

1 **Accurate *de novo* identification of biosynthetic gene clusters with GECCO**

2 Laura M. Carroll<sup>1\*</sup>, Martin Larralde<sup>1\*</sup>, Jonas Simon Fleck<sup>1,3\*</sup>, Ruby Ponnudurai<sup>1</sup>, Alessio

3 Milanese<sup>1,2</sup>, Elisa Cappio<sup>1</sup>, Georg Zeller<sup>1+</sup>

4

5 <sup>1</sup> Structural and Computational Biology Unit, EMBL, Heidelberg, Germany

6 <sup>2</sup> Department of Biology, ETH Zürich, Zürich, Switzerland

7 <sup>3</sup> Present address: Department of Biosystems Science and Engineering, ETH Zürich, Basel,  
8 Switzerland

9 \* These authors contributed equally to this work

10 + Corresponding author: Georg Zeller, [zeller@embl.de](mailto:zeller@embl.de)

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26 Biosynthetic gene clusters (BGCs) are enticing targets for (meta)genomic mining efforts, as  
27 they may encode novel, specialized metabolites with potential uses in medicine and  
28 biotechnology. Here, we describe GECCO (GEne Cluster prediction with COnditional  
29 random fields; <https://gecco.embl.de>), a high-precision, scalable method for identifying novel  
30 BGCs in (meta)genomic data using conditional random fields (CRFs). Based on an extensive  
31 evaluation of *de novo* BGC prediction, we found GECCO to be more accurate and over 3x  
32 faster than a state-of-the-art deep learning approach. When applied to over 12,000 genomes,  
33 GECCO identified nearly twice as many BGCs compared to a rule-based approach, while  
34 achieving higher accuracy than other machine learning approaches. Introspection of the  
35 GECCO CRF revealed that its predictions rely on protein domains with both known and  
36 novel associations to secondary metabolism. The method developed here represents a  
37 scalable, interpretable machine learning approach, which can identify BGCs *de novo* with  
38 high precision.

39

40

41

42

43

44

45

## 46 INTRODUCTION

47 Host-associated and environmental microbes alike are capable of producing a wide  
48 array of secondary metabolites through which they interact with their environments.<sup>1</sup> These  
49 metabolites equip their producer with a chemical repertoire to respond to stressors, which may  
50 confer competitive advantages over other organisms in their environmental niche.<sup>2,3</sup> In human  
51 host-associated microbial communities, secondary metabolites can also modulate host health  
52 via a range of processes, including immune system regulation, xenobiotic and nutrient  
53 metabolism, and cancer susceptibility/resistance.<sup>3,4</sup> Beyond their natural purposes, many  
54 microbial secondary metabolites have found important uses in medicine, including as first-in-  
55 class antimicrobial, anticancer, and antidiabetic drugs.<sup>1,5,6</sup>

56 Due to the biomedical and biotechnological interest in microbial secondary metabolites,  
57 there is a strong incentive to identify novel natural products. Genome mining efforts have  
58 successfully made use of the fact that a large proportion of the enzymatic pathways responsible  
59 for secondary metabolite production are encoded by physically clustered groups of genes called  
60 biosynthetic gene clusters (BGCs).<sup>1,7-9</sup> Recently, the development of computational tools for  
61 BGC detection has been further fueled by the ever-increasing availability of microbial genomic  
62 and metagenomic data.<sup>7-10</sup> Currently, *in silico* methods used to identify BGCs in  
63 (meta)genomic sequencing data can largely be categorized into two groups. “Rule-based”  
64 approaches (e.g., antiSMASH, PRISM)<sup>8,11,12</sup> use hard-coded BGC detection “rules” to identify  
65 BGCs in (meta)genomic data based on signature genes.<sup>9</sup> These approaches display a high  
66 degree of precision (i.e., low false positive rates) but are unable to detect novel BGCs of  
67 unknown architecture. To prioritize the detection of novel BGCs, “model-based” approaches  
68 have been developed.<sup>7,9</sup> The most widely used representative of this group, ClusterFinder,  
69 relies on a hidden Markov model (HMM) to segment (meta)genomic sequences into BGC and  
70 non-BGC regions based on local enrichment of protein domains characteristic of biosynthetic

71 genes.<sup>7</sup> More recently, DeepBGC, which employs a three-layer Bidirectional Long Short-Term  
72 Memory (BiLSTM) recurrent neural network (RNN)<sup>13</sup>, was shown to yield more accurate *de*  
73 *novo* BGC predictions than HMM-based ClusterFinder.<sup>9</sup>

74 Conditional random fields (CRFs) are an alternative machine learning (ML) approach  
75 to HMMs and BiLSTMs for sequence segmentation. These discriminative graphical models  
76 (Fig. 1 and Supplementary Figure S1) have been shown to outperform generative models, such  
77 as HMMs, in various application domains.<sup>14,15</sup> Furthermore, compared to their “black box”  
78 RNN counterparts, CRFs have the advantage of being inherently interpretable, an important  
79 feature in a biomedical context.<sup>16</sup> Here, we describe GECCO (GEne Cluster prediction with  
80 COnditional random fields; <https://gecco.embl.de>), a high-precision, scalable method for *de*  
81 *novo* BGC identification in microbial genomic and metagenomic data. On the basis of a newly  
82 developed, extensive *de novo* BGC prediction benchmarking framework, we show that  
83 GECCO is not only more accurate than state-of-the-art *de novo* BGC detection approaches, but  
84 also more computationally efficient. As an interpretable ML model, GECCO can moreover  
85 provide insights into BGC biology, architecture, and function.

## 86 RESULTS

### 87 GECCO: a CRF-based *de novo* BGC detection tool

88 To train a CRF that could identify novel BGCs in (meta)genomic sequences, a  
89 training/cross-validation (CV) data set was constructed by embedding known BGCs into long,  
90 BGC-negative fragments of prokaryotic genomes (Fig. 1 and Supplementary Figures S1 and  
91 S2). Briefly, known BGCs present in the Minimum Information about a Biosynthetic Gene  
92 cluster (MIBiG) database<sup>1</sup> were embedded into randomly selected prokaryotic contigs, in  
93 which other known and predicted BGCs had been masked (see section “Data acquisition and  
94 feature construction” below). To construct the feature matrix for training, open reading frames  
95 (ORFs) were identified and annotated with protein domains, using one of fourteen

96 combinations of databases in which protein families are represented by profile hidden Markov  
97 models (pHMMs; see Fig. 1, section “Data acquisition and feature construction” below, and  
98 Supplementary Figures S1 and S2). As the protein family resources are broader in scope than  
99 what may be needed for BGC identification and their combinations potentially redundant, an  
100 additional feature selection approach was implemented in GECCO: to identify domains that  
101 are either most strongly enriched or depleted in BGCs, we nested a two-sided Fisher’s Exact  
102 Test (FET) into the CV employed for fitting GECCO’s CRF. Within each CV fold, we  
103 iteratively retrained GECCO using only the top domains associated with BGC presence or  
104 absence to estimate how far the CRF feature space (i.e. the domain pHMMs used for  
105 annotation) could be reduced to gain speed while retaining optimal prediction accuracy  
106 (Supplementary Figures S3 and S4a).

