From sequence to molecules: Feature sequence-based genome mining uncovers the hidden diversity of bacterial siderophore pathways

Shaohua Gu[1,2*], Yuanzhe Shao[2*], Karoline Rehm[3], Laurent Bigler[3], Di Zhang[1], Ruolin He[1], Ruichen Xu[4], Jiqi Shao[1], Alexandre Jousset[5], Ville-Petri Friman[6], Xiaoying Bian[7], Zhong Wei[5#], Rolf Kümmerli[8#], Zhiyuan Li[1,2#]

[1] Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[2] Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China

[3] University of Zurich, Department of Chemistry, Winterthurerstr. 190, 8057 Zurich, Switzerland

[4] School of Life Science, Shandong University, Qingdao, Shandong 266237, China

[5] Jiangsu Provincial Key Lab for Organic Solid Waste Utilization, Key lab of organic-based fertilizers of China, Nanjing Agricultural University, Nanjing, P R China

[6] University of Helsinki, Department of Microbiology, 00014, Helsinki, Finland

[7] Helmholtz International Lab for Anti-infectives, State Key Laboratory of Microbial Technology, Shandong University, Qingdao, Shandong 266237, China

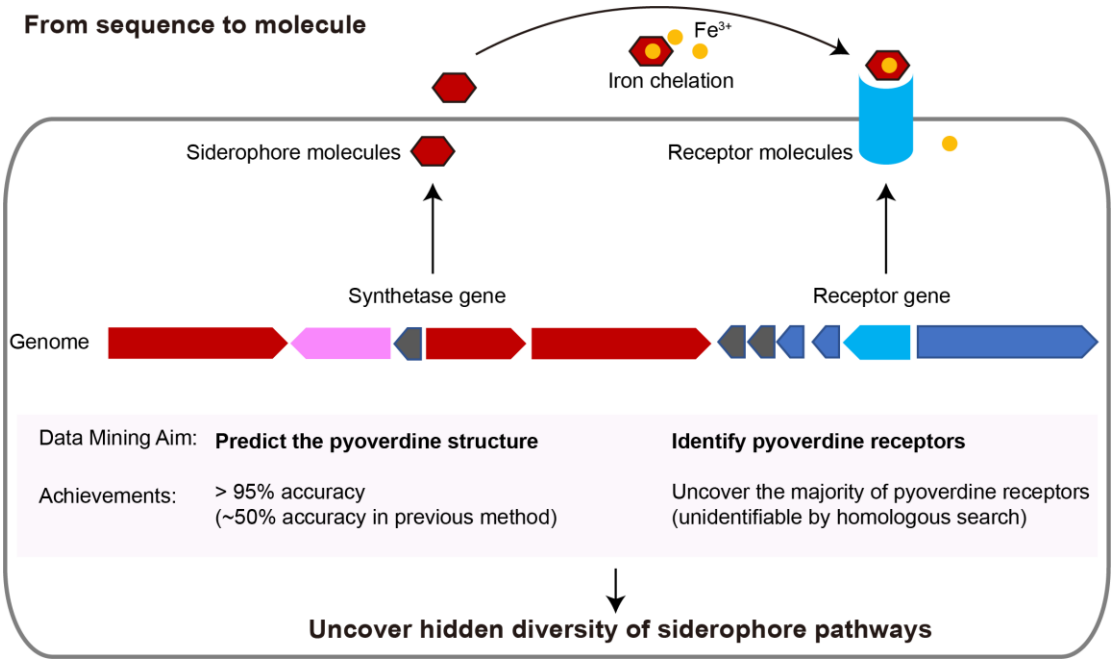
[8] University of Zurich, Department of Quantitative Biomedicine, Winterthurerstr. 190, 8057 Zurich, Switzerland

[*] These authors contributed equally to this article.

[#]Corresponding authors (email: weizhong@njau.edu.cn; rolf.kuemmerli@uzh.ch; zhiyuanli@pku.edu.cn)

Abstract (146 words count)

Microbial secondary metabolites are a rich source for pharmaceutical discoveries and play crucial ecological functions. While tools exist to identify secondary metabolite clusters in genomes, precise sequence-to-function mapping remains challenging because neither function nor substrate specificity of synthesis enzymes can accurately be predicted. Here we developed a knowledge-guided bioinformatic pipeline to solve these issues. We analyzed 1928 genomes of *Pseudomonas* bacteria and focused on iron-scavenging pyoverdines as model metabolites. Our pipeline predicted 188 chemically different pyoverdines with nearly 100% structural accuracy and the presence of 94 distinct receptor groups required for the uptake of iron-loaded pyoverdines. Our pipeline unveils an enormous yet overlooked diversity of siderophores (151 new structures) and receptors (91 new groups). Our approach, combining feature sequence with phylogenetic approaches, is extendable to other metabolites and microbial genera, and thus emerges as powerful tool to reconstruct bacterial secondary metabolism pathways based on sequence data.



Introduction

Rapid advancements in sequencing technologies have revolutionized our view on microbial communities. While amplicon sequencing provides information on community composition and diversity, shotgun and whole genome sequencing allow us to reliably anticipate evolutionary and ecological relationships between microbes and to obtain functional information on communities. Computational models assessing the metabolic capacity of individual members, or an entire consortium, have become very popular and powerful[1-3]. The major focus of such modelling approaches is typically on the primary metabolism of bacteria, as genes involved in core metabolic pathways are highly conserved and can be identified with relative ease[2,4]. Conversely, analysis of the secondary metabolites has attracted less attention, even though they include compounds such as antibiotics, toxins, siderophores, biosurfactants, all known to have important implications for microbial community assembly[5,6] and to be important sources for pharmaceutical discoveries[7-9] .

There are multiple challenges that currently prevent a detailed unravelling of secondary metabolism of bacteria based on genome data[10-12]. First, most secondary metabolites are produced by pathways comprised of modular enzymes such as non-ribosomal peptide synthetases (NRPSs) or polyketide synthases (PKS)[13,14]. Locating complete synthesis clusters and identifying all enzyme-encoding genes is challenging from highly fragmented metagenomic sequences or draft genomes with a high number of contigs. Second, functional predictions for coding regions within a cluster rely on homologous comparisons with experimentally characterized genes. Such information is often restricted to a limited number of model organisms, meaning that only a small portion of the existing secondary metabolism pathways

is covered by current data bases. Finally, given the complex multi-modular synthesis machineries, it is challenging to precisely predict the secondary metabolites produced even with accurately annotated NRPS or PKS clusters. The main challenge is that a large pool of non-proteinogenic amino acids is used as substrates and the specificity of an enzyme's A domain, connecting these unusual amino acids, is often poorly understood[15]. As a result, new computational methods are needed to accurately reconstruct bacterial secondary metabolism from sequence data.

Here, we present a new bioinformatic pipeline that overcomes these challenges. We specifically focus on a particular class of secondary metabolites (iron-scavenging siderophores) as a case study to develop a bioinformatic workflow that predicts the chemical structure of the produced metabolites with nearly 100% accuracy. Our pipeline is based on improved gene annotation combined with a phylogeny- and feature sequence-based substrate prediction techniques (Figure 1). In comparison with the currently available databases and bioinformatic tools[16], the main advancement of our method is the more accurate prediction of synthesized products based on NRPS clusters identified in genome data. In addition, we show that our workflow is capable of correcting previous mis-annotations and is extendable to other secondary metabolites (e.g., toxins, biosurfactants, virulence factors) and microbial genera.

Among siderophores, we focus on NRPS machineries that are responsible for the synthesis of pyoverdines, a class of chemically diverse siderophores with high iron affinity, produced by *Pseudomonas* bacteria[17,18]. While each *Pseudomonas* strain produces a single type of pyoverdine, an enormous structural diversity has been described across strains and species[19-22]. Pyoverdine types differ in their peptide backbone, meaning that the diversity

83 should be mirrored in NRPS enzyme diversity and their selectivity for the different amino acid

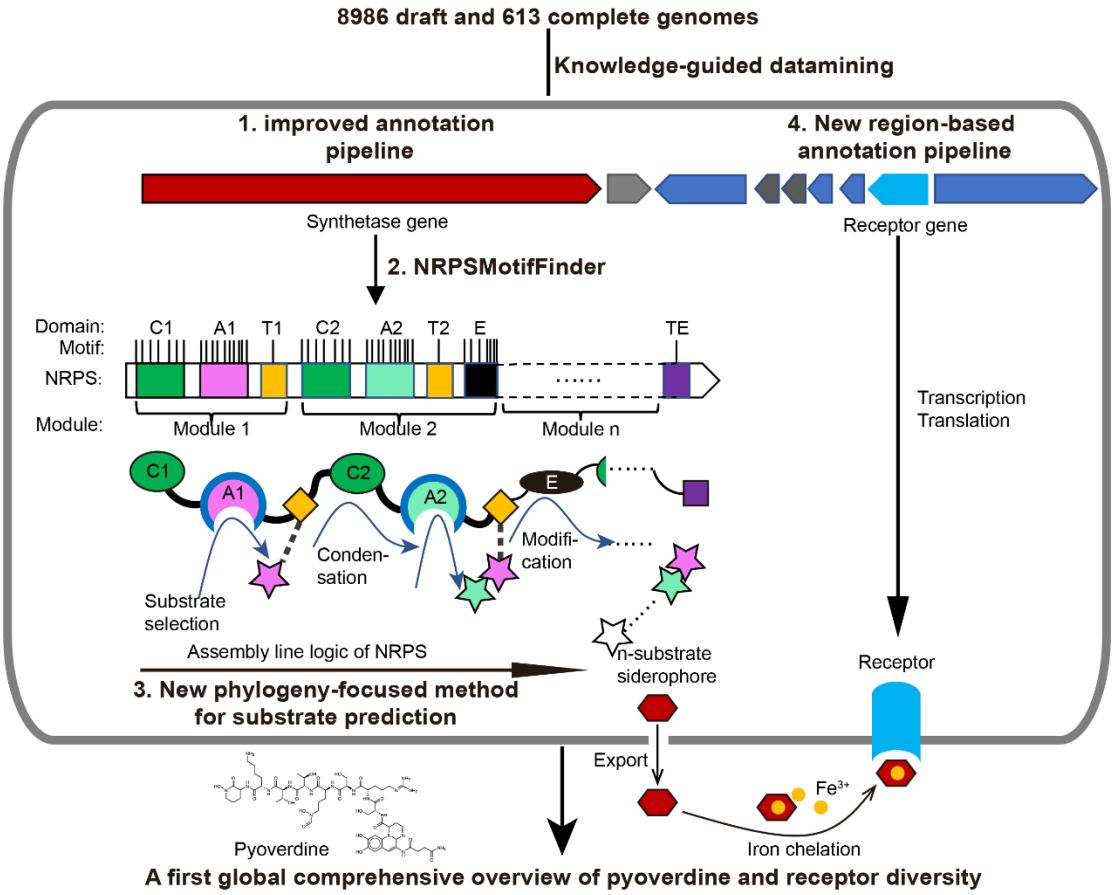
84 substrates[18]. Based on this knowledge, our pipeline entails the following steps (Figure 1): (i)

85 identification of the complete sequences of pyoverdine synthetase genes from fragmented draft

86 genomes, (ii) building the pyoverdine synthesis machinery *in silico* by extracting the feature

87 sequences for substrate specificity from motif-standardized NRPSs, and (iii) predicting the

88 precise chemical structure of pyoverdines followed by empirical verification.



