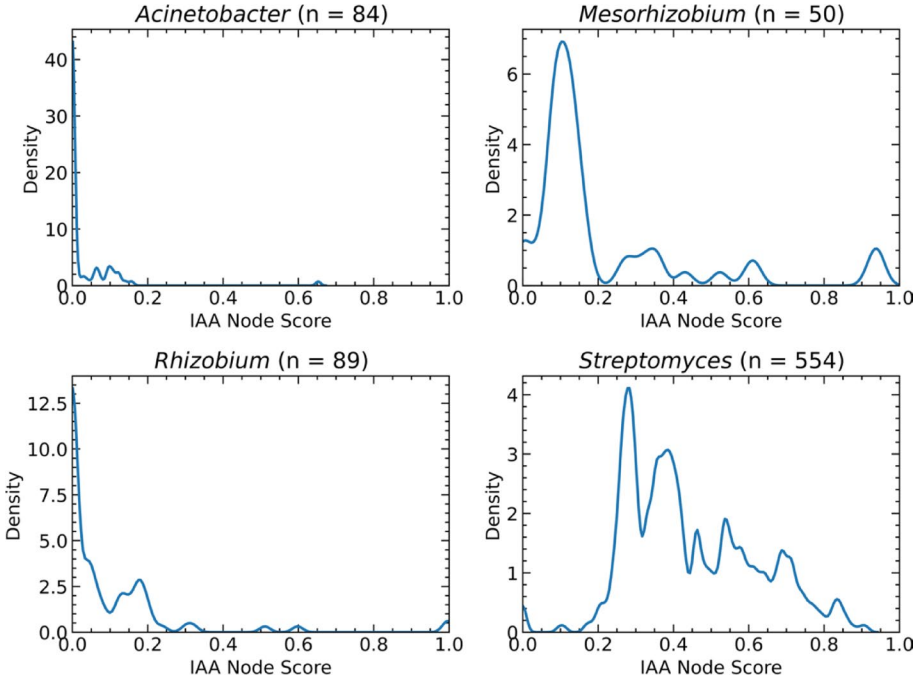


Fig. 7 IAA node score distribution of 4 genera in the positive set

Table 2 IAA node score of collected species and genera for validation

a. With alternative pathway		
Species	**IAA node score (mean ± sd)**	**Number of genomes**
Bacillus amyloliquefaciens	0.102 ± 0.033	168
b. Positive set		
Species	**IAA node score (mean ± sd)**	**Number of genomes**
Aeromonas hydrophila	0 ± 0	260
Azospirillum brasilense	0.055 ± 0.141	28
Bacillus subtilis	0.439 ± 0.082	665
Bradyrhizobium japonicum	0.466 ± 0.087	73
Burkholderia vietnamiensis	0.521 ± 0.049	128
Paenibacillus polymyxa	0.075 ± 0.118	97
Genus	**IAA node score (mean ± sd)**	**Number of genomes**
Acinetobacter	0.030 ± 0.080	84
Mesorhizobium	0.211 ± 0.230	50
Rhizobium	0.094 ± 0.174	89
Streptomyces	0.446 ± 0.173	554

with high potential to synthesize IAA from Trp from a collection of bacterial genomes. The source code of the prediction framework is freely available at https://github.com/ComputationalAgronomy/bcpip

Factors that affect prediction accuracy

The result from the Binary Model suggested that only 1.7% (255/15377) of the bacteria were capable of producing IAA, which is much lower than the previously reported 18.5% in 7282 genomes [30]. Several reasons could contribute to this discrepancy. Firstly, the difference in sequence identification and filtering criteria may play a role. An in-house script re-analysis of the previous study with more stringent criteria for the alignment

