FunGeneTyper: An Expandable Deep Learning-Based Framework for Ultra-Accurate Classification and Discovery of Functional Protein-Coding Genes

Guoqing Zhang[#1,2,3], Hui Wang[#1], Lu Zhang[1,2,3], Zhiguo Zhang[1,2,3], Guibing Guo[4], Jian Yang[5,6], Fajie Yuan[1,2,3]*, Feng Ju[1,2,3,5,6]*

[1]Research Center for Industries of the Future (RCIF), School of Engineering, Westlake University, Hangzhou, 310030, Zhejiang Province, China.

[2] Center of Synthetic Biology and Integrated Bioengineering, School of Engineering, Westlake University, Hangzhou 310030, Zhejiang Province, China

[3] Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, 310024, Zhejiang Province, China.

[4] Software College, Northeastern University, Shenyang, China

[5] School of Life Sciences, Westlake University, Hangzhou, China

[6] Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China

[#] The authors contributed equally to the work

*Correspondence to

Dr. Feng Ju, Email: jufeng@westlake.edu.cn

Dr. Fajie Yuan, Email: yuanfajie@westlake.edu.cn

Address:

Westlake University, 600 Dunyu Road, Hangzhou 310030, China

Tel.: 571-87963205 (lab), 571-87380995 (office)

Fax: 0571-85271986

Abstract

Next-generation metagenomic sequencing technologies have generated large quantities of protein-coding gene sequence data from microbial dark matter. However, accurately assigning protein functions to new gene sequences remains challenging. To this end, we developed FunGeneTyper, an expandable deep learning-based framework with models, structured databases, and tools for ultra-accurate (>0.99) and fine-grained classification of both antibiotic resistance genes (ARGs) and virulence factor genes. This new framework also achieves superior performance in discovering new ARGs from human gut (accuracy: 0.8512; and F1-score: 0.6948), wastewater (0.7273; 0.6072), and soil (0.8269; 0.5445) samples, beating the state-of-the-art bioinformatics approaches and protein sequence (F1-score: 0.0556-0.5065) and domain (0.2630-0.5224)-based alignment tools. We empowered the generalized application of the framework by implementing a lightweight, privacy-preserving and plug-and-play neural network module shareable among global developers and users. The FunGeneTyper* is released to promote the global discovery of novel and precious microbial, genetic, and enzymatic resources in environmental and animal microbiomes.

Keywords: Functional classification, Protein-coding gene, Deep learning, Structured database, Microbiome, Bioinformatics

* The codes and database resources are available at: https://github.com/emblab-westlake/FunGeneTyper.

45 **Main**

46 High-throughput DNA sequencing and metagenomics have generated extensive
47 protein-coding gene (PCG) sequences from diverse environmental and human
48 microbiomes[1-3]. Accurate classification of genes into related protein functions is the
49 key to effective gene discovery. However, these datasets pose significant computational
50 challenges in metagenomic studies. Sequence alignment (SA), implemented using
51 NCBI's BLAST[4], usearch[5], and Diamond[6], is commonly used for functional annotation
52 of PCGs[7]. To minimize false-positives, SA-based methods are routinely conducted with
53 strict user-defined cutoffs or thresholds (alignment identity, coverage, and bit scores) to
54 retain high-confidence best hits for each query sequence from validated databases. This
55 practice is widely implemented in the development of tools for categorizing genes,
56 including antibiotic resistance genes (ARGs)[8,9] and virulence factor genes (VFGs)[10].
57 SA-based approaches effectively predict functions between genes that share high
58 homology (>80% identity [8,9]), but exclude distantly homogeneous genes that fall below
59 arbitrarily-defined and one-size-fits-all cutoffs that may represent the majority of
60 targeted functional genes in environmental samples (e.g. core ARGs in activated
61 sludge[11] and soil[12]). Therefore, these SA approaches with stringent bioinformatic
62 cutoffs unavoidably generate numerous false-negative results and heavily
63 underestimate true novelties and diversity of functional genes in largely uncultured
64 bacteria, thus biasing research outcomes or conclusions. Therefore, it is crucial to
65 develop intelligent and accurate classification paradigm and bioinformatic tools to

66 overcome limitations of existing SA-based classification approaches. Importantly, this
67 endeavor will accelerate discovery of new genes in future metagenomic-based
68 environmental and human microbiome studies[13,14].