89

90 **Figure 1 Scheme depicting our new genome mining pipeline to precisely predict the**
91 **synthesis, the molecular structure and the uptake machinery of pyoverdines, a family of**
92 **iron-scavenging siderophores produced by members of the *Pseudomonas* genus.** The
93 grey rounded outer rectangle represents a bacterial cell. The red and blue arrow-shaped boxes stand for
94 the synthetase and receptor genes for pyoverdines, respectively. Synthetase genes are transcribed and
95 translated to form the *n*-modular NRPS enzymes. These enzymes synthesize the peptide backbone of
96 pyoverdine through an assembly line using their repeating module units, with the A domain being
97 responsible for substrate selection and the E domain for chirality. The *n*-substrate siderophores are then
98 exported to the extracellular space for iron chelation. Membrane-embedded TonB-dependent receptors
99 recognize the ferri-siderophore complex and import it into the cell. Bold black text and black arrows

describe our multi-step computational methods developed to reconstruct the entire process from genome sequence data. First, the annotation pipeline was improved (from antiSMASH[23]) to extract the complete sequence of pyoverdine synthetase genes from draft genomes. Second, NRPSMotifFinder was used to define A- and E-domains and to determine the exact motif-intermotif structure of the pyoverdine assembly line. Third, intermotif regions most indicative of substrate specificity were used to develop a phylogeny-focused method for precise product prediction. Fourth, a sequence-region-based annotation method was combined with genome architecture features to identify the FpvA, receptors responsible for ferri-pyoverdine import.

An additional element of iron metabolism is that when siderophores are secreted and bound to iron, bacteria rely on a specific receptor for their uptake into the cell. Pyoverdine receptors are annotated as FpvA and it is known that receptor diversity matches pyoverdine diversity[22,24]. Moreover, FpvA belongs to the family of TonB-dependent receptors and a single *Pseudomonas* species often has many gene copies encoding these receptors. This poses an additional bioinformatic challenge: how to find the gene encoding the specific pyoverdine receptor among several potential receptor genes? To overcome this, we develop an algorithm that focuses on sequence regions involved in pyoverdine recognition and translocation across the outer membrane with supervised learning methods that locate the *fpvA* genes in the fragmented genomes based on these regions (Figure 1). Not only did our method unveil a substantial number of previously unrecognized pyoverdine receptors, but it also revealed that receptors with the potential to uptake pyoverdine are distributed across a broad spectrum of bacteria. This suggests the possibility of pyoverdine serving as another 'shared language' among microbes. Altogether, our bioinformatic pipeline uses knowledge-guided insights empowered by supervised learning to construct a first systematic sequence-to-function mapping of a family of secondary metabolites (pyoverdine) and their corresponding receptors. Our analysis unveils a yet unrecognized extraordinary diversity of iron-scavenging machineries in pseudomonads.

Results

Section 1: Improved annotation pipeline reveals a vast reservoir of pyoverdine synthetase genes

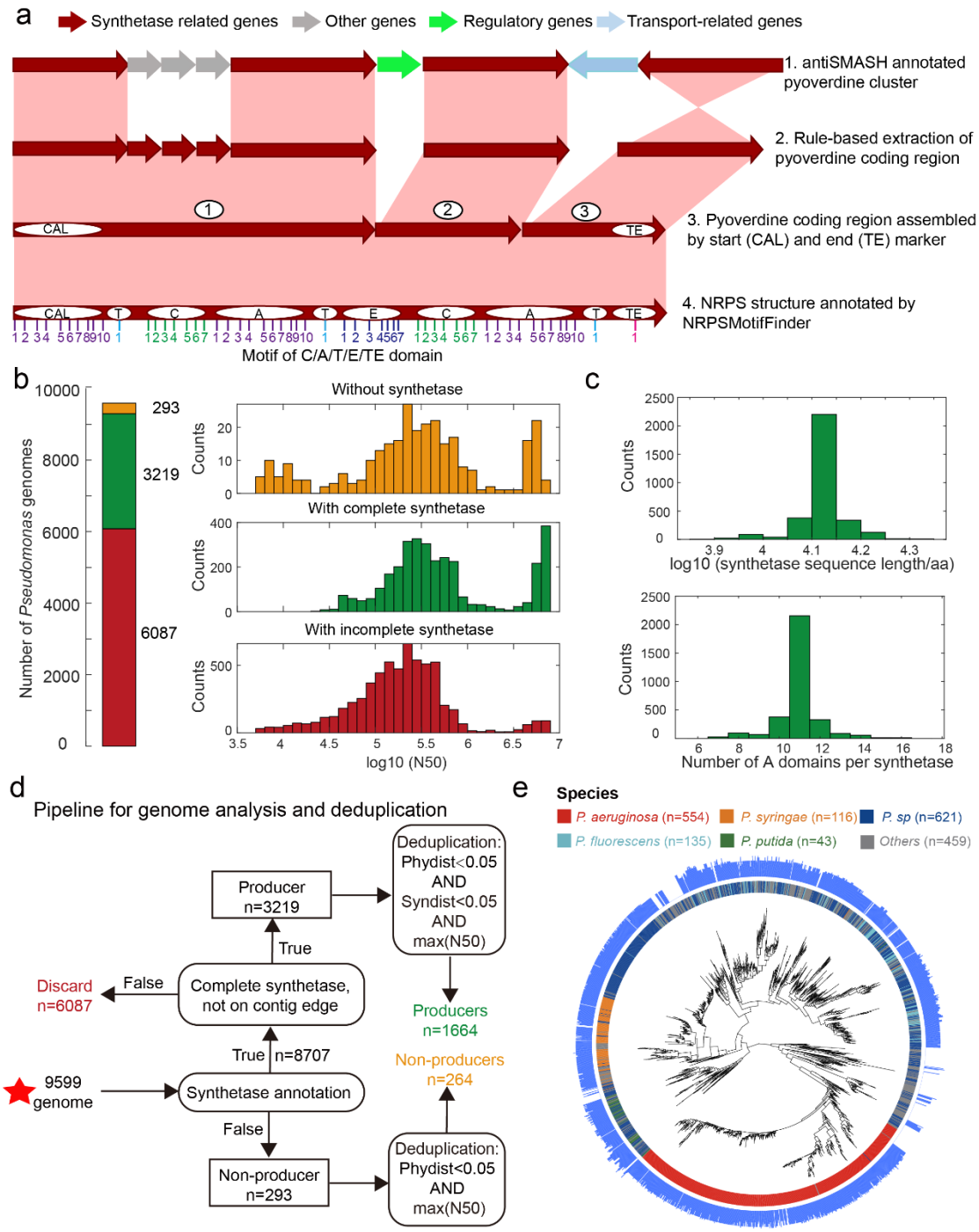


Figure 2 Improved annotation pipeline reveals a vast diversity of pyoverdine synthetase genes. a. Improved annotation pipeline based on the raw annotation from antiSMASH[23]. **b.** The

annotation pipeline was applied to 9599 *Pseudomonas* genomes (94% draft genomes). Genomes could be separated into three categories. Yellow: genomes without pyoverdine cluster. Green: genomes with a complete pyoverdine cluster. Red: genomes with incomplete pyoverdine synthetase cluster. The red category involved genomes with truly incomplete clusters (lacking Flu or Pep synthetic genes) or genomes with likely truncated synthetic genes at the edge of contigs. **c.** Distributions of the sequence length (upper panel) and the number of A domains (lower panel) across all the genomes with a complete synthetase cluster. **d.** Workflow applied to separate the 9599 *Pseudomonas* genomes into the three categories described in b and removing of redundant genomes with high phylogenic similarity and showing high similarity in pyoverdine synthetases. Red star indicates the start of the workflow. **e.** Phylogenetic tree depicting the relationship among the 1928 non-redundant *Pseudomonas* strains (1664 producers and 264 non-producers) based on the concatenated alignment of 400 single-copy conserved genes in their genomes. The inner ring depicts the taxonomical classification including the four most prevalent species. The outer ring shows the number of A domains present in the pyoverdine synthetase assembly line in each strain.

The first step of our bioinformatic pipeline was to improve the annotation of pyoverdine synthetase genes. The pyoverdine molecules is composed of a conserved fluorescent chromophore (Flu) and a peptide chain (Pep), which are both synthesized by NRPS enzymes[25]. There are already existing tools, such as antiSMASH that can find and annotate NRPS clusters in microbial genomes[23]. However, antiSMASH (and other popular annotation platforms[26,27]) rely on accurate gene predictions, which are typically problematic for fragmented genomes. Consequently, while antiSMASH can recognize and annotate certain genes of an NRPS cluster, the precise reconstruction of a complete NRPS assembly line often fails. This is particularly problematic because most available genomes are drafts and any analysis suffers from the unavoidable issue of incomplete or misannotation of gene fragments.

To overcome these issues, we developed an improved four-step annotation pipeline starting with the raw annotation of the pyoverdine cluster obtained from antiSMASH[23] (Figure 2a). First, we implemented a NRPS Hidden Markov Model (HMM) to re-annotate and extract the entire nucleotide sequence of the pyoverdine synthetase cluster[28], including the genes missed by antiSMASH. For this step, the nucleotide sequences were converted into amino acid

164 sequences to avoid erroneous gene predictions typically associated with antiSMASH. Second,
165 we assembled the entire re-annotated pyoverdine coding region into a single sequence with a
166 defined start (CAL) and/or end (TE) markers. Third, we used NRPSMotifFinder to identify the
167 C, A, T, E and TE motifs that are characteristic for the NRPS structure of pyoverdine[15]. Finally,
168 we applied a safety measure to ensure that the recovered NRPS assembly line is complete
169 (contains both Flu and Pep) and is not truncated, which can occur when a synthetase coding
170 region is at the edge of a contig. Consequently, we dismissed all pyoverdine synthesis clusters
171 located within 100 bp proximity to contigs' edges and either lacked Flu or Pep synthetic genes.
172 Next, we applied our improved pyoverdine synthesis annotation pipeline to 9599
173 *Pseudomonas* genomes (including 613 complete and 8986 draft genomes) retrieved from the
174 *Pseudomonas* Genome Database[29]. We found the pyoverdine synthesis machinery in 97% of
175 the genomes (Figure 2b), indicating that the machinery is ubiquitous in *Pseudomonas*. However,
176 since 94% of the analyzed genomes were in draft form, the pyoverdine synthesis machinery
177 was likely truncated (i.e., on the edge of the contig) in 63.4% (6087) of the genomes. These
178 genomes were excluded from further analysis. Around 3.1% of retained genomes (293) with
179 high assembly completeness were missing pyoverdine synthetic genes, indicating that these
180 *Pseudomonas* strains were not able to produce pyoverdine ('non-producers'). The rest of the
181 genomes (33.5%; 3219 genomes) were classified as 'producers' with complete pyoverdine
182 NRPS assembly lines that meet all our quality controls. For these 3219 genomes, we used
183 NRPSMotifFinder to find boundaries between the various synthesis domains and to determine
184 amino acid length and the number of A domains. The lengths of pyoverdine synthetic genes
185 ranged between 7690 and 21333 amino acids, and the number of A domains per synthetase

186 ranged between 6 and 17, with a total of 35,281 A domains being present across all strains
187 (Figure 2c and Figure S1). Overall, our analysis pipeline unveiled a vast diversity of pyoverdine
188 synthetase that goes far beyond of what has previously been described in the literature.

189 Finally, we conducted a phylogenetic analysis based on 400 conserved genes with the 293
190 non-producers and the 3219 producers. We first removed redundant non-producers by
191 retaining the most integrative genome among strains with high phylogenic similarity. Then, we
192 removed redundant producers by retaining the most integrative genome among strains with
193 high phylogenic and pyoverdine synthetase similarity (Figure 2d). This data cleaning yielded a
194 total of 1928 *Pseudomonas* strains (403 complete and 1525 incomplete genomes), segregating
195 into 1664 pyoverdine producers and 264 non-producers. The phylogenetic tree revealed that
196 all major *Pseudomonas* species clades were present in our data set (Figure 2e). Moreover, the
197 number of A domains varied widely among species and even between strains within species.
198 For example, the number of *Pseudomonas aeruginosa* A domains ranges between 7 and 14.
199 In summary, by improving the synthetase annotation method, we successfully obtained 1664
200 highly reliable pyoverdine synthetases (with a total of 18,292 A domains) and 264 non-
201 producers.