107 **GECCO provides superior precision and speed relative to state-of-the-art *de novo* BGC  
108 prediction methods**

109 To construct a benchmark data set in a way that guarantees that training and test data  
110 are disjoint, we partitioned MIBiG v2.0<sup>17</sup> into (i) BGCs for training that were already contained  
111 in an earlier MIBiG version (v1.3, which was also originally used to train DeepBGC and  
112 DeepBGC’s re-trained implementation of ClusterFinder)<sup>9</sup>, and (ii) selected BGCs for testing  
113 that were newly added in subsequent updates of MIBiG; from this test set we also removed  
114 BGCs that were very similar in architecture to any instance contained in MIBiG v1.3 (see  
115 section “Data acquisition and feature construction” below for additional details). This yielded  
116 a final test set of 376 prokaryotic contigs which each had an embedded BGC that was  
117 exclusively present in MIBiG v2.0 (referred to hereafter as the “376-genome test set”). We also  
118 used two additional, previously constructed test sets containing thoroughly annotated genomes  
119 of well-studied BGC producer organisms (the “six genome test set” used by Hannigan, et al.<sup>9</sup>  
120 to evaluate DeepBGC and the “nine genome bootstrapping set” used by Hannigan, et al.<sup>9</sup> for

121 hyperparameter tuning, validation, and testing of DeepBGC) and removed all instances of  
122 BGCs similar to those in these additional test sets from the MIBiG v1.3-based training set. To  
123 ensure a fair comparison of BGC detection methods, we retrained DeepBGC and GECCO on  
124 this very same training set, using BGCs from MIBiG v1.3 which were absent from all test sets.  
125 We evaluated the performance of both methods using (i) the 376-genome test set (the main test  
126 set presented in this study; Fig. 2a-c), as well as (ii) the six- and (iii) nine-genome test sets<sup>9</sup> and  
127 (iv) 10-fold CV using BGCs from MIBiG v1.3 (Supplementary Figures S4-S6). We  
128 additionally compared GECCO and the re-trained implementation of DeepBGC to the original  
129 DeepBGC, as well as DeepBGC’s “original” and “re-trained” implementations of the  
130 ClusterFinder algorithm (also trained on BGCs from MIBiG v1.3), using the three  
131 aforementioned test sets (Fig. 2a-c and Supplementary Figure S6).<sup>7,9</sup>

132 For direct comparability to previous evaluations,<sup>9</sup> we first conducted a receiver  
133 operating characteristic (ROC) analysis on the level of individual domains, i.e. based on a per-  
134 domain assessment of true positives, true negatives, false positives and false negatives (Fig.  
135 2a, Supplementary Figure S4a). Based on area under (AU) the ROC curve values, GECCO  
136 showed superior performance compared to all DeepBGC/ClusterFinder implementations on  
137 the 376- and six-genome test sets (Fig. 2a and Supplementary Figure S6) and during 10-fold  
138 CV (GECCO AUROC = 0.97; Supplementary Figure S5). On the nine-genome test set,  
139 GECCO and the original implementation of DeepBGC performed equally (AUROC = 0.94;  
140 Supplementary Figure S6). From the same true/false positive/negative metrics, we also  
141 constructed per-domain precision-recall (PR) curves (Fig. 2b, Supplementary Figure S4a).  
142 These evaluations showed superior performance of GECCO compared to all  
143 DeepBGC/ClusterFinder implementations for all three test sets (Fig. 2b and Supplementary  
144 Figure S6) and during 10-fold CV (GECCO AUPR = 0.73; Supplementary Figure S5). In  
145 addition to evaluating model performance at the domain level, we also assessed to which extent

146 predicted BGCs overlapped with known BGCs by calculating precision and recall from true  
147 and false positive BGC segments, as well as false negative non-BGC segments, for each model  
148 (referred to hereafter as the “segment overlap” metric; Supplementary Figure S4b). Based on  
149 PR curves constructed from this segment overlap metric, GECCO achieved substantially higher  
150 AUPR than all implementations of DeepBGC/ClusterFinder (Fig. 2c and Supplementary  
151 Figures S4-S7). These evaluations demonstrate that GECCO is capable of detecting BGCs *de*  
152 *novo* with unprecedented accuracy, primarily by more precisely locating their boundaries; this  
153 also greatly alleviates the problem of fragmented predictions, which other methods suffer from  
154 (Fig. 2e).

155 We moreover used the training data to optimize GECCO’s feature space. We found that  
156 feature inclusion thresholds  $T = [35,100]$  (percentage of retained domain features) achieved  
157 highly similar AUPR and  $F_1$  scores (Fig. 2d and Supplementary Figure S3), suggesting that  
158 65% of features can be discarded without noticeable sacrifices in accuracy. Among the domain  
159 resources used for feature generation, a combination of TIGRFAM v15.0<sup>18</sup> and Pfam v33.1<sup>19</sup>  
160 with  $T = 35\%$  achieved among the highest AUROC and AUPR scores, and was thus chosen as  
161 the final model for BGC detection in GECCO (10-fold CV AUROC = 0.96, AUPR = 0.89; Fig  
162 2d, Supplementary Figure S3). To explore GECCO’s ability to identify novel BGC classes not  
163 currently represented in MIBiG, leave-one-type-out (LOTO) CV was used. In LOTO, one  
164 biosynthetic class of BGCs is completely removed from the training set during CV to  
165 specifically assess its re-discovery in the test set. GECCO achieved LOTO AUROC scores  $>$   
166 0.98 for four of six classes and 0.91 and 0.88 AUROC for MIBiG’s ribosomally synthesized  
167 and post-translationally modified peptide (RiPP) and Saccharide classes, respectively  
168 (Supplementary Figure S8). To determine the MIBiG biosynthetic class for each newly  
169 predicted BGC, a separate random forest (RF) classifier was trained and evaluated using five-  
170 fold CV, as has been previously proposed<sup>9</sup> (see section “Prediction of biosynthetic class”

171 below; Fig. 1). Using the domain composition associated with each BGC as features, the RF  
172 classifier achieved AUROC scores > 0.90 for all classes (Supplementary Figure S9).

173 In a final benchmark, we compared the runtime between GECCO, DeepBGC, and  
174 antiSMASH using the three test sets, as well as all representative genomes in the proGenomes2  
175 database, a comprehensive resource of prokaryotic genome sequences (containing 627,182  
176 contigs from 12,221 genomes; Fig. 2f and Supplementary Figure S10).<sup>20</sup> Using a single CPU,  
177 GECCO was over three times faster than both antiSMASH and DeepBGC (Fig. 2f and  
178 Supplementary Figure S10).

179 Taken together, this comparative evaluation, to our knowledge, is the most  
180 comprehensive benchmarking of *de novo* BGC prediction tools conducted to date. It clearly  
181 demonstrates that GECCO greatly improves the accuracy of *in silico* BGC identification over  
182 the state of the art, while also being computationally efficient.

183 **GECCO's CRF-based approach provides insight into the biosynthetic potential of  
184 microbes**

185 To compare GECCO BGC predictions to those produced by other tools on a real-world  
186 data set, each of GECCO, DeepBGC, and antiSMASH were used to identify and classify BGCs  
187 among all 12,221 representative genomes in the proGenomes2 database. Notably, the majority  
188 of BGCs predicted by either GECCO or DeepBGC were not detected using the rule-based  
189 approach implemented in antiSMASH ( $n = 59,041$  antiSMASH BGCs using default  
190 parameters; Fig. 3ab). Overall, GECCO predicted nearly twice as many BGCs as antiSMASH,  
191 but far fewer than DeepBGC ( $n = 115,131$  and 470,137 GECCO and DeepBGC BGCs,  
192 respectively), consistent with the above evaluations showing a clear tendency of GECCO to  
193 produce fewer false positives and fragmented predictions (Fig. 2a-c,e and Supplementary  
194 Figures S4-S7).