step and filtering based on percentage identity and coverage suggested that over 90% of the sequenced entities identified in the previous study would not meet these criteria. Secondly, the query sequence is searched against a custom curated IAA database containing only the proteins present in the pathway instead of the entire Swiss-Prot database. Query sequences against the Swiss-Prot database showed that some sequences previously aligned to IAA entries were found to align to proteins in the same family with similar functions but not precisely to the intended targets. Nevertheless, searching against the whole database can be computationally intensive, and a small-scale customized database allows rapid screening across a large number of genomes. Thirdly, sequence similarity is the primary method used for enzyme identification. A 40% similarity to the target sequence is sufficient to identify the protein family with confidence. However, to align with its functions, a 60% similarity is recommended [43]. Setting the identity threshold at 40% prevents underestimating the prediction results, with the potential trade-off of reducing the specificity. It is important to note that a bacterium that can be experimentally validated to produce IAA may only be predicted to have an enzyme in the IAA pathway if the enzyme has sufficient sequence similarity. However, a limitation is that sequence similarity does not always directly translate to functional similarity, as highly similar sequences can still miss key catalytic domains or signal peptides essential for enzyme function. A more robust approach would involve protein structure prediction, such as AlphaFold, and 3D structural comparison with known enzymes to determine functional similarities [45, 46]. This would be computationally intensive with sufficiently powerful GPUs required for structure prediction and comparison of every sequence, and represents a completely different algorithm for pathway prediction research. Additionally, IAA production is strain-specific; therefore, wet lab experiments are still required to validate the true capabilities of IAA synthesis [47, 48].

The location of IAA genes varies across bacterial species, with some genes residing on plasmids and others on chromosomes. These plasmid genes have the potential to facilitate HGT, which plays an important role in evolutionary history. Although including evolutionary history would improve prediction accuracy, one of the limitations is accurately modeling the stochastic nature of HGT events and their variable rates across different bacterial lineages, and robustly transferring and incorporating this information into the probabilistic pathway prediction models. Therefore, the current prediction model did not include this information, and this would be beyond the scope of the current project.

These models focused on integrating four major Trp-dependent pathways and excluded the minor IAAT pathway. This pathway is excluded because it is only utilized by a small group of bacteria and it is not the main focus of this project [44].

This study only focused on and performed validation using complete genomes, which may introduce a small bias when applied to real-world data. When the prediction model is applied to incomplete genomes or data with poor sequencing quality, users should expect increased false negative rates due to missing or fragmented gene sequences. There are multiple metrics that can assess the genome completeness and quality, such as BUSCO scores and N50 values, and these should be checked before applying the prediction model [49, 50]. Ultimately, users should interpret results with caution when working with draft assemblies or low-quality sequences.

Model assumptions

There are several assumptions for this prediction framework. This framework assumes that all enzymes and chemical compounds in the pathway are independent. However, this is an oversimplified assumption and generally not the case. Most enzymes have complicated regulatory and feedback systems and are regulated by other proteins or by the concentration of a compound. For instance, in *Azospirillum brasilense*, the expression of the *ipdC* gene, which encodes IPDC in Fig. 1, is activated by IAA [25]. Therefore, a non-independent prediction framework requires integrating the full convoluted network of interactions and relationships with all other enzymes instead of focusing solely on the core pathway. Additionally, the framework assumes that all pathways have equal weighting, an assumption that is expected to be verified but not examined in this study. The current framework provides a valuable tool that can be adapted for all pathways without resolving all underlying mechanisms involved in the network. A comprehensive prediction model with non-independent components and unequal weighting between pathways would be ideal. Nevertheless, this specific model would be dedicated to a group of organisms that have mapped out all interactions and feedback mechanisms, a task which is beyond the scope of this project.

The Logistic model extracts genomic information and assigns high scores to bacteria with multiple complete pathways and highly identical proteins to reported IAA pathway enzymes. However, a high score or rank in the Logistic Model does not guarantee a positive result. For example, a genome with a score of 0.960 in the Logistic Model was found to have a negative result from the Binary Model. Upon closer inspection of the data, the latter genome showed the presence of multiple pathways; however, none of these pathways surpassed the 40% identity threshold, resulting in a negative result in the Binary Model. The main advantage of the Logistic Model is its ability to identify genomes with multiple and less confident enzymes, whereas the Binary Model excludes such genomes based on the 40% identity threshold. In the Logistic Model, a bacterium with a high score suggests that either the bacterium possesses multiple Trp-dependent pathways or a single pathway with a very high confidence level. The model does not have the capability to distinguish between these two scenarios, and downstream wet lab validation is required to determine the exact mechanism of the synthesis processes. The primary purpose of the ranking is to act as in-silico filtering for the downstream large-scale wet lab screening process. This is an effective method to reduce the number of candidates required for biological assays. In practice, the selection of top N-ranked bacterial candidates should be determined based on the project goals and downstream analysis requirements. This is often based on resource availability, such as established wet lab procedures and the capacity of wet lab equipment (i.e., plate readers, incubators). Ultimately, project scope and available resources should guide candidate selection.