202

203 **Section 2: Phylogeny-focused substrate prediction for pyoverdine A domains**

204 Our next goal was to precisely predict the molecular structure of the pyoverdines produced by
205 the 1664 strains with complete synthetase gene clusters. The first essential step towards this
206 goal was to reliably predict the substrate selectivity of all A domains in the NRPS assembly line.
207 The A domain of each module selects for a single substrate among 22 proteinogenic and

208 hundreds of non-proteinogenic amino acids[30,31]. Moreover, whenever an E domain exists
209 downstream of an A domain, the chirality of the amino acid incorporated into the peptide chain
210 gets modified from L to D. Thus, the modularity combined with the selectivity of A domains can
211 promote an enormous diversity of pyoverdine molecule structures. To date, 73 pyoverdine
212 structures have been reported (Supplementary_table1) out of which 13 have their synthetase
213 genes sequenced (Supplementary_table2). In order to make reliable predictions, two
214 challenges must be addressed: (i) the extraction of relevant information from A domain
215 sequences for which the substrate is known, and (ii) the effective application of this information
216 to predict specificity of A domain sequences for which the substrate is unknown.

217 To address the first challenge, we built our analysis on the NRPS assembly lines of the
218 known 13 pyoverdines to extract relevant information from the A domain sequences. From this
219 dataset, we could identify 101 A domains that could be experimentally linked to 13 amino acid
220 substrates (Supplementary_table3). We next performed multisequence alignment of the 101 A
221 domains to determine the “feature sequence distance”, which is the most informative for the
222 substrate selectivity. To this end, we tested three different A domain regions, three different
223 sequence similarity measurements and seven different clustering methods for their predictive
224 power (Figure 3a). We found that the full A domain sequence is not informative for substrate
225 prediction (Figure 3b, left panel). Instead, our analysis indicated that information-rich positions
226 start with motif A4 and end before motif A5, consistent with the known role of the A domain
227 pocket in substrate selectivity[15]. Overall, the sequence region from motifs A4 to A5 (termed
228 “Amotif4-5”), in conjunction with Jukes-Cantor distance and Ward linkage clustering, performed
229 best in accurately distinguishing between different substrates and maintaining homogeneity for

230 identical substrates (Figure 3b). Utilizing these parameters, we conducted a comprehensive
231 analysis of all A domains and their respective substrate information within the *Pseudomonas*
232 NRPS biosynthetic gene cluster available in the MIBIG database[32] (Supplementary_table4). To
233 our delight, our method can find and correct the incorrect substrate information corresponding
234 to A domain in the MIBIG database (Figure S2 and Supplementary_table5). This demonstrates
235 the effectiveness of performing substrate predictions for the A domain based on the selected
236 parameter combination.

237 To address the second challenge, we developed a “phylogeny-focused method” to apply
238 the feature sequence distance derived in the preceding paragraph to the 18,292 discovered A
239 domains. We realized that a direct construction of a phylogenetic tree including all 18,292 query
240 A domains and the 101 reference A domains would be computationally too demanding and
241 impossible to scale up. Furthermore, such an approach would result in phylogeny-interference
242 issues, where domains would cluster not only based on their substrate similarities but also
243 based on overall species relatedness[15]. To minimize the effect of phylogeny and speed up
244 calculation, we took each of the 18,292 query A domains and identified the two most similar A
245 domain clusters within the 101 reference A domain set. We then compared the feature distance
246 between each query A domain and the two most similar reference A domains in different clusters
247 and assigned the query A domain to a substrate specificity using the following rules (Figure 3c).
248 (1) If the feature distance is below the 0.7 threshold (corresponding to 50% identity) for only
249 one of the two reference A domains, then the substrate of the query A domain is matched to the
250 substrate of the more similar (lower distance) reference A domain. (2a) If the feature distance
251 is below the 0.7 threshold for both reference A domains, then we considered the relative

difference of the query A domain towards the two reference A domains. If the relative difference is larger than 0.2, the query A domain is matched to the substrate of the more similar reference A domain. (2b) If the relative difference is smaller than 0.2, the substrate of the query A domain cannot unambiguously be determined and is thus matched with both reference substrates. (3) If the feature distance is above the 0.7 threshold (below 50% identity) for both reference A domains, then the substrate of the query A domain is marked as “unknown”. For most query A domains, rule (1) could be applied (17880 cases), whereas rules (2) and (3) had to be used rarely (133 and 279 cases, respectively). We applied our methodology termed “phylogeny-focused method” to all following substrate and pyoverdine structure predictions.

Section 3: Experimental validation of the annotation and prediction pipeline

We tested whether our bioinformatic pipeline can reliably predict the structure of a set of yet uncharacterized pyoverdines. To achieve this objective, we selected 20 *Pseudomonas* strains, all known to produce pyoverdines, from a natural strain collection that was previously isolated from soil and water[33]. We sequenced their genomes and subsequently applied our annotation and prediction pipeline to generate predicted pyoverdine structures for all 20 strains harboring a total of 237 A domains. Notably, none of the predicted structures matched any of the 13 reference pyoverdine structures, and 9 out of the 20 structures were predicted to be novel (not yet characterized) pyoverdines. To verify our predictions, we elucidated the chemical structure of the 20 pyoverdines using culture-based methods combined with UHPLC-HR-MS/MS[34].

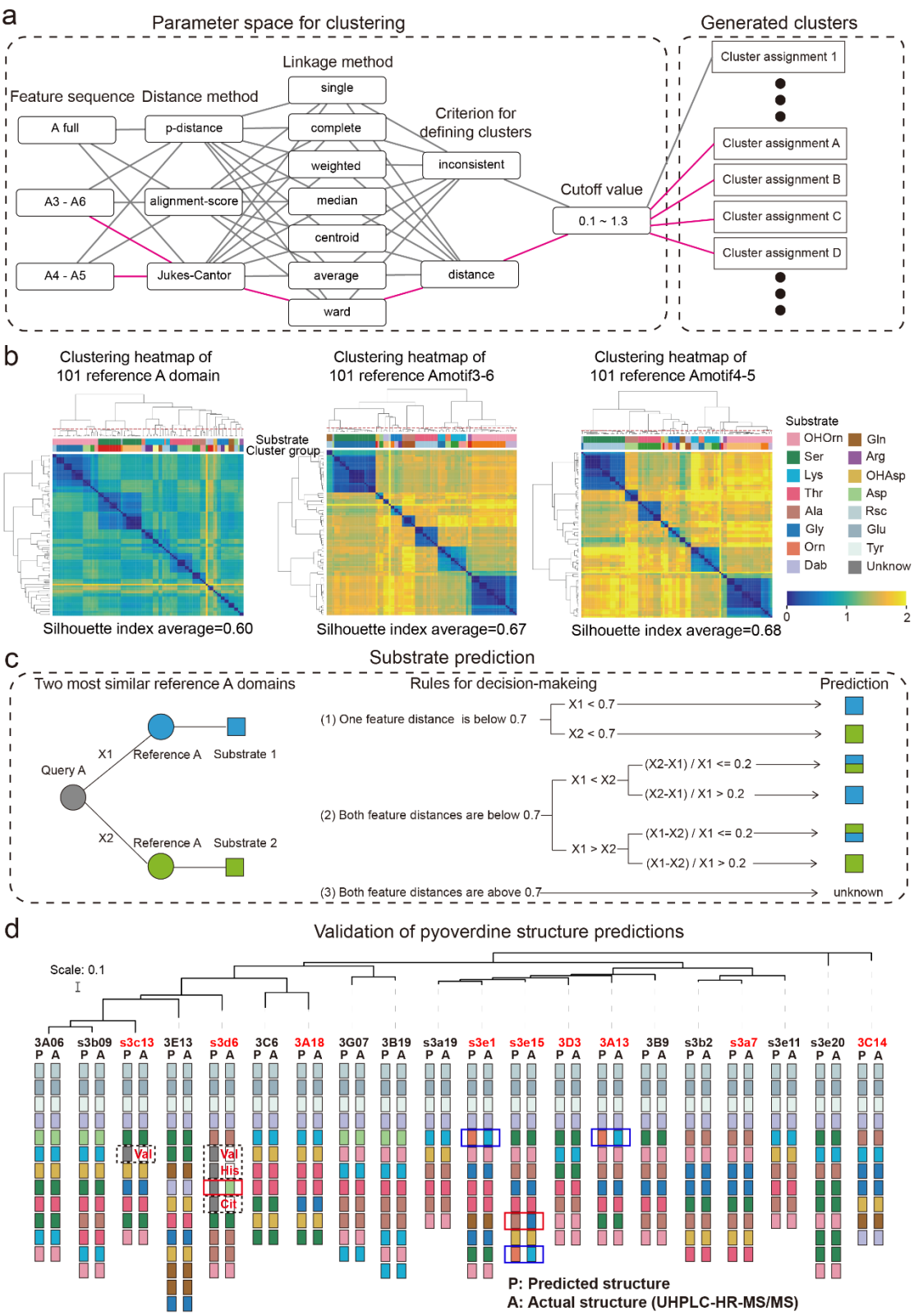


Figure 3 Phylogeny-focused substrate prediction for pyoverdine synthetase assembly lines. a. Information from 101 reference A domains with known amino acid substrates were used to develop an algorithm that predicts substrates from A domain sequence data with high accuracy. The challenge is to group the variable A domains into clusters that predict the same substrate (captured by the silhouette index). To find the most distinctive algorithm, we combined different feature sequences of A

domains (Amotif) with different distance and linkage methods in our hierarchical clustering analyses. The best performing path is shown in pink. **b.** Heatmap showing the hierarchically clustered distances of the 101 reference A domains as a function of the feature sequence used. Left panel: complete A domain sequences. Middle panel: Amotif3-6 sequences. Right panel: Amotif4-5 sequences. The experimentally validated substrates are shown on top of the heatmaps. The heatmaps show that hierarchical clustering, reliably associating sequence distances with substrate, worked best with the Amotif4-5. **c.** Phylogeny-focused substrate prediction pipeline for query A domains (grey circle) based on Amotif4-5 feature sequence comparisons. X1 and X2 represent the feature distance between the query A domain and two closest reference A domains (blue and green circles), respectively. Three rules are used, based on the feature distances X1 and X2 and a threshold value of 0.7 (50% similarity), to make substrate predictions for the query A domain. There are three possible outcomes: unambiguous substrate prediction (blue or green squares), ambiguous substrate prediction (dual-colored squares), and no prediction ("unknown). **d.** Phylogenetic tree of 20 *Pseudomonas* strains and visualization of their predicted and actual pyoverdine structures to validate our phylogeny-focused substrate prediction pipeline. Strains marked in red font indicate that the siderophore structures they generate were predicted to be novel (not yet characterized) pyoverdines. 228 out of the 237 substrates (96.2%) were correctly predicted. The nine inconsistencies are boxed in blue (Lysine and Ornithine are indistinguishable), in dashed black (correct detection of "unknown" substrates), and in red (true mismatches). Note that our prediction pipeline (as any other pipeline) cannot distinguish between modified variants of the same amino acid.