69 Hidden Markov models (HMM) with manual-crafted sequence alignments and
70 scoring functions are powerful tools for protein domain-based functional gene
71 annotation for detecting remote gene homologues with low sequence identity (< 30%)
72 to known proteins[15,16]. However, these methods rely on token (amino acid) matching,
73 which fail to detect high-level semantic representation similarity or structure-level
74 representation similarity, leading to false-positives[17], and thus cannot distinguish
75 functions of proteins in the same family[18]. In contrast, deep learning (DL) methods
76 excel at learning rich and high-level semantic representations when sufficient training
77 data are available, and are effective at identifying proteins with structural and functional
78 similarities[19-22]. Specifically, ground-breaking big language models initially developed
79 for natural language processing tasks have been successfully applied to protein function
80 prediction tasks[23,24], often termed protein language models (PLMs). The high-level
81 semantic representations learned from PLMs establish valid connections between
82 sequences and function[25,26]. Notwithstanding the power of PLMs, gene classification
83 tasks, particularly identifying fine-grained protein function subclasses, pose challenges
84 for data-hungry deep learning paradigms because of limited supervised training dataset
85 for certain genes. Additionally, it remains unclear whether advanced PLMs perform
86 better than state-of-the-art metagenomic bioinformatics tools at microbiome gene

classification and discovery.

Here, we propose FunGeneTyper, a PLM-based deep-learning framework for accurate and expandable prediction of PCG function. FunGeneTyper implements a two-stage pipeline that separately handles the assignment of the main types and subtypes of PCG functional classes, reducing issues associated with insufficient training data during subtype-level predictions. To improve conciseness, it first performs standard classification of genes of the main types and then performs fine-grained retrieval by comparing similarities between learned protein subtype representations. FunGeneTyper models classify ARGs with ultra-high accuracy (>0.99) and outperforms the state-of-the-art SA and HMM-based methods and tools. Furthermore, we also demonstrate the generalized application of FunGeneTyper models in ultra-high classification of VFGs and introduce the adapter module, a lightweight neural network that can be inserted into the current backbone architecture to realize parameter-efficient training. The adapter-tuning-based FunGeneTyper models are expandable to the classification of various categories of genes and enables sharing of both task-agnostic and task-specific parameters without accessing the private training dataset. Thus, FunGeneTyper offers a unified and innovative way of integrating the global efforts of the microbiome and bioinformatics communities, endowing the FunGeneTyper framework with the ability to conduct unlimited prediction of functional gene categories beyond the ARGs and VFGs demonstrated here, which is key to accelerating the global discovery of new and precious genetic and enzymatic resources from microbiomes.

Results

FunGeneTyper framework, structured database, and deep learning models

FunGeneTyper is the first unified framework that utilizes DL models and structured functional gene datasets (SFGD) to develop new DL-based classifiers for any gene category via transfer learning. This novel framework achieves highly accurate PCG classification from metagenomic studies and extends the models to efficiently predict broad categories of gene functions from large varieties of microbiomes with corresponding customizable SFGD.

Structured functional gene datasets

We deployed a transferable strategy to collect high-quality reference gene sequences to meet FunGeneTyper's training requirements with high reliability (Fig 1a). Experimentally-confirmed reference sequences of target genes from literature and/or expert-curated databases were used as the core dataset, and highly homologous protein sequences (at least 80% identity and 80% coverage) were extracted from Uniref100 database and used as the expanded functional genes dataset. A non-target sequence dataset was selected from Swiss-Prot database (version: June 2021) by excluding perfect matches to the target genes, and used as the negative training set so that FunGeneTyper could learn sufficient features of non-target genes. Core and expanded functional gene datasets and the non-target dataset were integrated to form the SFGD, which was organized hierarchically into a secondary structure based on gene ontology. The SFGD was divided into training, validation, and testing sets (ratio 6:2:2) and used

129 to train the following two DL models.