195 To investigate which protein domains GECCO relied on for BGC detection, we first  
196 analyzed which protein domains were retained in the first feature elimination step. Notably,  
197 nearly half of all GECCO protein domains (2,382 of 5,255 total GECCO protein domains,  
198 45.3%) were derived from the TIGRFAM database (Fig. 3c), highlighting the complementary  
199 nature of the Pfam and TIGRFAM databases for BGC prediction optimization. When compared  
200 to a collection of protein domains previously associated with secondary metabolism (used by  
201 BiG-SLiCE v1.1.0)<sup>21</sup>, nearly half of these core biosynthetic domains were included in  
202 GECCO's model (937 of 2,027 BiG-SLiCE core domains, 46.2%; Fig. 3c). Domains in the  
203 core biosynthetic set/GECCO intersection received more positive (i.e., BGC-associated) CRF  
204 weights relative to TIGRFAM domains not present in the core biosynthetic domain set, but not  
205 relative to Pfam domains not present in the core biosynthetic domain set (two-sided Mann-  
206 Whitney *U* test raw *P* = 3.12e-07 and 0.10, respectively; Fig. 3c). However, many other  
207 domains outside of the comparatively small core biosynthetic space received CRF weights with  
208 comparably high (absolute) values. Domains with negative weights were important for  
209 capturing non-BGC regions (Fig. 3c); however, some of the most highly weighted (i.e., BGC-  
210 associated) domains were not members of the core biosynthetic set (Fig. 3c, Supplementary  
211 Table S1). Among these (CRF weight > 4.0) were (i) terpene synthase family 2, C-terminal  
212 metal binding domain PF19086 and (ii) lantibiotic alpha domain PF14867, both of which have  
213 previously been associated with secondary metabolite production (Fig. 3c, Supplementary  
214 Table S1). Interestingly, among the highest-weighted, BGC-associated domains (CRF weight  
215 > 2.0) that were not members of the core biosynthetic set were three domains of unknown  
216 function (DUF): (i) PF19155 (DUF5837), which is associated with a cyanobactin (RiPP) BGC,  
217 tenuecyclamide A (MIBiG ID BGC0000480); (ii) PF11379 (DUF3182), a Proteobacteria-  
218 restricted protein of unknown function (InterPro ID IPR021519); (iii) PF17537 (DUF5455), a  
219 protein of unknown function found in Proteobacteria, which contains three predicted trans-

220 membrane regions (InterPro ID IPR035210; Supplementary Table S1). Their importance for  
221 BGC prediction with GECCO suggests that functional studies of these domains in the context  
222 of secondary metabolism are warranted.

223 To be able to observe coherent biological functions among the domain weights learned  
224 by the GECCO CRF, beyond the most strongly associated domains, we used Gene Ontology  
225 (GO)<sup>22</sup> and Pfam (structural) clan<sup>19</sup> annotations. This led to the identification of 33 biological  
226 processes (BPs) and 22 molecular functions (MFs) enriched in BGCs (topGO Kolmogorov-  
227 Smirnov  $P < 0.05$ ), with “defense response to bacterium” (GO:0042742), “secondary  
228 metabolite biosynthetic process” (GO:0044550), “isoprenoid biosynthetic process”  
229 (GO:0008299), and “toxin metabolic process” (GO:0009404) showcasing the strongest  
230 associations (all BGC enrichment scores  $> 2.5$ ; Fig. 3d and Supplementary Figure S11). Three  
231 Pfam clans were additionally enriched in BGCs (false discovery rate-corrected  $P < 0.10$ ):  
232 Alpha/Beta hydrolase fold (CL0028), CoA-dependent acyltransferase superfamily (CL0149),  
233 and Double-Glycine leader-peptide cleavage motif (CL0400; Fig. 3e and Supplementary  
234 Figure S12). Collectively, these results indicate that the GECCO CRF relies on domains  
235 associated with secondary metabolite production for BGC inference.

236 **DISCUSSION**

237 ML approaches have revolutionized numerous disciplines and are being increasingly  
238 employed to solve problems in biological and medical realms.<sup>23-25</sup> Models that can account for  
239 sequential data are particularly attractive when leveraging genomic data to make predictions,  
240 as feature context and order (e.g., for genes, domains) may be important.<sup>9</sup> CRFs specifically  
241 have played a crucial role in sequential modeling tasks and have been used extensively in areas  
242 such as natural language processing (NLP), where they frequently outperform their generative  
243 counterparts.<sup>14,15</sup>

244 Recently, deep learning approaches have become popular methods for processing  
245 sequential data. However, these models often require a great deal of training data and/or pre-  
246 training efforts to show marked improvements over classical ML models.<sup>16</sup> This is relevant for  
247 BGC identification, as the need for experimental characterization of “true” BGCs limits the  
248 amount of training data for these approaches; for example, the current version of MIBiG (v2.0)  
249 contains only 1,923 experimentally validated BGCs (with some being very closely related to  
250 one another, and thus of limited value as training data).<sup>17</sup> Here, we showed that, with the  
251 relatively limited amount of known BGCs available, the linear CRF implemented in GECCO  
252 outperforms DeepBGC’s BiLSTM approach, achieving higher accuracy at reduced training  
253 and prediction time.

254 An additional advantage of CRFs over deep learning approaches is that the former are  
255 inherently “simpler” and thus more interpretable (whereas “black box” RNNs require  
256 substantial additional efforts to “explain” their behavior).<sup>16,26,27</sup> In the context of BGC mining,  
257 an interpretable model can provide insights into genomic mechanisms of secondary  
258 metabolism; here, introspection of GECCO’s CRF identified numerous intuitive biological and  
259 molecular characteristics that were highly associated with BGC presence. The highly BGC-  
260 enriched GO:0042742 and CL0400 terms (corresponding to "defense response to bacterium"  
261 and "Double-Glycine leader-peptide cleavage motif", respectively), for example, are typical of  
262 bacteriocin RiPPs often exported by ABC transporters,<sup>28</sup> while BGC-enriched CL0149 (“CoA-  
263 dependent acyltransferase superfamily”) and GO:0008299 (“isoprenoid biosynthetic process”)  
264 are associated with polyketide synthases and terpenes, respectively.<sup>29,30</sup> Furthermore, we  
265 identified numerous BGC-associated domains, which had not been included among domain  
266 sets previously associated with secondary metabolism, including three highly BGC-associated  
267 domains of unknown function. These results not only provide insight into BGC architecture  
268 and function, but may be leveraged in the future to improve BGC annotation and identify “high-

269 confidence”, putative novel BGCs, which can be targeted by experimentalists. In conclusion,  
270 GECCO’s CRF-based approach used here showcases that model interpretability and  
271 computational efficiency can be realised with simultaneous gains in accuracy of *de novo* BGC  
272 identification.

## 273 METHODS

274 **Data acquisition and feature construction.** A total of 8,000 randomly selected host-  
275 associated prokaryotic contigs were downloaded from the proGenomes2 v12 database<sup>20</sup>  
276 (<https://progenomes.embl.de/index.cgi>) to serve as candidate BGC-negative instances for  
277 training, CV, and testing (accessed 15 July 2020). A Python implementation of the OrthoANI  
278 algorithm<sup>31</sup> (<https://github.com/althonos/orthoani>) was used to calculate average nucleotide  
279 identity (ANI) values between all pairs of candidate contigs. To eliminate the potential risk of  
280 training data leakage during CV and testing, a diverse subset of these prokaryotic contigs were  
281 selected in which all selected contigs were confirmed to share (i) < 85 ANI with each other and  
282 (ii) < 80 ANI with all contigs in the external test set used by Hannigan, et al.<sup>9</sup> (see section  
283 “Validation of CRF performance on external test data” below).