We found a near-perfect match (96.2%) between the predicted and the observed pyoverdine chemical structures and were able to accurately assign amino acids in 228 out of 237 cases (Figure 3d). Our method demonstrated a substantial improvement comparing to the prediction accuracy of antiSMASH[16] in pyoverdines (46.0%), which could accurately assign correct amino acids only in 109 out of 237 cases (Supplementary_table6). The nine non-matching cases in our analysis segregated into three groups. In three cases (1.3%), our algorithm could not distinguish between the substrates Lysine and Ornithine, as these two amino acids are highly similar both in terms of their chemical structures and corresponding A domain sequences. This is the only sensitivity issue that is associated with our approach. In four cases (1.7%), our technique assigned an "unknown" substrate to amino acids that turned out to be valine, citrulline and histidine. Indeed, these three amino acids have not been reported in pyoverdines before and are therefore not yet present in the reference dataset. These cases

310 show that our analysis pipeline can be used to identify new substrates. Once experimentally
311 verified, the new A domains and their substrates can expand the reference dataset, allowing
312 targeted improvement of our phylogeny-focused prediction technique. Finally, there were only
313 two cases (0.8%) that represented true mismatches between observed and predicted amino
314 acids. Altogether, our phylogeny-focused method is highly accurate in predicting pyoverdine
315 peptide structures and in identifying unknown substrates in *Pseudomonas*.

316 To assess the extendibility of our pipeline, we chose *Burkholderiales* as a test case,
317 leveraging a manually curated dataset that we have accumulated from literatures (see Method
318 for detail). This dataset contains 203 A domain sequences and corresponding substrate
319 information for 34 NRPS metabolites (e.g., toxins, biosurfactants, virulence factors) across 7
320 genera of *Burkholderiales* (Supplementary_table7). We also integrated experimentally
321 validated biosynthetic gene cluster data from *Burkholderiales* in MIBIG 3.0[32]
322 (Supplementary_table8) with our manual dataset. Upon analyzing these datasets, we observed
323 that our pipeline consistently maintains high prediction accuracy within *Burkholderiales* (83%
324 vs. 67% in antiSMASH[16], Figure S3 and Supplementary_table9). This outcome underscores
325 the robustness of our developed prediction algorithm, highlighting its potential extension to
326 other NRPS secondary metabolites and microbial genera.

327

328 **Section 4: Application of the annotation and prediction pipelines to a full dataset**

329 After successful validation, we applied our bioinformatic pipeline to the 1664 complete NRPS
330 assembly lines annotated in our genome analysis (Figure 2). Across all assembly lines, we
331 were able to predict the substrates of 17,880 A domains (97.75%) without ambiguity, whereas

133 A domains (0.73%) were associated with two different substrates, and 279 A domains (1.52%) predicted an unknown substrate (similar to the case of valine above). After considering the presence/absence of an E domain in each module, we derived the structure of 1664 pyoverdines according to method at section 2 (Figure 4).

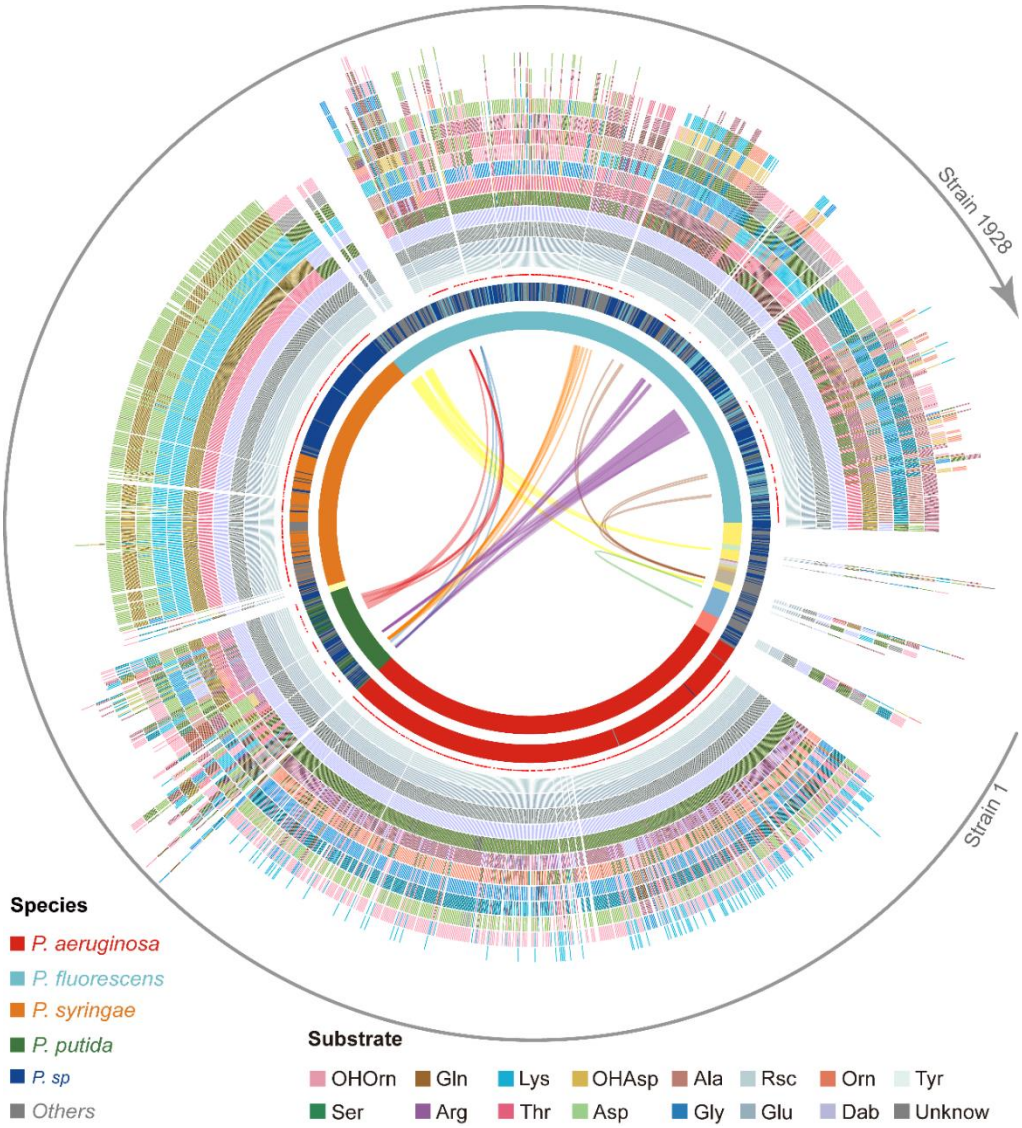


Figure 4 Predicted pyoverdine structural diversity based on our developed algorithm mapped onto the phylogenetic tree comprising all 1928 (non-redundant) *Pseudomonas* strains. The stacked boxes in the outermost circle show the predicted structure of pyoverdines, whereby each color represents a specific amino acid substrate. Strains without boxes represent non-producers (n = 264). Boxes with two colors indicate cases of ambiguous (dual) substrate prediction. The red dots at the basis of the stacked boxes indicate experimentally validated pyoverdine structures. The inner circle shows the taxonomic species classification following Figure 2e. Because the allocation of strains to species names is often imprecise, we divided the 1928 strains by their phylogenetic distance into 18 clades (color shadings in inner-most circle), out of which 13 contained more than 1 strain. Lines within the inner-most

346 circle link strains from different clades that share the same pyoverdine structures, whereby line colors
347 represent the shared unique pyoverdines. The bending of the lines represents the phylogenetic sequence
348 distances of the connected strain pairs.
349

350 Our prediction yielded 188 different pyoverdine molecules, out of which only 37 structures
351 had been previously reported. While these 37 reported structures were highly abundant across
352 strains (1103 out of 1664), our pipeline was powerful in identifying many of the rarer pyoverdine
353 variants. Agreeing with previous studies, we observed that the fluorophore is highly conserved
354 among the 188 predicted structures. Moreover, our analysis confirmed that 13 amino acid
355 substrates form the core of all the 188 pyoverdine structures, with most of the variation being
356 attributable to different substrate combinations, peptide lengths, and substrate chirality (Figure
357 4). Notably, pyoverdine structural diversity was not strongly linked to phylogeny because the
358 same pyoverdine structure could be found in completely unrelated species, while closely related
359 species often had different pyoverdine structures (Figure 4). These observations suggest that
360 there may be both frequent recombination and horizontal gene transfer of pyoverdine
361 synthetase clusters between species. Taken together, the bioinformatics methods developed in
362 our study can predict a suit of secondary metabolites (pyoverdines) from sequence data with
363 high accuracy, revealing an unprecedented richness and evolutionary history of siderophores
364 within pseudomonads and the discovery of 151 novel pyoverdine candidate variants.
365

366 **Section 5: Development of a region-based identification method for annotation of the**
367 **FpvA receptors**

368 In pseudomonads, iron-loaded pyoverdines are recognized by FpvA, a TonB-dependent
369 receptor, that transports the ferri-siderophore into the periplasm[19,35-37] . The protein structure of
370 characterized FpvA variants consists of three domains: The Secretin and TonB N-terminus short

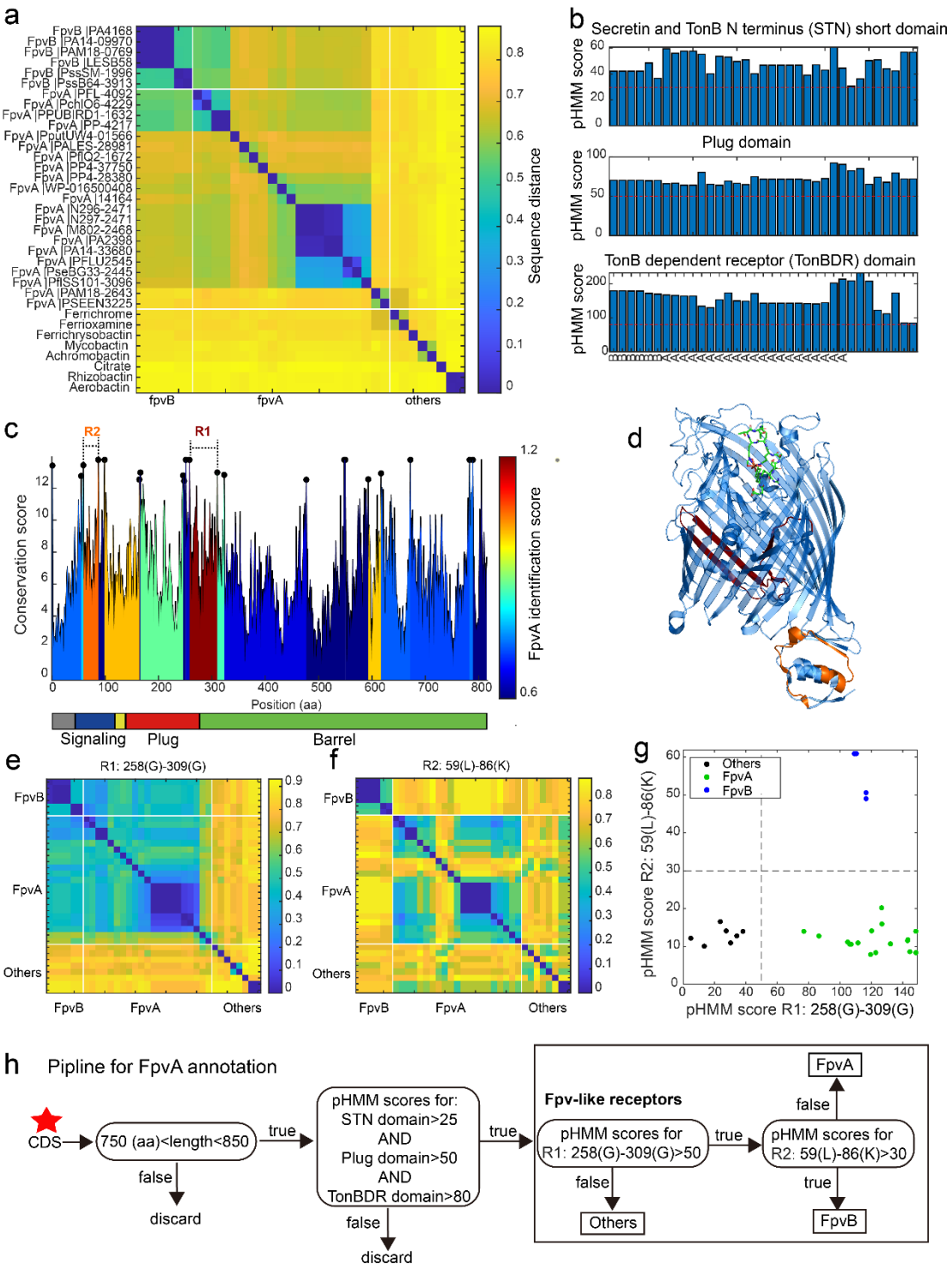
371 domain (STN), the Plug domain (Plug), and the TonB dependent receptor domain (TonB)[19].
372 While these domains are conserved across FpvA variants and other siderophore receptors,
373 there is substantial variation at the sequence level. This makes it challenging to reliably identify
374 FpvA receptors from sequence data by homologous search. As an example, we were unable
375 to find FpvA genes (with a 60% identity threshold) by homologous search in several genomes
376 although they had complete pyoverdine synthesis machineries. Moreover, there are many other
377 TonB-dependent receptors with fairly high sequence identity to FpvA but that transport other
378 siderophores than pyoverdine (e.g. FpvB, 55% identity, transporting pyoverdine, ferrichrome
379 and ferrioxamine B[38]. Therefore, it is imperative to develop a new comprehensive method for
380 identifying FpvA receptors in *Pseudomonas* genomes.