130 **Deep learning models:** The framework has a top-down protein function prediction
131 workflow featuring two DL models (Fig. 1b), FunTrans and FunRep, which
132 progressively classify protein sequences from the upper (type) to lower (subtype)
133 functional levels. FunGeneTyper was pre-trained on ESM-1b, a large-scale pre-trained
134 protein sequence model based on the transformer architecture released by Facebook[22].
135 ESM-1b is composed of the 33-layer transformer architecture consisting of 650 million
136 parameters trained on Uniref50. It has the superior capacity to infer fundamental
137 structural and functional characteristics of proteins from sequences that can
138 significantly increase the performance metrics for sequence-function tasks. Despite
139 sharing the 33-layer transformer architecture, FunTrans and FunRep were
140 independently constructed, trained, and optimized, to complete two-level functional
141 classification tasks that successively assigned a PCG to its best matches of functional
142 type and subtype in a structured database, respectively.

143 FunTrans distinguishes protein sequences and classify them into specific functional
144 types equivalent to gene families with the same or similar functions. It is inspired by
145 the fact that proteins with similar structures and functions clustered closer together in
146 the embedding space. The main FunTrans structure is a 33-layer transformer that
147 implements initial classification of input data (Fig 1c). Adapter modules are inserted
148 into each transformer block as trainable parameters. The adapter enables efficient fine-
149 tuning of parameters for different gene classification tasks. High parameter sharing is

150 achieved under the premise that the parameters of the original network remain
151 unchanged. Adapter modules enable flexible and parameter-efficient transfer learning
152 and prevent overfitting[27]. FunTrans adds a nonlinear classification layer at the end of
153 the sequence semantic representation for functional classification.

154 FunRep has a structure similar to that of FunTrans and can be used to embed
155 representation retrieval for further subtype classification of protein function (Fig 1d). It
156 uses embedding representation retrieval to accurately predict functional subtypes of the
157 FunTrans output results for classification. FunRep also adds an adapter layer to increase
158 robustness and insight into a broader range of gene classification.