284 Prodigal v2.6.3<sup>32</sup> was used to identify ORFs within each of the selected contigs in  
285 metagenomic mode (“-p meta”; Supplementary Figure S2). For each contig, the hmmsearch  
286 command in HMMER v3.3.1<sup>33</sup> was used to identify protein domains within the resulting amino  
287 acid sequences, using pHMMs from each of the following databases/combinations of  
288 databases: (i) Pfam v31.0<sup>34</sup>; (ii) Pfam v32.0<sup>34</sup>; (iii) Pfam v33.1<sup>19</sup>; (iv) TIGRFAM v15.0<sup>18</sup>; (v)  
289 PANTHER v15.0<sup>35</sup>; (vi) Pfam v32.0, TIGRFAM v15.0, and PANTHER v15.0; (vii) Pfam  
290 v33.1, TIGRFAM v15.0, and PANTHER v15.0; (viii) Pfam v33.1 and TIGRFAM v15.0; (ix)  
291 Pfam v33.1, TIGRFAM v15.0, ASPeptides (from antiSMASH v5.1)<sup>11</sup>, smCOGs (from  
292 antiSMASH v5.1),<sup>11</sup> and dbCAN v3.0<sup>36</sup>; (x) Pfam v33.1, TIGRFAM v15.0, and Resfams  
293 v1.2<sup>37</sup>; (xi) Pfam v33.1, TIGRFAM v15.0, dbCAN v3.0, smCOGs v5.1, and Resfams v1.2;

294 (xii) Pfam v33.1, TIGRFAM v15.0, and smCOGs v5.1; (xiii) Pfam v33.1, TIGRFAM v15.0,  
295 smCOGs v5.1, and Resfams v1.2; (xiv) Pfam v33.1 and TIGRFAM v15.1 (Supplementary  
296 Figure S2). The resulting ORFs and their respective domains were stored in tabular format and  
297 ordered by their start coordinates (referred to hereafter as the “feature table”), and domains  
298 with an E-value < 1E-5 were maintained. The command-line implementation of antiSMASH  
299 v4.2.0<sup>8</sup> was then used to identify the coordinates of known BGCs in all selected contigs (using  
300 default settings), and ORFs/domains that overlapped with the resulting known BGC regions  
301 were removed from the feature table, yielding a final BGC-negative feature table for each  
302 prokaryotic contig (Supplementary Figure S2).

303 To construct a set of BGC-positive instances, the amino acid sequences and metadata  
304 for all BGCs within MIBiG v2.0<sup>17</sup> (<https://mibig.secondarymetabolites.org/download>) were  
305 downloaded ( $n = 1,923$ ). To prevent training data leakage during testing, the diamond blastp  
306 command in DIAMOND v0.9.13<sup>38</sup> was used to align the amino acid sequences of all genomes  
307 present in the external test data set (see section “Validation of CRF performance on external  
308 test data” below) to the MIBiG BGC amino acid sequences, using minimum amino acid identity  
309 (id) and query coverage thresholds (query-cover) of 50% each, and a maximum E-value  
310 threshold of 1E-5. MIBiG BGCs were removed from the training set if 50% or more of their  
311 amino acid sequences were detected in any test set contigs using DIAMOND and the  
312 aforementioned thresholds, yielding a final set of 1,137 MIBiG v2.0 BGCs for training and  
313 CV. HMMER was used to identify Pfam domains within the amino acid sequences of the BGCs  
314 as described above, producing a BGC-positive feature table for each of 1,137 MIBiG v2.0  
315 BGCs.

316 To construct a final training set that contained both negative and positive BGC  
317 instances, the feature table for a randomly selected MIBiG v2.0 BGC (i.e., a positive instance)  
318 was randomly embedded into the feature table of a randomly selected member of the masked,

319 BGC-negative contigs (i.e., a negative instance). This approach yielded a final set of 1,137  
320 contigs that each contained a single MIBiG v2.0 BGC with known coordinates (Supplementary  
321 Figure S2).

322 **CRF training and cross-validation.** For each pHMM database combination ( $n = 14$ ; see  
323 section “Data acquisition and feature construction” above), a two-state CRF was trained using  
324 the CRF architecture implemented in CRFsuite v0.12.<sup>39</sup> Briefly, for each CRF, features  
325 consisted of an ordered list of Python dictionaries, each containing domains identified in each  
326 amino acid sequence using the respective pHMMs. Output states corresponded to the  
327 probability that a given domain was part of a BGC or not, coded as 1 and 0, respectively (Fig.  
328 1). Additionally, for each pHMM database combination, a feature selection approach was  
329 employed, in which the two-sided Fisher’s Exact Test (FET) implemented in the fisher v0.1.9  
330 Python package (implemented as a Cython extension; <https://pypi.org/project/fisher>) was  
331 nested into training fold(s) and used to identify domains associated with BGC  
332 presence/absence; the top domains that were associated with the binary outcome variable at a  
333 threshold  $T$  after employing a false-discovery rate correction remained in the model. For each  
334 pHMM database combination, values of  $T$  ranging from 0.05 to 1.0 in increments of 0.05 were  
335 tested.

336 Each combination of pHMM database(s)/feature selection threshold  $T$  was evaluated  
337 using ten-fold CV, using the Kfold function in scikit-learn v0.22.1<sup>40</sup> and the sequence ID of  
338 each ORF treated as a group (i.e., to ensure that each ORF was contained within a single fold  
339 and not split across multiple folds; Supplementary Figure S3). For all models that employed it,  
340 the FET feature selection approach was nested into training fold(s) to avoid overfitting.  
341 Optimization of the  $c1$  and  $c2$  CRF hyperparameters (which correspond to L1 and L2  
342 regularization coefficients, respectively) was additionally performed within CV folds, in which  
343 either  $c1$  or  $c2$  was set to 0.15, while the value of the other hyperparameter was set to one of

344 [0, 0.1, 0.15, 1, 2, 10]. Model performance was evaluated using the following metrics, with  
345 scikit-learn and Matplotlib v3.3.4<sup>41</sup> used to construct all curves: (i) per-protein ROC curves;  
346 (ii) per-protein PR curves; (iii)  $F_1$  and (iv) AUPR score versus fraction of FET-selected  
347 features. The model selected as the final CRF to be implemented in GECCO (i.e., the CRF  
348 trained on BGCs derived from MIBiG v2.0, using domains from Pfam v33.1 and TIGRFAM  
349 v15.0, FET inclusion threshold  $T = 0.35$ , and  $c1 = c2 = 0.15$ ; Supplementary Figure S3) was  
350 additionally evaluated using LOTO CV for each MIBiG biosynthetic class, with BGCs  
351 assigned to multiple biosynthetic classes excluded (Supplementary Figure S8).

352 **Validation of CRF performance on external test data.** The (i) six genome test set and (ii)  
353 nine genome bootstrap set used by Hannigan, et al.<sup>9</sup> (see Supplementary Tables S4 and S3 of  
354 Hannigan, et al.<sup>9</sup>, respectively) were used as external test sets to evaluate the performance of  
355 the GECCO CRF (see section “CRF training and cross-validation” above). To construct an  
356 extensive third external test set comprising known BGC and non-BGC regions, BGCs that were  
357 present in MIBiG v2.0 but absent from MIBiG v1.3 were each embedded into a randomly  
358 selected prokaryotic contig as described above (see section “Data acquisition and feature  
359 construction” above; Supplementary Figure S2). For this external test set, DIAMOND was  
360 used to identify potentially redundant BGCs in MIBiG v2.0 that aligned to BGCs in MIBiG  
361 v1.3, using the blastp thresholds described above (see section “Data acquisition and feature  
362 construction” above); contigs that contained these potentially redundant BGCs were removed  
363 from the external test set to avoid training data leakage during testing, yielding a final set of  
364 376 contigs that each contained a single BGC present in MIBiG v2.0 but absent from MIBiG  
365 v1.3 (referred to as the “376-genome test set”).