381 We started our approach by comparing the sequences of 35 reported siderophore
382 receptors, including 21 FpvA, 6 FpvB, and 8 TonB-dependent siderophore receptor sequences
383 often found in *Pseudomonas* genomes, encoding receptors for the uptake of heterologous
384 siderophores (Supplementary_table10). We found that all receptor sequences share a similar
385 length of around 800 amino acids (FpvA and FpvB sequences: 809 ± 10 amino acids). We then
386 used the complete sequences to calculate the pair-wise distances by global alignment before
387 applying hierarchical clustering (Figure 5a). We found substantial divergence between FpvA
388 variants to an extent that was comparable to the distance between FpvA and other siderophore
389 receptors. Moreover, FpvB variants clustered with FpvA variants, showing that FpvA
390 identification based on full sequence distances is unachievable. We hence focused on the three
391 typical receptor domains (TonB, Plug, and STN, retrieved from the Pfam database) and applied
392 Profile Hidden Markov Models (pHMM) to calculate the pHMM probability scores for each

393 domain and reference sequence. The probability scores (calculated as the log-odd ratios for

394 emission probabilities and log probabilities for state transitions) had reasonably high scores but

395 no distinction was apparent between the three receptor classes (Figure 5b).



396

397 **Figure 5 A sequence-region-based identification pipeline for annotating FpvA receptors.**

398 **a.** Heatmap displaying the hierarchically clustered sequence distances (p-distance calculation method,

identity (%) = (1-sequence distance) * 100) of 35 reference siderophore receptors identified in *Pseudomonas* spp., based on full sequences. No clear discrimination between FpvA, FpvB and other receptors is possible. The order of receptors is consistent across panels (**b**), (**e**), and (**f**). **b.** The pHMM scores of the three standard receptor domains (STN, Plug, and TonBDR) vary across the 35 reference sequences (A: FpvA, B: FpvB and NA: others), but do not allow to distinguish between receptor groups. **c.** FpvA region-based conservation scores from a multi-alignment of the 35 reference sequences mapped to the FpvA sequence of strain *P. aeruginosa* PAO1. All residues within the top 10% of the conservation score denoted with black dots. For each region flanked by two black dots, we calculated the FpvA identification score (heatmap), representing the ability to distinguish FpvA from non-FpvA receptors. **d.** Mapping of the two regions with the highest FpvA identification scores R1(dark red) and R2 (orange) to the crystal structure of FpvA from PAO1 conjugated with pyoverdine (PDB 2IAH). **e.** Heatmap showing the hierarchically clustered sequence distances of 35 reference siderophore receptors based on the R1 sequence region. A clear discrimination between FpvA/FpvB and other receptors emerges. **f.** Heatmap showing the hierarchically clustered sequence distances of 35 reference siderophore receptors based on the R2 sequence region. A clear discrimination between FpvA and FpvB receptors emerges. **g.** The pHMM scores of regions R1 and R2 for the 35 siderophore reference receptors are contrasted against each other, yielding a clear separation between FpvA, FpvB and other receptors. Dashed lines indicate the pHMM threshold scores used for later analysis. **f.** Flowchart showing all steps involved in the FpvA annotation from genome sequence data. The red star indicates the start of the workflow.

We next asked whether there are specific regions within the receptor sequences that are characteristic of FpvA. To address this, we conducted a multiple sequence alignment (MSA) with all 35 reference receptor sequences and mapped them onto the sequence of the well-characterized FpvA of *P. aeruginosa* PAO1 (Figure 5c). MSA allows to identify conserved sites (Figure 5c, black dots representing the top 10% most conserved sites) that are shared by the majority of the reference sequences. We then used these conserved sites to partition the MSA into variable regions which were flanked by two conserved sites. For each variable region, we assessed its predictive power to differentiate FpvA from non-FpvA sequences. For this we defined the "FpvA identification score" analogous to the intercluster-vs-intracluster Calinski-Harabasz variance ratio, as

$$I_{\mathrm{FpvA}} = d_{\mathrm{A:non}}/d_{\mathrm{A:A}}$$

431 where $d_{\mathrm{A:A}}$ is the sequence distance among all 21 FpvA sequences, and $d_{\mathrm{A:non}}$ is the
432 sequence distance between all 21 FpvA and the 14 non-FpvA sequences.

433 Our analysis yielded two locations with noticeably high FpvA identification scores (Figure
434 5c). The region with the highest FpvA identification score (referred to as R1) locates at the
435 intersection of the Plug domain and the barrel structure of the TonB domain (Figure 5d, between
436 258 Gly and 309 Gly in the PAO1 FpvA). According to the sequence distance matrix, the R1
437 region allows to distinguish heterologous siderophore receptors from FpvA and FpvB receptors
438 (Figure 5e). The region with the second highest FpvA identification score (referred to as R2)
439 was located in the C-terminal signaling domain (Figure 5d, between 59 Leu and 86 Lys in the
440 PAO1 FpvA). The sequence distance matrix revealed that R2 allows to distinguish FpvB from
441 FpvA receptors (Figure 5f).

442 We then constructed two pHMM by (i) the alignment of the 21 FpvA sequences in the R1
443 region, termed R1(FpvA), and (ii) the alignment of the 6 FpvB sequences in the R2 region,
444 termed R2(FpvB). Running R1(FpvA) and R2(FpvB) against all 35 reference sequences
445 revealed a clear separation between the three receptor categories (Figure 5g). Along the
446 R1(FpvA) axis, FpvA and FpvB reference sequences have high R1 scores (minimal score 77.0)
447 that separate them from other siderophore receptors (maximal score 38.1), whereas FpvAs
448 references have substantially lower R2 scores (maximal 20.2) than FpvBs (minimal 49.0) along
449 the R2(FpvB) axis.

450 Based on these insights, we developed a decision flow chart for annotating FpvAs in
451 *Pseudomonas* genomes (Figure 5h): First, we considered sequences as Fpv-like receptors that
452 share similar properties to the ones identified in our reference database. Particularly, protein

453 coding sequence (CDS) length has to be between 750 and 850 amino acids and the phMM
454 scores for the three typical receptor domains STN, Plug, and TonB have to be greater than 25,
455 50, and 80, respectively (Figure 5b, red dashed lines). Second, we used the pHMM threshold
456 scores obtained for R1(FpvA) and R2(FpvB) (Figure 5g) to differentiate other siderophore
457 receptors (R1(FpvA) score < 50) from FpvB receptors (R1(FpvA) score > 50 and R2(FpvB)
458 score > 30) and FpvA receptors (R1(FpvA) score > 50 and R2(FpvB) score < 30). Our method
459 effectively identifies FpvA receptors from sequence data and can be readily applied to the entire
460 *Pseudomonas* dataset.
461
462 **Section 6: Application of the receptor annotation pipeline to the full dataset.**
463 The region-based receptor identification pipeline was applied to all 1928 *Pseudomonas*
464 genomes. The analysis identified 4547 FpvAs, 615 FpvBs, and 9139 other TonB-dependent
465 Fpv-like receptors across the dataset (Figure 6a). The 4547 FpvA sequences clustered
466 hierarchically into 114 groups, defined by an identity threshold of 60%. When comparing to the
467 21 reference FpvAs (Figure 6b), we found that 2293 FpvA sequences have close homologues
468 in the reference data base, while 2254 FpvA sequences lack such close homologues (sequence
469 identity < 50%). These latter sequences, termed as "dissimilar to reference", may represent
470 novel subtypes of FpvA receptors that could not be found by simple homology search. Our
471 analysis further shows that many strains have more than one FpvA receptor.
472 We then asked whether the 4547 FpvAs are found in proximity of pyoverdine Pep
473 synthetase genes as it is commonly the case for cognate FpvA receptors [39]. We thus calculated
474 the proximity between pyoverdine Pep genes and the Fpv-like receptor genes by counting the

475 number of base pairs between the two coding regions. All TonB-dependent receptors that have
476 not been classified as FpvAs were more than 20 kb away from the Pep genes (Figure 6c). In
477 contrast, 92% of the nearest FpvA genes were indeed located within 20 kb of their pyoverdine
478 Pep genes (Figure 6d, called proximate receptors). These proximate receptors encompassed
479 both those with close (66%) and more distant (34%) resemblance to the reference receptor
480 types. Overall, this proximity analysis confirmed that our region-based gene identification
481 method can reliably identify FpvA receptors.

482 We next explored the diversity among FpvA receptors in more detail by focusing on the
483 1534 strains that had proximate-receptors within 20 kb of the pyoverdine Pep genes (Figure 6d)
484 and using high-confidence FpvAs for sequence feature extraction. When considering the whole
485 gene sequences, these receptors segregated into 44 groups according to single-linkage
486 clustering with an identity threshold of 60% (Figure S4a). To investigate which sequence
487 regions were the most informative for reliable clustering, we used a similar approach as with
488 FpvAs detection by quantifying the "group identification score" for variable regions flanked by
489 highly conserved sites. The higher the score, the stronger a region's capacity to discriminate
490 between FpvA groups. We found that the four regions with the top discrimination capacities all
491 located near the Plug domain surrounding the pyoverdine transmission channel (Figure 6e).
492 The plug domain is known to undergo conformational changes and is involved in pyoverdine
493 selectivity and import[40,41], suggesting that the four high-score regions are responsible for
494 pyoverdine specificity.