Evaluation of IAA synthesis via Trp-dependent pathways in bacteria

Experimental validation of bacterial IAA synthesis typically requires culturing and assaying bacteria in wet laboratory experiments [51]. However, the limited availability of bacterial genomes and the resource-intensive nature of large-scale wet lab screening procedures pose significant challenges to researchers in this field. Therefore, a bioinformatics-based approach presents a promising alternative strategy. By leveraging sequence similarities and investigating the homologous proteins involved in IAA synthesis across

a wide range of bacterial genomes, the prediction framework developed in this study offers valuable insights into the potential for IAA synthesis in individual bacterial species. Moreover, it can serve as an effective pre-screening tool and reduce the effort required for downstream resource-intensive wet lab experiments. We acknowledge that experimental validation of these predicted strains would be valuable future work, which represents an important opportunity for future investigation. For instance, if the prediction framework identified a group of bacteria with a high likelihood of synthesizing IAA that have not been reported previously, these bacteria would be potential candidates for further wet-lab validation. We also believe that computational prediction studies play an important role in generating reliable testable candidates for the research community.

Among the top 100 ranked bacteria in the prediction results, the four most abundant genera were *Paraburkholderia*, *Nonomuraea*, *Rhodococcus*, and *Nocardia*. Nevertheless, they were not in the positive set. Notably, the species *Paraburkholderia phytofirmans* has been reported to utilize IAA as an energy and carbon source [52]. Furthermore, Qin et al. [53] isolated a strain closely related to *Nonomuraea endophytica* and demonstrated its ability to promote plant growth, possibly through the activity of 1-aminocyclopropane-1-carboxylate (ACC) deaminase, an enzyme that degrades the precursor ACC of ethylene. The plant pathogen *Rhodococcus fascians* is able to excrete IAA and increase IAA levels in infected tissues, which may contribute to gall formation [54]. In addition, the *Nocardia* strain BMG51109 has been shown to promote root nodule formation and synthesize IAA [55]. Although previous studies on these genera have not primarily focused on their role in IAA synthesis, the prediction results suggest that some species within these genera may possess the potential to synthesize IAA. Consequently, additional research on these strains could unveil their contribution to plant growth promotion through IAA production and metabolism.

Conclusion

This study presents a probabilistic framework for estimating the potential of bacteria to synthesize the plant hormone IAA via Trp-dependent pathways. This framework integrates genomic data, sequence homology analysis, and probabilistic modeling to predict the likelihood of IAA synthesis capability. The Binary Model provides a binary prediction, while the Logistic Model ranks bacteria based on their relative potential for IAA synthesis. Validation against a comprehensive dataset of more than 15,000 bacterial genomes available on NCBI demonstrates the framework's efficacy in identifying promising candidates for further experimental validation in the wet lab. The proposed framework offers a cost-effective in-silico screening approach and enables targeted exploration of bacteria with potential IAA synthesis capabilities. This contribution has significant implications for advancing the understanding of plant–microbe interactions and the discovery of novel plant growth-promoting rhizobacteria. Furthermore, it provides a gateway for developing sustainable agricultural practices using novel biostimulants.

Abbreviations

IAA	Indole-3-acetic acid
Trp	Tryptophan
PGPR	Plant growth-promoting rhizobacteria
NCBI	National Center for Biotechnology Information
KEGG	Kyoto Encyclopedia of Genes and Genomes
DFS	Depth-first search
IAM	Indole-3-acetamide

IPA	Indole-3-pyruvate
TAM	Tryptamine
IAN	Indole-3-acetonitrile
IAAT	IAA acetyltransferase
IAOX	Indole-3-acetaldoxime
IAAld	Indole-3-acetaldehyde
ACC	1-Aminocyclopropane-1-carboxylate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06445-9.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Acknowledgements
The authors thank four anonymous reviewers for their constructive comments and valuable feedback, which significantly enhanced the quality of this manuscript.