366 To avoid training leakage into the 376-genome test set, the GECCO CRF (see section  
367 “CRF training and cross-validation” above) was re-trained on BGCs available in MIBiG v1.3  
368 and was used to predict BGC presence/absence in each genome in the three test sets (i.e., the

369 six-, nine-, and 376-genome test sets; Fig. 2 and Supplementary Figures S4-S7). The ability of  
370 each of the following methods to predict BGC presence/absence was additionally evaluated on  
371 each of the three test sets (Fig. 2a-c,e and Supplementary Figure S6): (i) DeepBGC v0.1.18<sup>9</sup>;  
372 (ii) the original ClusterFinder<sup>7</sup> algorithm, implemented in DeepBGC v0.1.18; (iii) the retrained  
373 version of the ClusterFinder algorithm, implemented in DeepBGC v0.1.18 (re-trained on BGCs  
374 available in MIBiG v1.3); (iv) a re-trained implementation of DeepBGC, which was trained on  
375 the exact positive and negative BGC instances used to retrain the GECCO CRF, using BGCs  
376 available in MIBiG v1.3 (using DeepBGC’s “train” function). The re-trained implementation  
377 of DeepBGC (iv) was additionally evaluated relative to the re-trained implementation of the  
378 GECCO CRF (i.e., trained on BGCs from MIBiG v1.3) using 10-fold CV, where both models  
379 were trained and tested on identical folds (see section “CRF training and cross-validation”  
380 above; Supplementary Figure S5). For all models, performance was evaluated using: per-  
381 domain (i) ROC and (ii) PR curves; (iii) segment overlap PR curves (Supplementary Figure  
382 S4), using minimum overlap thresholds of 25, 50, and 75% (Fig. 2a-c, Supplementary Figure  
383 S6-S7).

384 **Prediction of biosynthetic class.** To assign the BGCs that the GECCO CRF predicted to one  
385 or more of the six biosynthetic classes in MIBiG v2.0 (with MIBiG’s “Other” class excluded),  
386 the following classifiers were trained (Supplementary Figure S9): (i) a random forest classifier,  
387 using the scikit-learn RandomForestClassifier function; (ii) an ExtraTrees classifier, using the  
388 scikit-learn ExtraTreesClassifier function; (iii) a *k*-nearest neighbors (kNN) classifier, using  
389 the scikit-learn KNeighborsClassifier function, a cosine distance metric, and number of  
390 neighbors  $n = 3$ ; (iv) the aforementioned kNN, with  $n = 15$ . For each classifier, BGCs were  
391 represented by compositional vectors, where individual features corresponded to the fraction  
392 of a particular domain present in the BGC. For example, a predicted BGC with 2 domains *A*,  
393 one domain *B*, and one domain *C* would be represented by domain composition vector [*A*: 0.5,

394  $B: 0.25, C: 0.25]$ , assuming  $A$ ,  $B$ , and  $C$  are the only possible domains. The ability of each  
395 classifier to predict MIBiG biosynthetic class was evaluated using five-fold CV via the  
396 cross\_val\_predict function in scikit-learn, and the random forest was implemented as the final  
397 biosynthetic classifier in GECCO (Supplementary Figure S9).

398 **BGC identification in prokaryotic genomes.** Each of the following methods was used to  
399 identify BGCs in all representative genomes available in the proGenomes2 v12 database<sup>20</sup> ( $n$   
400 = 12,221; accessed 15 July 2020): (i) the GECCO CRF trained on BGCs available in MIBiG  
401 v2.0 (i.e., the final model implemented in GECCO, run using default parameters); (ii)  
402 antiSMASH v4.2.0<sup>8</sup> (run using default parameters); (iii) DeepBGC v0.1.18<sup>9</sup> (run using default  
403 parameters with the addition of DeepBGC’s “--prodigal-meta-mode” option, as GECCO uses  
404 this option for BGC detection by default; Fig. 3ab). antiSMASH-to-MIBiG type mappings  
405 from BiG-SLiCE v1.1.0<sup>21</sup> were used to map antiSMASH biosynthetic types to MIBiG  
406 biosynthetic types (used by DeepBGC and GECCO; Supplementary Table S2). The three  
407 aforementioned BGC detection/classification methods were additionally applied to the  
408 following data sets to assess their speed using a single CPU (Fig. 2f and Supplementary Figure  
409 S10): (i) contigs in each of the three test sets (i.e., the six-, nine-, and 376-genome test sets,  $n$   
410 = 395 contigs; see section “Validation of CRF performance on external test data” above); (ii)  
411 the 12,221 proGenomes2 representative genomes. Plots were constructed in R v3.6.1<sup>42</sup> using  
412 ggplot2 v3.3.3.<sup>43</sup>

413 **Comparison of GECCO and BiG-SLiCE domain sets.** Domains that were included in the  
414 optimized GECCO pHMMs based on their FET association with BGC presence/absence (see  
415 section “CRF training and cross-validation” above) were compared to protein domains used by  
416 BiG-SLiCE v1.1.0.<sup>21</sup> BiG-SLiCE, which is designed to cluster antiSMASH BGCs into Gene  
417 Cluster Families, relies on a set of core biosynthetic domains for BGC annotation and  
418 clustering. Domains within the GECCO pHMMs were compared to all publicly available core

419 biosynthetic BiG-SLiCE domains with a reported accession number, as well as BiG-SLiCE's  
420 larger set of "BioPfam" domains (identical to the Pfam v33.1 database) by finding the union of  
421 the three domain sets and plotting via venn.js (<https://github.com/benfred/venn.js/>) and  
422 Matplotlib (Fig. 3c and Supplementary Table S1). Three independent, two-group Mann-  
423 Whitney *U* tests were used to compare CRF weights associated with the following GECCO  
424 domain sets, using the "wilcox.test" function in R, with parameters set to perform an unpaired  
425 (paired = F), two-sided (alternative = "two.sided") test using a normal approximation (exact =  
426 F) and a continuity correction (correct = T) : (i) GECCO domains included in BiG-SLiCE's  
427 core biosynthetic domain set; (ii) GECCO Pfam domains excluded from BiG-SLiCE's core  
428 biosynthetic domain set; (iii) GECCO Tigrfam domains excluded from BiG-SLiCE's core  
429 biosynthetic domain set. Tests between groups (i)/(iii) and (ii)/(iii) were statistically significant  
430 after a Bonferroni correction (raw *P* = 3.12e-07 and 5.75e-06, respectively), but not groups  
431 (i)/(ii) (raw *P* = 0.10).

432 **GO term enrichment.** Weights associated with each protein domain were extracted from the  
433 trained GECCO CRF instance, and all available GO terms for each domain were retrieved from  
434 InterPro (*n* = 2,722 domains with one or more assigned GO terms, out of 5,255 total  
435 domains).<sup>22,44</sup> To identify over-represented GO terms associated with BGC presence (i.e.,  
436 BGC-enriched GO terms), domains were assigned ranks based on their weights, where the  
437 domain with the highest weight (i.e., PF14867, with weight 4.190953) was assigned a value of  
438 "1", and the domain with the lowest weight (i.e., PF02881, with weight -1.798162) was  
439 assigned a value of "2722". For each of the (i) Biological Process and (ii) Molecular Function  
440 GO ontologies, the runTest function in the topGO v2.36.0 package<sup>45</sup> in R v3.6.1 was used to  
441 perform a Kolmogorov-Smirnov (KS) test (statistic = "ks"), using the "weight01" algorithm  
442 (algorithm = "weight01") to account for the GO graph topology.<sup>45</sup> Enrichment scores were  
443 calculated for all statistically significant (*P* < 0.05) GO terms by negating the base-10 logarithm

444 of the resulting *P*-values. The aforementioned steps were repeated to identify over-represented  
445 GO terms associated with BGC-absence, using (i) domains ranked by weight from lowest-to-  
446 highest (i.e., the domain with the lowest weight was assigned a value of “1”, and the domain  
447 with the highest weight was assigned a value of “2722”) and (ii) enrichment scores  
448 corresponding to the non-negated base-10 logarithms of the resulting *P*-values. topGO’s  
449 weight01 algorithm calculates the *P*-value of a GO term conditioned on neighbouring GO  
450 terms; therefore, tests were considered not independent, and *P*-values were interpreted as  
451 inherently corrected.<sup>45</sup> Enrichment scores were plotted using the ggplot2 package in R (Fig. 3d  
452 and Supplementary Figure S11).