495

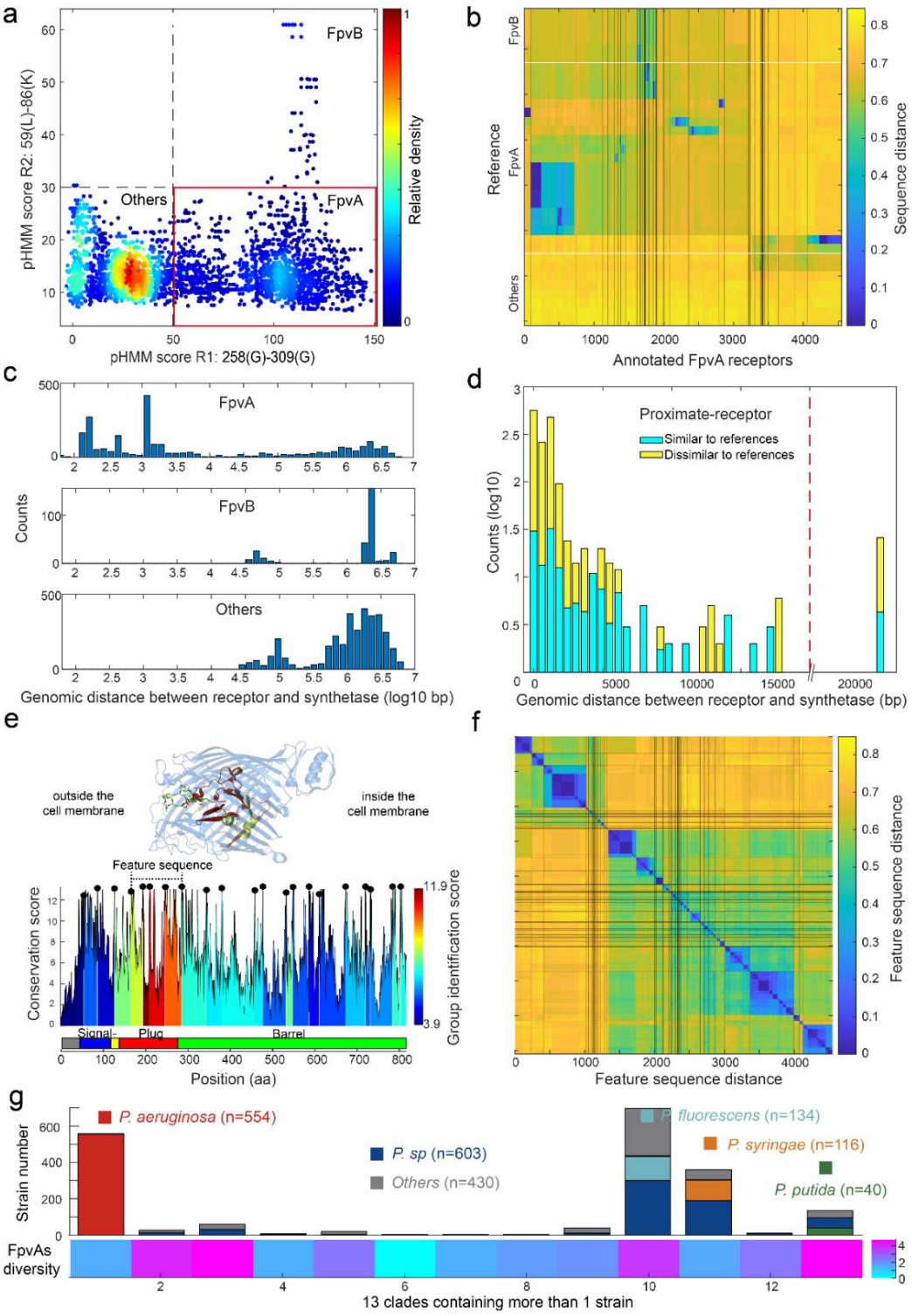


Figure 6 Application of the receptor annotation pipeline to the full database. a. Applying the receptor annotation pipeline to the genomes of the 1928 non-redundant *Pseudomonas* strains yields 14301 Fpv-like receptors, which segregate into 4547 FpvA receptors (red box), 615 FpvB receptors, and 9139 other receptors, based on the pHMM score thresholds for regions R1 and R2. The heatmap indicates receptor density. **b.** Sequence distance matrix between the 35 reference sequences (y-axis) and the 4547 annotated FpvA sequences in the full database (x-axis). Database sequences were ordered by hierarchically clustering and segregated into 114 groups. 2254 of the annotated FpvA sequences have

sequence identity < 60% compared to the reference receptors, pointing at novel subtypes of FpvA receptors. **c.** Genomic distance (in base pairs) between each Fpv-like receptor sequence and its pyoverdine peptide synthetase gene (Pep) for annotated FpvA receptors (upper panel), FpvB receptors (middle panel) and other receptors (lower panel). **d.** Distribution of the genomic distance between each FpvA receptor and its nearest pyoverdine peptide synthetase depending on whether the annotated FpvA receptor has high sequency similarity (blue, ≥ 50%) or low sequence similarity (yellow, < 50%) with at least one of the 21 reference FpvAs. **e.** FpvA region-based conservation scores from a multi-alignment of all the annotated FpvA receptors that are proximate (< 20 kbp) to the pyoverdine synthetase cluster mapped to the FpvA sequence of strain *P. aeruginosa* PAO1. All residues within the top 10% of the conservation score denoted with black dots. For each region flanked by two black dots, we calculated the group identification score (heatmap, lower panel), representing the ability of the region to distinguish between different groups of FpvA receptors. Four regions in the plug domains had a particularly high group identification score (called the feature sequence). They are mapped to the crystal structure of FpvA from PAO1 conjugated with pyoverdine (PDB 2IAH, up panel). All four regions surround the pyoverdine transmission channel and are shown in the respective heatmap color. **f.** Heatmap showing the hierarchically clustered distances between the 4547 annotated FpvA receptors based on the feature sequence (comprising the four groups with the highest identification scores). The analysis identifies 94 receptor groups with a 70% identity threshold. **e.** The diversity of FpvA receptors along the 13 phylogeny clades containing more than 1 strain. Receptor diversity was calculated by the Shannon entropy, similar to the alpha-diversity in microbial community.

Based on the above insights, we concatenated the four high-score regions (from 168 Pro to 295 Ala in PAO1) into a single "feature sequence". The feature sequence could characterize 98% of the distance matrix compared to the whole sequence (1534 FpvAs, r=0.98, Figure S4a-b) and substantially reduced within-group distance. We applied the concatenated feature sequence approach to all the 4547 annotated FpvAs to calculate the sequence distance matrix. Single-linkage clustering with an identity threshold of 70% revealed a total of 94 groups, out of which 43 groups contained more than 10 members (Figure 6f). The diversity of receptors is hence much larger than currently anticipated as only 3 groups of FpvAs have previously been reported. Finally, we calculated the diversity of receptor FpvAs for each of the 13 phylogenetic clades with more than one strain by the Shannon entropy, which is similar to the alpha-diversity in microbial community (Figure 6g). We noticed large differences in FpvA diversity across the clades and species: clades with *P. aeruginosa* and *P. syringae* species had lower FpvAs

diversity (1.55 and 1.60) than clades containing *P. putida* and *P. fluorescens* species (4.82 and 3.77). Taken together, the region-based identification method developed in our study can reliably mine the FpvAs (pyoverdines receptors) from genome data, revealing undiscovered diversity of FpvA pyoverdine receptors that are unequally distributed across the different phylogenetic clades of pseudomonads.

We further extended the search into the entire bacterial world, by applying the same pipeline in Figure 5h. Among the >30,000 complete bacterial genomes in NCBI, 31,936 receptors pass the R1 and R2 thresholds (Figure S5). To our surprise, these FpvAs are dispersed across 12,944 strains spanning 13 phyla, 22 classes, 54 orders, 139 families, and 468 genera, encompassing 2,598 species (Figure S6 and Supplementary_table11). While *Pseudomonas* remains the genus with the highest distribution of FpvA, other genera with notable distributions comprise *Escherichia*, *Salmonella*, *Enterobacter*, *Acinetobacter*, *Xanthomonas*, *Bordetella*, and *Achromobacter*. This study marks the first systematic exploration of FpvA's distribution throughout the bacterial kingdom. The extensive distribution may implie that pyoverdine serves as another 'shared language' in the bacterial community and plays a more important role than previously thought.

Discussion

The rapid expansion of sequencing data offers exciting opportunities for microbiology[42-44]. One key challenge of current research in the field is to infer biological functions of microbial communities from genome sequence data[45-47]. While this endeavor is increasingly successful for biological functions involving the primary metabolism and the associated complex metabolic

559 flux, reconstructing aspects of the secondary metabolism is much more challenging. The main
560 issue is that neither the function of a secondary metabolite enzyme nor the resulting metabolite
561 can be precisely predicted from gene sequence data. In our study, we tackled this challenge
562 and developed a bioinformatic pipeline to reconstruct the complete secondary metabolism
563 pathway of pyoverdines, a class of iron-scavenging siderophores produced by *Pseudomonas*
564 spp. These secondary metabolites are synthetized by a series of non-ribosomal peptide
565 synthetases and require a specific receptor (FpvA) for uptake. We combined knowledge-guided
566 learning with phylogeny-based methods to predict with high accuracy: (i) the full pyoverdine
567 assembly line, (ii) the substrate specificity for each enzyme within the assembly lines, (iii) the
568 complete chemical structure of pyoverdines, and (iv) the FpvA receptors from genome
569 sequences. After validation, we tested our pipeline with sequence data from 1664
570 phylogenetically distinct *Pseudomonas* strains and were able to determine 18,292 enzymatic A
571 domains involved in pyoverdine synthesis, reliably predicted 97.8% of their substrates,
572 identified 188 different pyoverdine molecule structures and 4547 FpvA receptor variants
573 belonging to 94 distinct groups. The uncovered diversity is stunning and goes far beyond
574 currently known levels of variation (73 pyoverdines and 3 FpvA groups). The molecular diversity
575 of iron scavenging capacity highlights its importance among pseudomonads.

576 We show that knowledge-guided learning is an extremely powerful tool to predict enzyme,
577 metabolite, and receptor properties. The establishment of our entire pipeline is based on only
578 101 previously known enzymatic A domains (from 13 known pyoverdine assembly lines) and
579 21 FpvA receptor sequences. Even with this limited amount of information, we were able to
580 predict the substrates of almost all the 18,292 enzymatic A domains and to identify 4547 FpvA

581 receptors from the sequence data. A key insight from our knowledge-guided learning is that
582 comparisons based on the full gene sequences (e.g., for pyoverdine synthetase or receptor)
583 are likely non-informative and unsuitable for obtaining functional information. This is because
584 overall diversity does not stand for functional diversity, meaning that A domains recognizing the
585 same substrate can diverge substantially in their full sequences. The same holds true for
586 receptor sequences: whole-sequence alignments can neither accurately identify FpvA
587 receptors nor reliably separate them into functional groups. Instead, it is imperative to extract
588 informative feature sequences that are defined as sequence stretches within a gene whose
589 diversity is tightly linked to variation in its functioning. We successfully extracted and applied
590 feature sequence comparisons for both A domain substrate prediction and FpvA identification.
591 It is important to note that a knowledge-guided pipeline does not have to be perfect right from
592 the start. For example, our pipeline for pyoverdine structure prediction returned unknowns for
593 several amino acid positions within the PEP. Our experimental verifications then revealed
594 indeed new substrates such as valine and citrulline. This information can then be used to refine
595 our prediction algorithm in a feedback loop.

596 Another main advantage of our bioinformatic pipeline is that it can be applied to draft
597 genomes. This reflects a major improvement compared to existing annotation tools such as
598 antiSMASH[23], which typically has difficulties in recognizing NRPS structures in fragmented
599 genome assemblies. However, draft genomes are the most common data source in
600 microbiology. While our pipeline shows high performance, we need to acknowledge that we still
601 lose many genomes (6087 out of 9599, 63.4%). The reason for the loss is that the pyoverdine
602 synthesis machinery is large, which increases the probability that it is positioned at the end of

603 a contig. We decided to exclude those cases because the annotated synthesis machinery might
604 be truncated and thus incomplete. Thus, the high loss rate of draft genomes is rather due to
605 limitations in sequence quality (too many short contigs) and not due to a limitation of our
606 bioinformatic pipeline. We believe that this limitation may disappear in the future as long-read
607 sequencing technologies are quickly becoming cheaper and more reliable.