Author contributions
Z.Y.: Methodology, Software, Data Curation, Validation, Writing. S.W.: Conceptualization, Methodology, Software, Validation, Writing.

Funding
This project is supported by the Ministry of Science and Technology, Taiwan, under Grant Number MOST 111–2313-B-002–007-MY3.

Data availability
All data generated or analysed during this study are publicly available. The source code for the implementation of this prediction framework is freely available on GitHub: https://github.com/ComputationalAgronomy/bcpip All the bacterial genomes used in this analysis are available on Zenodo: https://doi.org/10.5281/zenodo.19237602 The list of all the accession numbers and other information for the bacteria reference genomes is available in Additional File 2. The list of all EC numbers and keywords used for UniProtKB/Swiss-Prot entries are listed in Table S1 in Additional File 1 and Additional File 3.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 June 2024 / Accepted: 8 April 2026
Published online: 26 April 2026

References

1. Woodward AW. Auxin: regulation, action, and interaction. Ann Bot. 2005;95:707–35.
2. Zhao Y. Auxin biosynthesis and its role in plant development. Annu Rev Plant Biol. 2010;61:49–64.
3. Ludwig-Müller J. Auxin conjugates: their role for plant development and in the evolution of land plants. J Exp Bot. 2011;62:1757–73.
4. Ouyang J, Shao X, Li J. Indole-3-glycerol phosphate, a branchpoint of indole-3-acetic acid biosynthesis from the tryptophan biosynthetic pathway in *Arabidopsis thaliana*. Plant J. 2000;24:327–34.
5. Wright AD, Moehlenkamp CA, Perrot GH, Neuffer MG, Cone KC. The maize auxotrophic mutant orange pericarp is defective in duplicate genes for tryptophan synthase beta. Plant Cell. 1992;4:711–9.
6. Patten CL, Glick BR. Bacterial biosynthesis of indole-3-acetic acid. Can J Microbiol. 1996;42:207–20.
7. Escobar MA, Dandekar AM. *Agrobacterium tumefaciens* as an agent of disease. Trends Plant Sci. 2003;8:380–6.
8. Rodríguez-Moreno L, Jiménez AJ, Ramos C. Endopathogenic lifestyle of *Pseudomonas savastanoi* pv. *savastanoi* in olive knots. Microb Biotechnol. 2009;2:476–88.
9. Bhattacharyya PN, Jha DK. Plant growth-promoting rhizobacteria (PGPR): emergence in agriculture. World J Microbiol Biotechnol. 2011;28:1327–50.