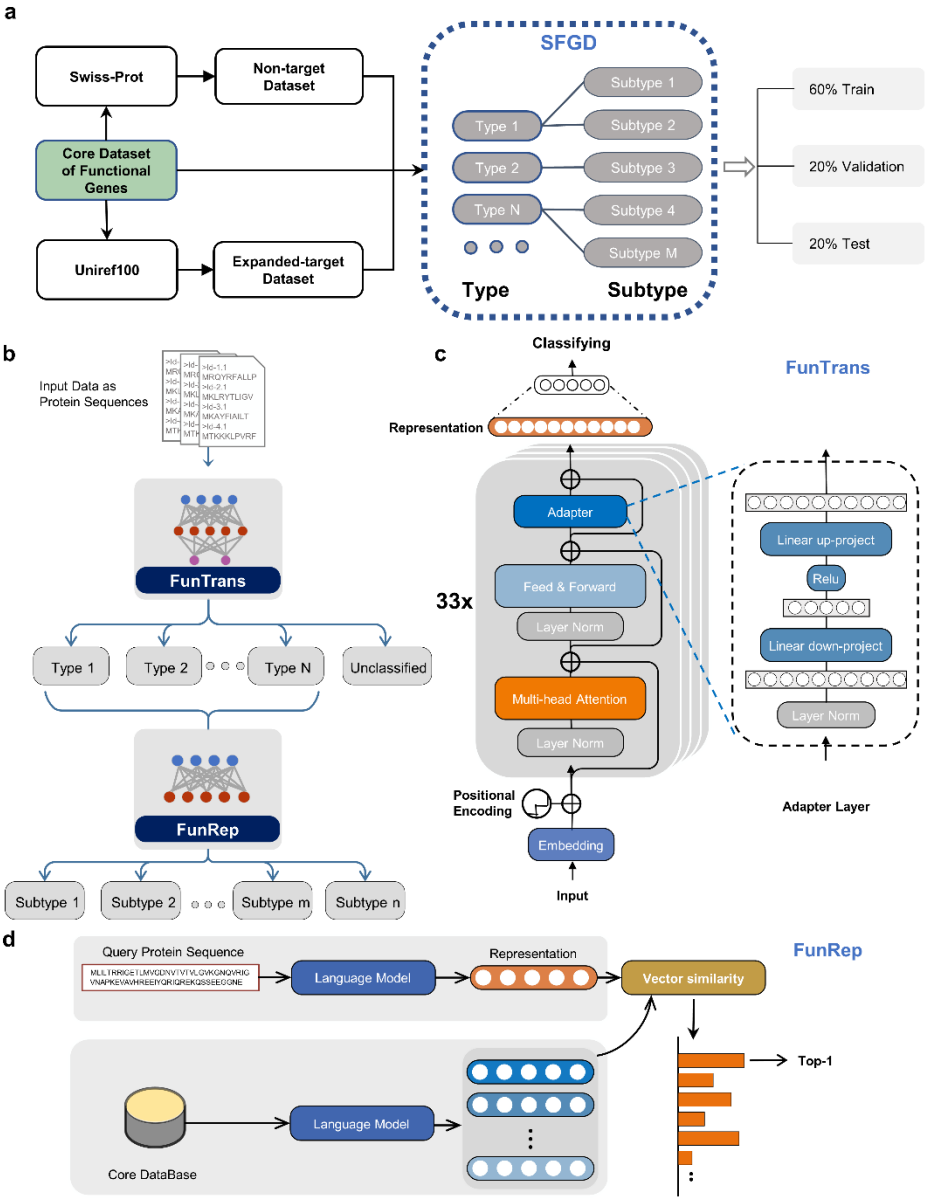


Fig. 1 FunGeneTyper model design and database construction workflows. **a**, Two-level hierarchical structure of FunGeneTyper. **b**, Schematic representation of FunTrans model. **c**, Schematic representation of FunRep model. **d**, Process of preparing a structured functional gene dataset (SFGD). The data set is divided into the training set, validation set and testing set in a 6:2:2 ratio.

166 **FunGeneTyper classification performance and learning ability**

167 The spread of antibiotic resistance has raised public health concerns globally[28]. Reliable
168 ARG model classification is important for surveillance and control of the spread of
169 antibiotic resistance, and achieving sufficient model sensitivity to remote homologues
170 is key to discovering new ARGs. Therefore, we first classified ARGs and demonstrated
171 the ability of the FunGeneTyper framework to achieve this goal. Before building the
172 ARGs classification models, we constructed a hierarchical structured ARG database
173 (SARD) based on antibiotic resistance ontology of the comprehensive antibiotic
174 resistance database (CARD)[7]. Based on CARD's ontological rules, ARGs were
175 assigned to class and group hierarchies based on the types of drugs to which they confer
176 resistance, and the subtypes of genes with the same resistance function, respectively
177 (Dataset S1). To test and improve the sensitivity of the model, we used different identity
178 thresholds to collect four non-target sequence sets from Swiss-Prot database —
179 excluding ARGs — as negative training datasets, for model training (Supplementary
180 Figure 1, see Methods). The addition of a negative training set allows the model to learn
181 features of non-targeted genes., which gives the model the ability to directly classify
182 targeted (e.g., ARGs) and non-targeted genes (e.g., non-ARGs) from new datasets to be
183 tested. We evaluated the impact of four identity thresholds of the negative datasets on
184 the learning features of the model. The results of five-fold cross-validation revealed that
185 the model with 0% identity as the threshold for recruiting non-target sequences had the
186 best performance metrics, including accuracy, recall, precision, and F1-score (Fig. 2a).