608 We further show that knowledge-guided learning combined with a phylogeny-focused
609 approach is a powerful tool for predicting the substrate specificity of A domains of synthetases.
610 It outperforms currently known bioinformatics prediction tools of NRPS substrates such as
611 antiSMASH[23]. Most current algorithms[48-52]perform poorly when applied to pyoverdines,
612 particularly when encountering non-proteogenic amino acids. The high accuracy of our
613 algorithm can largely be attributed to our reference set, composing only 13 pyoverdines from
614 *Pseudomonas spp*., yet capturing most of the substrate diversity. Similarly accurate
615 predictions based on a handful of known substrates among closely related species were
616 observed in several fungal NRPS systems[53]. It is worth noting that when the algorithm output
617 is "unknown," it actually signifies uncharacterized A domains not yet incorporated into the
618 reference data set. This should prompt researchers to pay attention to these A domains, and
619 like in our case, subject them to further experimental investigation. This approach helped us
620 discover new substrates (valine, histidine, citrulline), which had not been previously
621 documented in pyoverdines and were therefore absent from the reference A domains. The
622 novel substrates identified through our structural assessment and mass spectrometry
623 experiment can subsequently be used to enhance the precision of our phylogeny-centered
624 substrate prediction technique in the future, creating a progressive feedback loop of expanding

625 knowledge. Taken together, supervised learning based on a few known compounds produced
626 by species from the same genus probably outperforms generalized prediction algorithms
627 trained on many products from a diverse set of microbes for NRPS substrate predictions.

628 Our results show that both pyoverdine and receptor diversity has been vastly
629 underestimated. While considerable pyoverdine diversity (n=73) has already been captured in
630 previous studies, here we discovered 151 new variants. On the receptor side, the uncovered
631 novel diversity is more dramatic. One reason for this is that research on receptors has mainly
632 focused on the pathogen *P. aeruginosa*[19-22,54]. For this species, three different pyoverdine types
633 were described[21] together with three structurally different FpvA receptor types that each
634 recognize one of the pyoverdine types[22]. While our study confirmed that *P. aeruginosa* strains
635 (n = 554) indeed have only 3 pyoverdine-receptor systems, we also discovered 91 new FpvA
636 groups among environmental *Pseudomonas* spp. Our findings raise the question why there are
637 so many different pyoverdine and receptor variants. One potential explanation is that the benefit
638 of specific siderophores could be context-dependent and locally adapted to multitude of
639 different environmental conditions pseudomonads are exposed to. For example, experimental
640 work has revealed that pyoverdines can be cooperatively shared among strains with matching
641 receptors[36,55], or conversely, pyoverdines can serve as competitive agents by locking away iron
642 from species that have non-matching receptors[56]. Given that bioavailable iron is limited in most
643 natural and host-associated habitats[57-59], the unraveled functional diversity is likely a direct
644 evolutionary consequence of the struggle and competition of microbes for iron. While
645 experimental work is often restricted to a low number of strains, we propose that our
646 bioinformatic pipeline can be used to predict pyoverdine-mediated interaction networks across

thousands of strains and across different habitats. We will address this point in a future study.

We believe our pipeline could be easily expanded to study iron competition in multi-species communities in the future and perhaps in plant-microbe ecosystems, as siderophores exist ubiquitously and are shared among microbes[60]. To move further, a key question is whether our knowledge-guided approach can be applied to other microbes and important secondary metabolites, such as antibiotics, toxins, biosurfactants and pigments? Our preliminary investigation, focused on the curated dataset in *Burkholderiales* as a test case, provides a highly promising response to this query. The outcomes from this evaluation predominantly showcased the method's potential for extension to diverse microbial genera and metabolites, given sufficiently accurate dataset. Meanwhile, we observed a notable decline in accuracy when compared to pyoverdine. This indicates that our algorithm possesses potential for refinement, particularly concerning specific secondary metabolites or microbial genera. As soon as sufficient case-by-case knowledge on a specific system is available, the annotation strategies together with the feature sequence extraction and the phylogeny-focused approach developed in our paper can be applied. For most of the secondary metabolites listed above, there are no receptors as the compounds have purely extra-cellular functions, which substantially simplifies the development of bioinformatic pipelines. In the long run, it will certainly be possible to automate the steps implemented in our workflow so that the algorithms can be applied to a large set of secondary metabolites when fed with an appropriate training set.

Method

Construction of phylogeny tree

The phylogenetic tree depicted in Figure 1e was constructed utilizing the PhyloPhlAn3 pipeline[61]. PhyloPhlAn is a comprehensive pipeline that encompasses the identification of phylogenetic markers, multiple sequence alignment, and the inference of phylogenetic trees. In this analysis, we employed over 400 universal genes defined by PhyloPhlAn as our selected phylogeny markers. Subsequently, the taxonomic cladogram was generated using the iTOL web tool (http://huttenhower.sph.harvard.edu/galaxy/).

Apply our pipeline to identify errors in the substrate information associated with the A domain of *Pseudomonas* in the MIBIG database

We downloaded all A domain and substrate data from all NRPS secondary metabolism instances within *Pseudonomas* from MIBIG 3.0[32](Supplementary_table4). Following our pipeline, we initiate the process by utilizing NRPSMotifFinder to extract all Amotif4-5 sequences of the A domain. Subsequently, we calculate the Jukes-Cantor sequence distance between these Amotif4.5 pairs. Finally, employing Ward linkage clustering, we conduct a cluster analysis on all Amotif4.5 sequences, resulting in the generation of Figure S2. We observed that within the same cluster, substrates often exhibit some very subtle impurities. We speculate that such clusters might contain some errors in substrate information. Consequently, we focused on these potential errors, manually searched for their original literature, and verified that the A domain and its corresponding substrate information entered in MIBIG do indeed contain errors (Supplementary_table5).

690 **Apply our pipeline to predict the structural composition of diverse NRPS metabolites**

691 **across various genera of *Burkholderiales***

692 To evaluate the extendibility of our bioinformatic pipeline to various metabolites and microbial

693 genera, we utilized a previously curated dataset of A domain sequences with experimentally

694 confirmed substrates from *Burkholderiales* [62] (Supplementary_table7). This manually curated

695 dataset, encompassing 7 genera, 34 secondary products, 203 A domains, and 25 substrates

696 within *Burkholderiales*, serves as a valuable resource and mainly as a test set. Simultaneously,

697 we compiled the training set by aggregating A domain and substrate data from all NRPS

698 secondary metabolism instances within *Burkholderiales*, utilizing experimentally validated

699 biosynthetic gene cluster data from MIBIG 3.0[32] (Supplementary_table8). This training set

700 covers 6 genera, 19 secondary products, 99 A domains, and 21 substrates. To ensure the

701 inclusivity of all substrates from the test set, we randomly selected an A domain from each

702 substrate in the manually curated dataset, incorporating it into our training set. Consequently,

703 the finalized training set comprises 124 A domains (Supplementary_table12), while the test set

704 encompasses 178 A domains (Supplementary_table9). Then, we concurrently employed our

705 pipeline and antiSMASH[16] to predict the corresponding substrate based on the sequence data

706 of the A domain in the test set. Subsequently, we compared these predictions with the actual

707 substrate information, calculating the accuracy rate for each algorithm.

708 **Apply our pipeline to annotate the receptor FpvA across the bacterial domain**

709 To explore the distribution of FpvA across the bacterial domain, we acquired all complete

710 bacterial genomes (33,207) from NCBI (https://www.ncbi.nlm.nih.gov). Employing the pipeline

711 depicted in Figure 5h to examine the CDS of these genomes, we identified a total of 357,790

712 TonB-dependent receptors, comprising 31,936 for FpvA and 623 for FpvB. Subsequently,
713 leveraging the taxonomy information from NCBI, we conducted a comparative analysis of the
714 distribution differences between FpvA and FpvB among various species. The findings revealed
715 that FpvA exhibited a widespread distribution across different species, whereas FpvB was
716 exclusive to the genus *Pseudomonas*, with a maximum occurrence of one in a given genome.
717
718

719 **Data Availability**

720 The source code and parameters used are available in the supplementary material.

721

722 **Acknowledgements**

723 We thank Richard Allen for genome sequencing of the 20 strains. We also appreciated Vera
724 Vollenweider for sample preparation for the experiment of pyoverdine structure elucidation.

725

726 **Funding**

727 This work was supported by the National Key Research and Development Program of China
728 (No. 2021YFF1200500, 2021YFA0910700), National Natural Science Foundation of China (No.
729 42107140, No. 32071255, No.41922053, No. T2321001), National Postdoctoral Program for
730 Innovative Talents (No. BX2021012). R.K. is supported jointly by a grant from the Swiss
731 National Science Foundation no. 310030_212266. V-P.F. is supported jointly by a grant from
732 UKRI, Defra, and the Scottish Government, under the Strategic Priorities Fund Plant Bacterial
733 Diseases program (BB/T010606/1), Research Council of Finland, and The Finnish Research
734 Impact Foundation.

735

736 **Author contributions**

737 Shaohua Gu performed the majority of computational analysis in this research and drafted the
738 manuscript. Yuanzhe Shao built the pipeline for retrieving synthetase sequence information and
739 developed the phylogeny-centered method for predicting substrates. Karoline Rehm and
740 Laurent Bigler performed the experiment of pyoverdine structure elucidation. Di Zhang depicted
741 the circular plot. Ruolin He standardized of all NRPS structures into motif-intermotif format. Jiqi
742 Shao and Ruichen Xu assisted in revising the manuscript. Xiaoying Bian provided curated
743 dataset of A domain sequences with experimentally confirmed substrates from *Burkholderiales*.
744 Xiaoying Bian, Alexandre Jousset and Ville-Petri Friman offered insightful comments and
745 assisted in revising and writing of the manuscript. Rolf Kümmerli and Zhong Wei oversaw the
746 project, designed experiments and revised the manuscript. Zhiyuan Li conceptualized and
747 oversaw the project, conducted the analysis of the receptor annotated method, and revised the
748 manuscript.

749

750 **Competing interests**

751 The authors declare no competing interests.

752

753 **References**

754 1 Zengler, K. & Palsson, B. O. A road map for the development of community systems
755 (CoSy) biology. *Nature Reviews Microbiology* **10**, 366-372, doi:10.1038/nrmicro2763
756 (2012).

2 Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models. *Genome Biology* **20**, 121, doi:10.1186/s13059-019-1730-3 (2019).

3 García-Jiménez, B., Torres-Bacete, J. & Nogales, J. Metabolic modelling approaches for describing and engineering microbial communities. *Computational and Structural Biotechnology Journal* **19**, 226-246, doi:10.1016/j.csbj.2020.12.003 (2021).

4 Colarusso, A. V., Goodchild-Michelman, I., Rayle, M. & Zomorrodi, A. R. Computational modeling of metabolism in microbial communities on a genome-scale. *Current Opinion in Systems Biology* **26**, 46-57, doi:10.1016/j.coisb.2021.04.001 (2021).

5 Scherlach, K. & Hertweck, C. Mediators of mutualistic microbe–microbe interactions. *Natural Product Reports* **35**, 303-308, doi:10.1039/C7NP00035A (2018).

6 Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology* **18**, 607-621, doi:10.1038/s41579-020-0412-1 (2020).

7 Price-Whelan, A., Dietrich, L. E. P. & Newman, D. K. Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nature Chemical Biology* **2**, 71-78, doi:10.1038/nchembio764 (2006).

8 Thirumurugan, D., Cholarajan, A., Raja, S. & Vijayakumar, R. An introductory chapter: secondary metabolites. (2018).

9 Chaturvedi, K. S., Hung, C. S., Crowley, J. R., Stapleton, A. E. & Henderson, J. P. The siderophore yersiniabactin binds copper to protect pathogens during infection. *Nature Chemical Biology* **8**, 731-736, doi:10.1038/nchembio.1020 (2012).