10. Mayak S, Tirosh T, Glick BR. Effect of wild-type and mutant plant growth-promoting rhizobacteria on the rooting of mung bean cuttings. J Plant Growth Regul. 1999;18:49–53.
11. Patten CL, Glick BR. Role of *Pseudomonas putida* Indoleacetic Acid in development of the host plant root system. Appl Environ Microbiol. 2002;68:3795–801.
12. Lebrazi S, Niehaus K, Bednarz H, Fadil M, Chraibi M, Fikri-Benbrahim K. Screening and optimization of indole-3-acetic acid production and phosphate solubilization by rhizobacterial strains isolated from *Acacia cyanophylla* root nodules and their effects on its plant growth. J Genet Eng Biotechnol. 2020;18:71.
13. Etesami H, Alikhani HA, Hosseini HM. Indole-3-acetic acid (IAA) production trait, a useful screening to select endophytic and rhizosphere competent bacteria for rice growth promoting agents. MethodsX. 2015;2:72–8.
14. Spaepen S, Vanderleyden J, Remans R. Indole-3-acetic acid in microbial and microorganism-plant signaling. FEMS Microbiol Rev. 2007;31:425–48.
15. Comai L, Kosuge T. Involvement of plasmid deoxyribonucleic acid in indoleacetic acid synthesis in *Pseudomonas savastanoi*. J Bacteriol. 1980;143:950–7.
16. Theunis M, Kobayashi H, Broughton WJ, Prinsen E. Flavonoids, NodD1, NodD2, and Nod-Box NB15 modulate expression of the y4wEFG locus that is required for Indole-3-Acetic Acid synthesis in *Rhizobium* sp. strain NGR234. Mol Plant-Microbe Interact. 2004;17:1153–61.
17. Schröder G, Waffenschmidt S, Weiler EW, Schröder J. The T-region of Ti plasmids codes for an enzyme synthesizing indole-3-acetic acid. Eur J Biochem. 1984;138:387–91.
18. Baca BE, Soto-Urzua L, Xochihua-Corona YG, Cuervo-Garcia A. Characterization of two aromatic amino acid aminotransferases and production of indoleacetic acid in *Azospirillum strains*. Soil Biol Biochem. 1994;26:57–63.
19. Costacurta A, Keijers V, Vanderleyden J. Molecular cloning and sequence analysis of an *Azospirilium brasilense* indole-3-pyruvate decarboxylase gene. Mol Gen Genet. 1994;243:463–72.
20. Koga J, Adachi T, Hidaka H. Molecular cloning of the gene for indolepyruvate decarboxylase from *Enterobacter cloacae*. Mol Gen Genet. 1991;226–226:10–6.
21. Tang J, Li Y, Zhang L, Mu J, Jiang Y, Fu H, et al. Biosynthetic pathways and functions of Indole-3-Acetic Acid in microorganisms. Microorganisms. 2023;11:2077.
22. Howden AJM, Rico A, Mentlak T, Miguet L, Preston GM. *Pseudomonas syringae* pv. *syringae* B728a hydrolyses indole-3-acetonitrile to the plant hormone indole-3-acetic acid. Mol Plant Pathol. 2009;10:857–65.
23. Kobayashi M, Izui H, Nagasawa T, Yamada H. Nitrilase in biosynthesis of the plant hormone indole-3-acetic acid from indole-3-acetonitrile: cloning of the *Alcaligenes* gene and site-directed mutagenesis of cysteine residues. Proc Natl Acad Sci U S A. 1993;90:247–51.
24. Shao J, Li S, Zhang N, Cui X, Zhou X, Zhang G, et al. Analysis and cloning of the synthetic pathway of the phytohormone indole-3-acetic acid in the plant-beneficial *Bacillus amyloliquefaciens* SQR9. Microb Cell Fact. 2015;14:130.
25. Broek AV, Gysegom P, Ona O, Hendrickx N, Prinsen E, Van Impe J, et al. Transcriptional analysis of the *Azospirillum brasilense* Indole-3-Pyruvate Decarboxylase gene and identification of a cis-acting sequence involved in auxin responsive expression. Mol Plant-Microbe Interact. 2005;18:311–23.
26. Ramos C, Matas IM, Bardaji L, Aragón IM, Murillo J. *Pseudomonas savastanoi* pv. *savastanoi*: some like it knot. Mol Plant Pathol. 2012;13:998–1009.
27. Mashiguchi K, Hisano H, Takeda-Kamiya N, Takebayashi Y, Ariizumi T, Gao Y, et al. *Agrobacterium tumefaciens* enhances biosynthesis of two distinct auxins in the formation of crown galls. Plant Cell Physiol. 2019;60:29–37.
28. Aubin E, El Baidouri M, Panaud O. Horizontal gene transfers in plants. Life. 2021;11:857.
29. Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol Biol. 2011;11:276.
30. Zhang P, Jin T, Kumar Sahu S, Xu J, Shi Q, Liu H, et al. The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis. Molecules. 2019;24:1411.
31. Jin T, Wang Y, Huang Y, Xu J, Zhang P, Wang N, et al. Taxonomic structure and functional association of foxtail millet root microbiome. Gigascience. 2017;6:1–12.
32. Babalola OO. Beneficial bacteria of agricultural importance. Biotechnol Lett. 2010;32:1559–70.
33. Asghar H, Zahir Z, Arshad M, Khaliq A. Relationship between in vitro production of auxins by rhizobacteria and their growth-promoting activities in *Brassica juncea L*. Biol Fertil Soils. 2002;35:231–7.
34. Keswani C, Singh SP, Cueto L, García-Estrada C, Mezaache-Aichour S, Glare TR, et al. Auxins of microbial origin and their use in agriculture. Appl Microbiol Biotechnol. 2020;104:8549–65.
35. Khalid A, Tahir S, Arshad M, Zahir ZA. Relative efficiency of rhizobacteria for auxin biosynthesis in rhizosphere and non-rhizosphere soils. Soil Res. 2004;42:921.
36. Huddedar SB, Shete AM, Tilekar JN, Gore SD, Dhavale DD, Chopade BA. Isolation, characterization, and plasmid pUPI126-mediated indole-3-acetic acid production in *Acinetobacter* strains from rhizosphere of wheat. Appl Biochem Biotechnol. 2002;102–103:21–39.
37. Chen F, Wang M, Zheng Y, Luo J, Yang X, Wang X. Quantitative changes of plant defense enzymes and phytohormone in biocontrol of cucumber *Fusarium* wilt by *Bacillus subtilis* B579. World J Microbiol Biotechnol. 2010;26:675–84.
38. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res. 2021;49:D1020–8.
39. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11(1):119.
40. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
41. Bateman A, Martin M-J, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2022;51:D523–31.
42. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2022;51:D587–92.
43. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003;333:863–82.