453 **Pfam clan enrichment.** Weights associated with each Pfam protein domain were extracted  
454 from the trained GECCO CRF instance, and all available Pfam domain-to-clan mappings were  
455 retrieved for Pfam v33.1 via FTP ( $n = 1,907$  Pfam domains with an assigned clan, out of 2,873  
456 total Pfam domains).<sup>19</sup> A (i) vector of raw Pfam domain weights (ordered from highest-to-  
457 lowest) and (ii) list of clan-to-domain mappings were supplied to the fgsea function from the  
458 fgsea v1.10.1 R package<sup>46,47</sup>, which was used to identify BGC- and non-BGC-enriched Pfam  
459 clans, using 1 million permutations ( $nperm = 1000000$ ), a minimum clan size of three ( $minSize$   
460 = 3), and no maximum clan size limit. For significantly enriched clans (false discovery rate-  
461 corrected  $P < 0.10$ ), ggplot2 was used to plot (i) fgsea normalized enrichment scores (NES)  
462 and (ii) the negated base-10 logarithm of the false discovery rate-corrected *P*-values (Fig. 3e  
463 and Supplementary Figure S12)

464 **Data availability.** Training and test data can be downloaded from  
465 <https://github.com/zellerlab/GECCO/releases/tag/v0.6.0>. GECCO CRF weights are available  
466 in Supplementary Table S1.

467 **Code availability.** GECCO code is free and publicly available at <https://gecco.embl.de>.

468 **ACKNOWLEDGMENTS**

469 We are grateful to Tobias Gulder, Maximilian Hohmann, and members of the Zeller Team for  
470 fruitful discussions, as well as EMBL IT Services for support with high-performance  
471 computing. This work was funded by the European Molecular Biology Laboratory, the German  
472 Research Foundation (Deutsche Forschungsgemeinschaft, DFG, grant no. 395357507 – SFB  
473 1371), and the German Federal Ministry of Education and Research (BMBF, grant no.  
474 031L0181A, and the de.NBI network, grant no. 031A537B).

475 **AUTHOR CONTRIBUTIONS**

476 Software development and computational analyses were performed by JSF, ML, and LMC with  
477 contributions of data or tools from all authors. GZ conceived and funded the study. LMC and  
478 GZ co-wrote the manuscript with input from all authors.

479 **COMPETING INTERESTS**

480 The authors declare no competing interests.

481

482

483

484

485

486

487

488

489

490

491

492

493

494 **REFERENCES**

495 1 Medema, M. H. *et al.* Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol* **11**, 625-631, doi:10.1038/nchembio.1890 (2015).

496 2 Tyc, O., Song, C., Dickschat, J. S., Vos, M. & Garbeva, P. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends Microbiol* **25**, 280-292, doi:10.1016/j.tim.2016.12.002 (2017).

497 3 Milshteyn, A., Colosimo, D. A. & Brady, S. F. Accessing Bioactive Natural Products from the Human Microbiome. *Cell Host Microbe* **23**, 725-736, doi:10.1016/j.chom.2018.05.013 (2018).

498 4 Sharon, G. *et al.* Specialized metabolites from the microbiome in health and disease. *Cell Metab* **20**, 719-730, doi:10.1016/j.cmet.2014.10.016 (2014).

499 5 Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* **83**, 770-803, doi:10.1021/acs.jnatprod.9b01285 (2020).

500 6 Eder, J., Sedrani, R. & Wiesmann, C. The discovery of first-in-class drugs: origins and evolution. *Nat Rev Drug Discov* **13**, 577-587, doi:10.1038/nrd4336 (2014).

501 7 Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412-421, doi:10.1016/j.cell.2014.06.034 (2014).

502 8 Blin, K. *et al.* antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res* **45**, W36-W41, doi:10.1093/nar/gkx319 (2017).

503 9 Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research* **47**, e110-e110, doi:10.1093/nar/gkz654 (2019).

519 10 Milshteyn, A., Schneider, J. S. & Brady, S. F. Mining the metabiome: identifying novel  
520 natural products from microbial communities. *Chem Biol* **21**, 1211-1223,  
521 doi:10.1016/j.chembiol.2014.08.006 (2014).

522 11 Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining  
523 pipeline. *Nucleic Acids Res* **47**, W81-W87, doi:10.1093/nar/gkz310 (2019).

524 12 Skinnider, M. A., Merwin, N. J., Johnston, C. W. & Magarvey, N. A. PRISM 3:  
525 expanded prediction of natural product chemical structures from microbial genomes.  
526 *Nucleic Acids Res* **45**, W49-W54, doi:10.1093/nar/gkx320 (2017).

527 13 Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**,  
528 1735–1780, doi:10.1162/neco.1997.9.8.1735 (1997).

529 14 Lafferty, J. D., McCallum, A. & Pereira, F. C. N. in *Proceedings of the Eighteenth*  
530 *International Conference on Machine Learning* 282–289 (Morgan Kaufmann  
531 Publishers Inc., 2001).

532 15 Yu, B. & Fan, Z. A comprehensive review of conditional random fields: variants,  
533 hybrids and applications. *Artificial Intelligence Review* **53**, 4289-4333,  
534 doi:10.1007/s10462-019-09793-6 (2020).

535 16 Rudin, C. Stop explaining black box machine learning models for high stakes  
536 decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206-  
537 215, doi:10.1038/s42256-019-0048-x (2019).

538 17 Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known  
539 function. *Nucleic Acids Res* **48**, D454-D458, doi:10.1093/nar/gkz882 (2020).

540 18 Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41**,  
541 D387-395, doi:10.1093/nar/gks1234 (2013).

542 19 Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**,  
543 D412-D419, doi:10.1093/nar/gkaa913 (2021).

544 20 Mende, D. R. *et al.* proGenomes2: an improved database for accurate and consistent  
545 habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids*  
546 *Res* **48**, D621-D625, doi:10.1093/nar/gkz1002 (2020).

547 21 Kautsar, S. A., van der Hooft, J. J. J., de Ridder, D. & Medema, M. H. BiG-SLiCE: A  
548 highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters.  
549 *Gigascience* **10**, doi:10.1093/gigascience/giaa154 (2021).

550 22 The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still  
551 GOing strong. *Nucleic Acids Research* **47**, D330-D338, doi:10.1093/nar/gky1055  
552 (2018).

553 23 Lopatkin, A. J. & Collins, J. J. Predictive biology: modelling, understanding and  
554 harnessing microbial complexity. *Nat Rev Microbiol* **18**, 507-520,  
555 doi:10.1038/s41579-020-0372-5 (2020).

556 24 Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat Med* **25**, 24-29,  
557 doi:10.1038/s41591-018-0316-z (2019).

558 25 Myszczynska, M. A. *et al.* Applications of machine learning to diagnosis and  
559 treatment of neurodegenerative diseases. *Nat Rev Neurol* **16**, 440-456,  
560 doi:10.1038/s41582-020-0377-8 (2020).

561 26 Topcuoglu, B. D., Lesniak, N. A., Ruffin, M. T. t., Wiens, J. & Schloss, P. D. A  
562 Framework for Effective Application of Machine Learning to Microbiome-Based  
563 Classification Problems. *mBio* **11**, doi:10.1128/mBio.00434-20 (2020).

564 27 Quinn, T. P. & Erb, I. Examining microbe-metabolite correlations by linear methods.  
565 *Nat Methods* **18**, 37-39, doi:10.1038/s41592-020-01006-1 (2021).

566 28 Beis, K. & Rebuffat, S. Multifaceted ABC transporters associated to microcin and  
567 bacteriocin export. *Res Microbiol* **170**, 399-406, doi:10.1016/j.resmic.2019.07.002  
568 (2019).

569 29 Keatinge-Clay, A. T. The Uncommon Enzymology of Cis-Acyltransferase Assembly  
570 Lines. *Chemical Reviews* **117**, 5334-5366, doi:10.1021/acs.chemrev.6b00683  
571 (2017).