10 Penn, K. *et al.* Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME Journal* **3**, 1193-1203, doi:10.1038/ismej.2009.58 (2009).

11 Yee, D. A. *et al.* Genome mining for unknown–unknown natural products. *Nature Chemical Biology* **19**, 633-640, doi:10.1038/s41589-022-01246-6 (2023).

12 Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nature Chemical Biology* **1**, 265-269, doi:10.1038/nchembio731 (2005).

13 Andryukov, B., Mikhailov, V. & Besednova, N. The Biotechnological Potential of Secondary Metabolites from Marine Bacteria. *Journal of Marine Science and Engineering* **7** (2019).

14 Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Research* **49**, D490-D497, doi:10.1093/nar/gkaa812 (2021).

15 He, R. *et al.* Knowledge-guided data mining on the standardized architecture of NRPS: Subtypes, novel motifs, and sequence entanglements. *PLOS Computational Biology* **19**, e1011100, doi:10.1371/journal.pcbi.1011100 (2023).

16 Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Research* **51**, W46-W50, doi:10.1093/nar/gkad344 (2023).

17 Ringel, M. T. & Brüser, T. The biosynthesis of pyoverdines. *Microb Cell* **5**, 424-437, doi:10.15698/mic2018.10.649 (2018).

801 18 Hopkinson, B. M. & Morel, F. M. M. The role of siderophores in iron acquisition by
802 photosynthetic marine microorganisms. *BioMetals* **22**, 659-669, doi:10.1007/s10534-
803 009-9235-2 (2009).
804 19 Cobessi, D. *et al.* The Crystal Structure of the Pyoverdine Outer Membrane Receptor
805 FpvA from Pseudomonas aeruginosa at 3.6Å Resolution. *Journal of Molecular Biology*
806 **347**, 121-134, doi:10.1016/j.jmb.2005.01.021 (2005).
807 20 Diggle, S. P. & Whiteley, M. Microbe Profile: Pseudomonas aeruginosa: opportunistic
808 pathogen and lab rat. *Microbiology (Reading)* **166**, 30-33, doi:10.1099/mic.0.000860
809 (2020).
810 21 Meyer, J.-M. *et al.* Use of Siderophores to Type Pseudomonads: The Three
811 Pseudomonas Aeruginosa Pyoverdine Systems. *Microbiology* **143**, 35-43,
812 doi:10.1099/00221287-143-1-35 (1997).
813 22 Bodilis, J. *et al.* Distribution and evolution of ferripyoverdine receptors in Pseudomonas
814 aeruginosa. *Environmental Microbiology* **11**, 2123-2135, doi:10.1111/j.1462-
815 2920.2009.01932.x (2009).
816 23 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining
817 pipeline. *Nucleic Acids Research* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).
818 24 Kümmerli, R. Iron acquisition strategies in pseudomonads: mechanisms, ecology, and
819 evolution. *BioMetals*, doi:10.1007/s10534-022-00480-8 (2022).
820 25 Meyer, J. M. Pyoverdines: pigments, siderophores and potential taxonomic markers of
821 fluorescent Pseudomonas species. *Arch Microbiol* **174**, 135-142,
822 doi:10.1007/s002030000188 (2000).

823 26 Xu, Z., Park, T. J. & Cao, H. Advances in mining and expressing microbial biosynthetic
824 gene clusters. *Crit Rev Microbiol*, 1-20, doi:10.1080/1040841x.2022.2036099 (2022).
825 27 Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery.
826 *Nature Reviews Microbiology* **17**, 167-180, doi:10.1038/s41579-018-0121-1 (2019).
827 28 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Research* **32**,
828 D138-D141, doi:10.1093/nar/gkh121 (2004).
829 29 Winsor, G. L. *et al.* Enhanced annotations and features for comparing thousands of
830 Pseudomonas genomes in the Pseudomonas genome database. *Nucleic Acids*
831 *Research* **44**, D646-D653, doi:10.1093/nar/gkv1227 (2016).
832 30 Felnagle, E. A. *et al.* Nonribosomal peptide synthetases involved in the production of
833 medically relevant natural products. *Mol Pharm* **5**, 191-211, doi:10.1021/mp700137g
834 (2008).
835 31 Süssmuth, R. D. & Mainz, A. Nonribosomal Peptide Synthesis—Principles and
836 Prospects. *Angewandte Chemie International Edition* **56**, 3770-3821,
837 doi:10.1002/anie.201609079 (2017).
838 32 Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally
839 validated biosynthetic gene clusters. *Nucleic Acids Research* **51**, D603-D610,
840 doi:10.1093/nar/gkac1049 (2023).
841 33 Butaitė, E., Kramer, J., Wyder, S. & Kümmerli, R. Environmental determinants of
842 pyoverdine production, exploitation and competition in natural Pseudomonas
843 communities. *Environmental Microbiology* **20**, 3629-3642, doi:10.1111/1462-
844 2920.14355 (2018).

845 34 Rehm, K., Vollenweider, V., Kümmerli, R. & Bigler, L. A comprehensive method to
846 elucidate pyoverdines produced by fluorescent Pseudomonas spp. by UHPLC-HR-
847 MS/MS. *Analytical and Bioanalytical Chemistry* **414**, 2671-2685, doi:10.1007/s00216-
848 022-03907-w (2022).
849 35 Butaitė, E., Baumgartner, M., Wyder, S. & Kümmerli, R. Siderophore cheating and
850 cheating resistance shape competition for iron in soil and freshwater Pseudomonas
851 communities. *Nature Communications* **8**, 414, doi:10.1038/s41467-017-00509-4
852 (2017).
853 36 Kramer, J., Özkaya, Ö. & Kümmerli, R. Bacterial siderophores in community and host
854 interactions. *Nature Reviews Microbiology* **18**, 152-163, doi:10.1038/s41579-019-
855 0284-4 (2020).
856 37 Jin, Z. *et al.* Conditional privatization of a public siderophore enables Pseudomonas
857 aeruginosa to resist cheater invasion. *Nature Communications* **9**, 1383,
858 doi:10.1038/s41467-018-03791-y (2018).
859 38 Chan, D. C. K. & Burrows, L. L. Pseudomonas aeruginosa FpvB
860 is a high-affinity transporter for xenosiderophores ferrichrome and ferrioxamine B.
861 *bioRxiv*, 2022.2009.2020.508722, doi:10.1101/2022.09.20.508722 (2022).
862 39 González, J. *et al.* Loss of a pyoverdine secondary receptor in Pseudomonas
863 aeruginosa results in a fitter strain suitable for population invasion. *The ISME Journal*
864 **15**, 1330-1343, doi:10.1038/s41396-020-00853-2 (2021).
865 40 Schalk, I. J., Mislin, G. L. & Brillet, K. Structure, function and binding selectivity and
866 stereoselectivity of siderophore-iron outer membrane transporters. *Curr Top Membr* **69**,

867 37-66, doi:10.1016/B978-0-12-394390-3.00002-1 (2012).

868 41 Greenwald, J. *et al.* FpvA bound to non-cognate pyoverdines: molecular basis of
869 siderophore recognition by an iron transporter. *Molecular Microbiology* **72**, 1246-1259,
870 doi:10.1111/j.1365-2958.2009.06721.x (2009).

871 42 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**,
872 499-504, doi:10.1038/s41586-019-0965-1 (2019).

873 43 Schloss, P. D. & Handelsman, J. Metagenomics for studying unculturable
874 microorganisms: cutting the Gordian knot. *Genome Biology* **6**, 229, doi:10.1186/gb-
875 2005-6-8-229 (2005).

876 44 Handelsman, J. Metagenomics: Application of Genomics to Uncultured
877 Microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685,
878 doi:10.1128/MMBR.68.4.669-685.2004 (2004).

879 45 Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome:
880 fundamentals, tools, and challenges. *Annals of Epidemiology* **26**, 330-335,
881 doi:10.1016/j.annepidem.2016.03.002 (2016).

882 46 Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in
883 sensitivity and precision. *ISME J* **10**, 1669-1681, doi:10.1038/ismej.2015.235 (2016).

884 47 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Reviews*
885 *Microbiology* **10**, 538-550, doi:10.1038/nrmicro2832 (2012).

886 48 Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. Classification of the
887 Adenylation and Acyl-Transferase Activity of NRPS and PKS Systems Using
888 Ensembles of Substrate Specific Hidden Markov Models. *PLOS ONE* **8**, e62136,

889 doi:10.1371/journal.pone.0062136 (2013).

890 49 Minowa, Y., Araki, M. & Kanehisa, M. Comprehensive Analysis of Distinctive Polyketide
891 and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *Journal*
892 *of Molecular Biology* **368**, 1500-1517, doi:10.1016/j.jmb.2007.02.099 (2007).

893 50 Prieto, C., García-Estrada, C., Lorenzana, D. & Martín, J. F. NRPSsp: non-ribosomal
894 peptide synthase substrate predictor. *Bioinformatics* **28**, 426-427,
895 doi:10.1093/bioinformatics/btr659 (2012).

896 51 Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation
897 domain specificity. *Nucleic Acids Research* **39**, W362-W367, doi:10.1093/nar/gkr323
898 (2011).

899 52 Zierep, P. F., Ceci, A. T., Dobrusin, I., Rockwell-Kollmann, S. C. & Günther, S. SeMPI
900 2.0-A Web Server for PKS and NRPS Predictions Combined with Metabolite Screening
901 in Natural Product Databases. *Metabolites* **11**, doi:10.3390/metabo11010013 (2020).

902 53 Fan, J. *et al.* Biosynthetic diversification of peptaibol mediates fungus-mycohost
903 interactions. *bioRxiv* (2022).

904 54 Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R. & Olson, M. V. Evidence for
905 diversifying selection at the pyoverdine locus of Pseudomonas aeruginosa. *J Bacteriol*
906 **187**, 2138-2147, doi:10.1128/jb.187.6.2138-2147.2005 (2005).

907 55 Gu, S. *et al.* Competition for iron drives phytopathogen control by natural rhizosphere
908 microbiomes. *Nature Microbiology* **5**, 1002-1010, doi:10.1038/s41564-020-0719-8
909 (2020).

910 56 Figueiredo, A. R. T., Özkaya, Ö., Kümmerli, R. & Kramer, J. Siderophores drive invasion

dynamics in bacterial communities through their dual role as public good versus public bad. *Ecology Letters* **25**, 138-150, doi:10.1111/ele.13912 (2022).

57 Andrews, S. C., Robinson, A. K. & Rodríguez-Quiñones, F. Bacterial iron homeostasis. *FEMS Microbiology Reviews* **27**, 215–237, doi:10.1016/s0168-6445(03)00055-x (2003).

58 Boyd, P. W. & Ellwood, M. J. The biogeochemical cycle of iron in the ocean. *Nat Geosci* **3**, 675-682, doi:10.1038/ngeo964 (2010).

59 Emerson, D., Roden, E. & Twining, B. The microbial ferrous wheel: iron cycling in terrestrial, freshwater, and marine environments. *Frontiers in Microbiology* **3**, doi:10.3389/fmicb.2012.00383 (2012).

60 Ruolin, H. *et al.* SIDERITE: Unveiling Hidden Siderophore Diversity in the Chemical Space Through Digital Exploration. *bioRxiv*, 2023.2008.2031.555687, doi:10.1101/2023.08.31.555687 (2023).

61 Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nature Communications* **11**, 2500, doi:10.1038/s41467-020-16366-7 (2020).

62 Chen, H. *et al.* Biosynthesis and engineering of the nonribosomal peptides with a C-terminal putrescine. *Nature Communications* **14**, 6619, doi:10.1038/s41467-023-42387-z (2023).