44. Idris EE, Iglesias DJ, Talon M, Borriss R. Tryptophan-dependent production of Indole-3-Acetic Acid (IAA) affects level of plant growth promotion by *Bacillus amyloliquefaciens* FZB42. Mol Plant-Microbe Interact. 2007;20:619–26.
45. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
46. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: an overview of protein structure prediction. Front Bioinforma. 2023;3:1120370.
47. Pliego C, Cazorla FM, González-Sánchez MÁ, Pérez-Jiménez RM, de Vicente A, Ramos C. Selection for biocontrol bacteria antagonistic toward *Rosellinia necatrix* by enrichment of competitive avocado root tip colonizers. Res Microbiol. 2007;158:463–70.
48. Sachdev DP, Chaudhari HG, Kasture VM, Dhavale DD, Chopade BA. Isolation and characterization of indole acetic acid (IAA) producing *Klebsiella pneumoniae* strains from rhizosphere of wheat (*Triticum aestivum*) and their effect on plant growth. Indian J Exp Biol. 2009;47:993–1000.
49. Jauhal AA, Newcomb RD. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour. 2021;21:1416–21.
50. Wang P, Wang F. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 2023;39:175–86.
51. Mohite B. Isolation and characterization of indole acetic acid (IAA) producing bacteria from rhizospheric soil and its effect on plant growth. J Soil Sci Plant Nutr. 2013;13(3):638–49.
52. Laird TS, Flores N, Leveau JHJ. Bacterial catabolism of indole-3-acetic acid. Appl Microbiol Biotechnol. 2020;104:9535–50.
53. Qin S, Miao Q, Feng W-W, Wang Y, Zhu X, Xing K, et al. Biodiversity and plant growth promoting traits of culturable endophytic actinobacteria associated with *Jatropha curcas* L. growing in Panxi dry-hot valley soil. Appl Soil Ecol. 2015;93:47–55.
54. Vandeputte O, Öden S, Mol A, Vereecke D, Goethals K, El Jaziri M, et al. Biosynthesis of auxin by the gram-positive phytopathogen Rhodococcus fascians is controlled by compounds specific to infected plant tissues. Appl Environ Microbiol. 2005;71(3):1169–77.
55. Ghodhbane-Gtari F, Nouioui I, Hezbri K, Lundstedt E, D'Angelo T, McNutt Z, et al. The plant-growth-promoting actinobacteria of the genus *Nocardia* induces root nodule formation in *Casuarina glauca*. Antonie Van Leeuwenhoek. 2018;112:75–90.

Publisher's Note