572 30 Hillier, S. G. & Lathe, R. Terpenes, hormones and life: isoprene rule revisited. *J  
573 Endocrinol* **242**, R9-R22, doi:10.1530/JOE-19-0084 (2019).

574 31 Lee, I., Ouk Kim, Y., Park, S.-C. & Chun, J. OrthoANI: An improved algorithm and  
575 software for calculating average nucleotide identity. *International Journal of*  
576 *Systematic and Evolutionary Microbiology* **66**, 1100-1103,  
577 doi:<https://doi.org/10.1099/ijsem.0.000760> (2016).

578 32 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site  
579 identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

580 33 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195,  
581 doi:10.1371/journal.pcbi.1002195 (2011).

582 34 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res*  
583 **47**, D427-D432, doi:10.1093/nar/gky995 (2019).

584 35 Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based  
585 classification tool, enhancer regions and extensive API. *Nucleic Acids Res* **49**, D394-  
586 D403, doi:10.1093/nar/gkaa1106 (2021).

587 36 Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme  
588 annotation. *Nucleic Acids Res* **40**, W445-451, doi:10.1093/nar/gks479 (2012).

589 37 Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic  
590 resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* **9**,  
591 207-216, doi:10.1038/ismej.2014.106 (2015).

592 38 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using  
593 DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

594 39 CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007).

595 40 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*  
596 **12**, 2825–2830 (2011).

597 41 Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science &*  
598 *Engineering* **9**, 90-95, doi:10.1109/MCSE.2007.55 (2007).

599 42 R: A Language and Environment for Statistical Computing v. 3.6.1 (R Foundation for  
600 Statistical Computing, Vienna, Austria, 2019).

601 43 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New  
602 York, 2016).

603 44 Blum, M. *et al.* The InterPro protein families and domains database: 20 years on.  
604 *Nucleic Acids Res* **49**, D344-D354, doi:10.1093/nar/gkaa977 (2021).

605 45 Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups  
606 from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**,  
607 1600-1607, doi:10.1093/bioinformatics/btl140 (2006).

608 46 Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv*, 060012,  
609 doi:10.1101/060012 (2021).

610 47 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach  
611 for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**,  
612 15545-15550, doi:10.1073/pnas.0506580102 (2005).

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627 **FIGURE LEGENDS**

628 **Figure 1.** Graphical depiction of the biosynthetic gene cluster (BGC) identification and  
629 classification approach developed here and implemented in GECCO (GEne Cluster prediction  
630 with COnditional random fields). Briefly, GECCO identifies open reading frames (ORFs) in  
631 an assembled prokaryotic (meta)genome (Step 1). Protein domains are annotated in the  
632 resulting ORFs using profile hidden Markov models (pHMMs; Step 2). The resulting ordered  
633 domain vectors are treated as features, and a conditional random field (CRF) is used to predict  
634 whether each feature belongs to a BGC or not (Step 3). Predicted BGCs are classified into one  
635 of six major biosynthetic classes as defined in the Minimum Information about a Biosynthetic  
636 Gene cluster (MIBiG) database using a Random Forest classifier (Step 4).

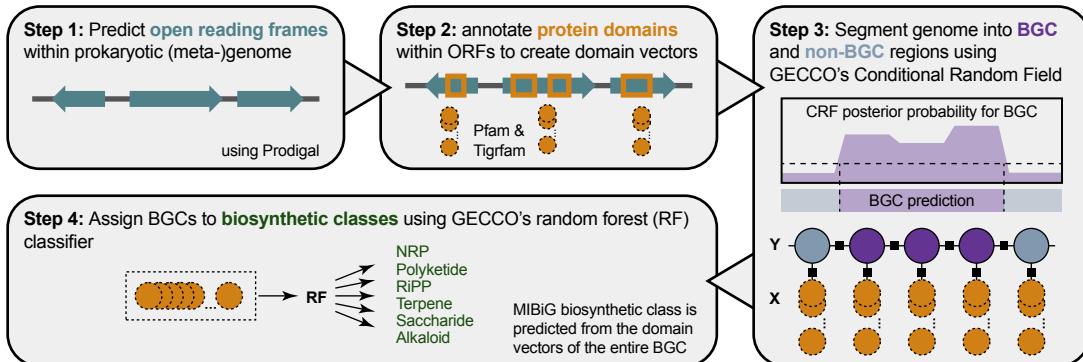
637

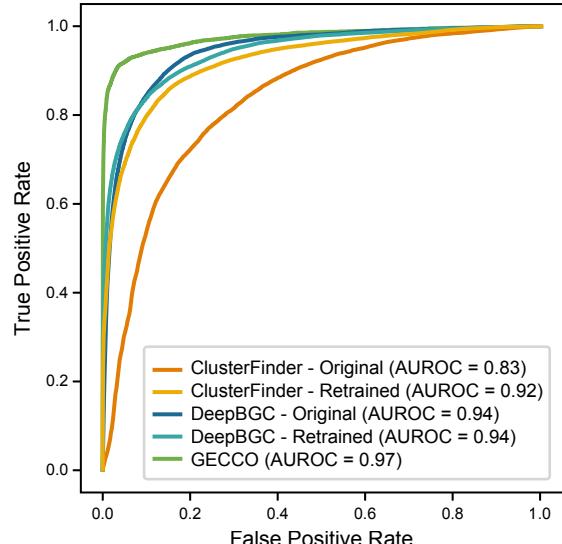
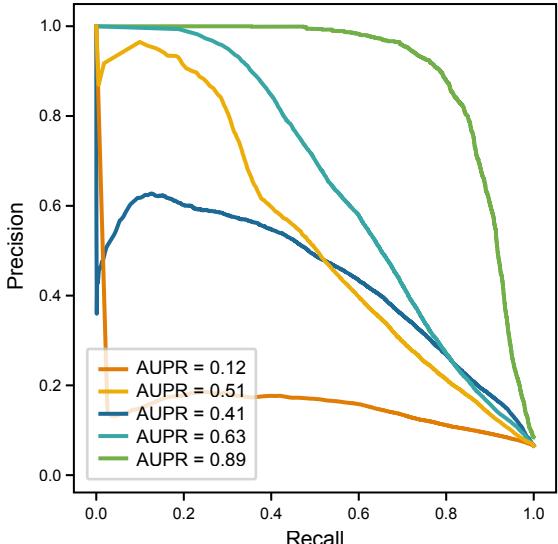
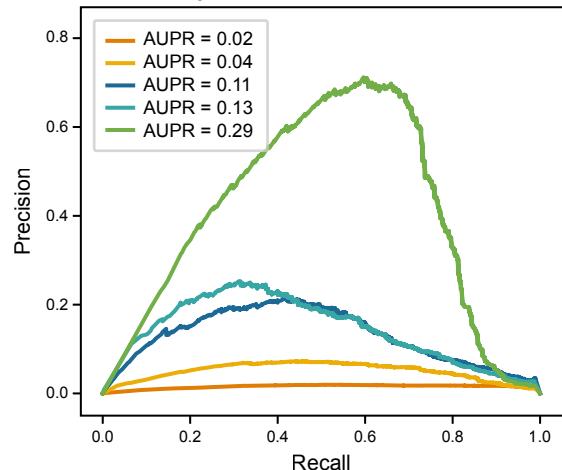
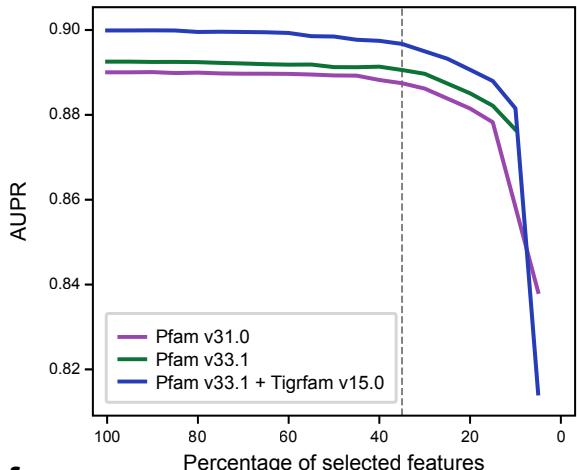
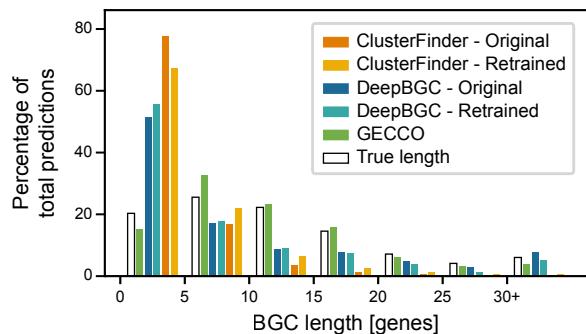
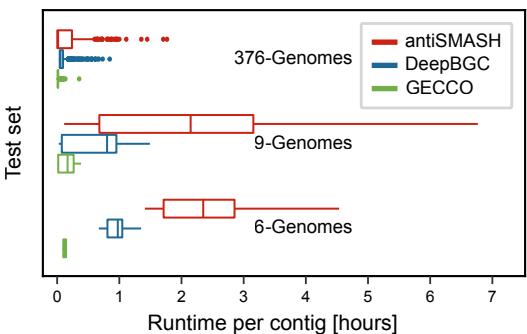
638 **Figure 2.** **(a)** Domain-level receiver-operating characteristic (ROC) curves, **(b)** domain-level  
639 precision-recall (PR) curves comparing original and retrained implementations of  
640 ClusterFinder (ClusterFinder-Original and ClusterFinder-Retrained, respectively) and  
641 DeepBGC (DeepBGC-Original and DeepBGC-Retrained, respectively) with GECCO (trained  
642 on a subset of Pfam v33.1 and Tigrfam v15.0 domains). **(c)** PR curves calculated from segment  
643 overlap (>50%) of predicted and known BGCs (see Supplementary Figures S4 and S7). All  
644 models (a-c) were trained on BGCs from MIBiG v1.3, evaluated on BGCs from MIBiG v2.0  
645 not contained in v1.3 (i.e., the 376-genome test set); area under the curve (AUROC and AUPR)  
646 values are reported in legends. **(d)** AUPR values (Y-axis) versus percentage of Fisher's Exact  
647 Test-selected features ( $T$ ; X-axis) included in CRFs trained on BGCs from MIBiG v2.0, using  
648 domains from several (combinations of) databases (see inset). The default value of  $T$  chosen  
649 for GECCO is denoted by the dashed line ( $T = 0.35$ ). **(e)** Histogram of predicted BGC lengths  
650 (in number of genes; X-axis) relative to true lengths among genomes in the 376-genome test  
651 set. The Y-axis denotes the percentage of total BGC predictions for each method. **(f)** Runtime

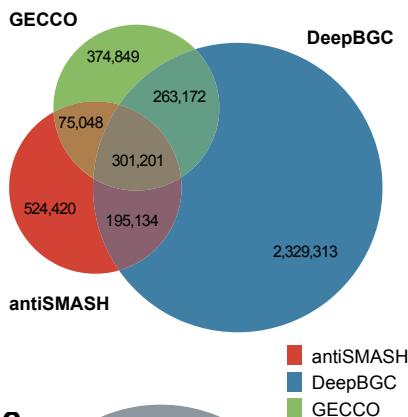
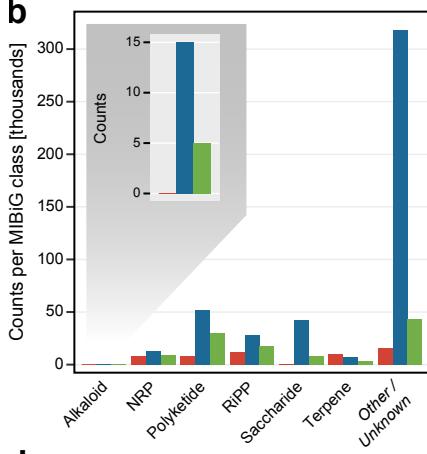
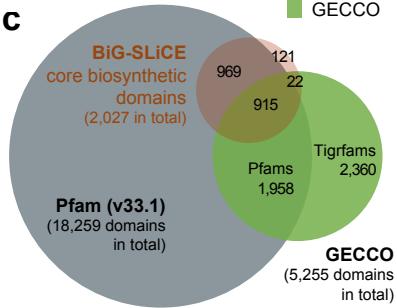
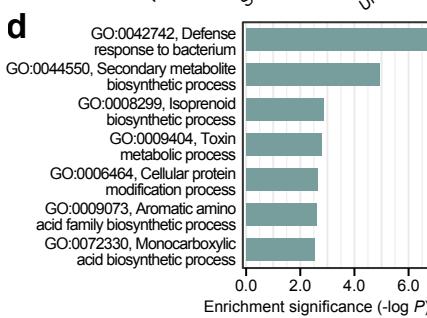
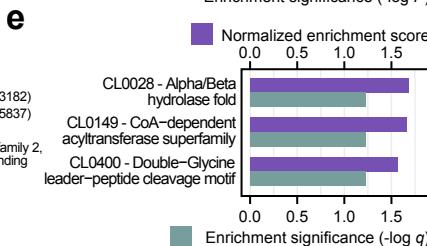
652 per contig required to detect and classify BGCs in each test set using antiSMASH, DeepBGC,  
653 and GECCO.

654

655 **Figure 3. (a)** Venn diagram of biosynthetic gene cluster (BGC) overlap, constructed using the  
656 presence and absence of individual genes in BGCs identified in 12,221 representative microbial  
657 genomes available in the proGenomes2 database using each of antiSMASH, DeepBGC, and  
658 GECCO. If a gene was contained within BGC predictions of more than one method, it was  
659 counted in the respective intersection area. **(b)** Predicted MIBiG biosynthetic classes (X-axis)  
660 associated with BGCs identified in the same 12,221 genomes using each of antiSMASH,  
661 DeepBGC, and GECCO. The Y-axis denotes the number of BGCs assigned to a given  
662 biosynthetic class. BGCs assigned to multiple classes are omitted. **(c)** Venn diagram and  
663 boxplots of GECCO CRF weights (X-axis), constructed using protein domains used by (i)  
664 GECCO, (ii) BiG-SLiCE, and (iii) and Pfam v33.1. GECCO domains were derived from either  
665 Pfam v33.1 or Tigrfam v15.0 and were selected based on their association with BGC  
666 presence/absence using Fisher's Exact Test (FET) and an FET-inclusion threshold ( $T$ ) of 35%  
667 ( $T = 0.35$ ). BiG-SLiCE domains correspond to those present in the core biosynthetic domain  
668 set used by BiG-SLiCE v1.1.0. **(d)** Top Gene Ontology (GO) terms (Y-axis) enriched in BGCs,  
669 obtained using the Kolmogorov-Smirnov test/weight01 algorithm implemented in topGO  
670 (enrichment significance  $> 2.75$ ). **(e)** Pfam clans (Y-axis) enriched in BGCs (X-axis; false  
671 discovery rate [FDR]-adjusted  $P < 0.10$ ). Normalized Enrichment Scores (NES) were obtained  
672 using the fgsea R package. For (d and e), enrichment significance values correspond to the  
673 negated base-10 logarithm of each term's  $P$ -value.



**a Domain-level evaluation****b Domain-level evaluation****c BGC overlap evaluation****d****e****f**

**a****b****c****d****e**

GECCO Domain Subsets