187 Under these optimized conditions, the positive SARD set contained 61874 ARG
188 sequences, including 2972 experimentally-confirmed core sequences inherited from the
189 CARD and 58902 homology-predicted (>80% identity and >80% coverage) expanded
190 ARG sequences from Uniref100. All ARG reference sequences were hierarchically
191 assigned to 19 classes and 2972 groups (Dataset S2 and Supplementary Figure 2).

192 To demonstrate the powerful utility of FunGeneTyper, we used the reference
193 protein sequences in SARD to train two transformer models (FunTrans and FunRep)
194 and developed them as a new deep-learning ARGTyper deep-learning classifier. We
195 used the trained ARGTyper to classify the testing set to validate the performance of the
196 ARGTyper. The overall ARGTyper performance metrics prove that FunGeneTyper
197 provides an excellent and robust framework for gene classification. Specifically, the
198 optimal FunTrans model at the ARG class level reached an accuracy of 0.9979, a
199 precision of 0.9830, a recall rate of 0.9683, and an F1 score of 0.9756 (Fig. 2b).
200 Moreover, the prediction precision and recall of all 17 ARG classes exceeded 0.96 (Fig.
201 2c), apart from fusidic acid and triclosan, which showed lower precision and recall
202 because they have only 21 and 53 reference sequences, respectively, in SARD (Dataset
203 S3). More training data helps the model learn more features. Nonetheless, the power of
204 FunTrans to classify these temporarily less-represented classes of ARGs will improve
205 as more functionally-verified sequences will be available for model training.

206 The vector space generated by FunGeneTyper was semantically rich and encoded
207 structural, evolutionary, and functional information. To explain what our model

208 intuitively learns, we obtained representations of all classes of ARG and non-ARG
209 sequences in the training set. We used uniform manifold approximation and projection
210 (UMAP) to reduce data dimensions in each layer to two. Visualizations performed in
211 the four essential representative layers (1st, 5th, 32nd, and 33rd) revealed the learning
212 process of the model (Fig. 2d). All ARG sequences were highly entangled at the first
213 level of encoding input. However, they became increasingly separated as the
214 transformer model got deeper. Each type of ARG undergoes a process from dispersion
215 to aggregation. This finding verified that FunTrans can efficiently learn the
216 representation features of sequences from raw input data with high entanglement.

217 Prediction multiclass confusion matrix was used to represent the effect of FunTrans
218 on the learning features of each ARG class. The results indicated that the FunTrans
219 model was excellent at predicting all ARG classes (Fig. 2e). We continued to locate
220 significant classification errors in the ARG classes using error detection counts (Fig.
221 2f). Prediction error was concentrated in the multidrug class. Specifically, 33 non-ARG
222 sequences were mispredicted as multidrug resistance, whereas 39 multidrug resistance
223 protein sequences were mispredicted as non-ARG sequences. The poor prediction
224 performance of these proteins is mainly due to their high structural differences and
225 diverse biological functions that include roles other than multidrug resistance[29], making
226 it challenging for a DL model to effectively learn sufficient discriminative features in
227 the absence of sufficient training data. Multidrug efflux pumps[29] export antibiotics and
228 other diverse extraneous substrates, including organic solvents, toxic heavy metals, and

229 antimicrobials, and also fulfill other key biological functions such as biofilm formation,
230 quorum sensing and survival and pathogenicity of bacteria[29]. Therefore, multidrug
231 resistance proteins or efflux pumps were not seriously considered as ARGs[17,30] and we
232 recommend excluding their sequences from ARG analysis unless they can be reliably
233 or unambiguously assigned to resistance functions of certain antibiotic classes.

234 Once the FunTrans model was shown to be robust and accurate in identifying ARGs
235 and classifying them into 19 classes, we trained FunRep, which conducted a more
236 detailed lower-level classification of ARGs into 2972 groups (Dataset S3). FunRep
237 achieved an overall prediction accuracy of 0.9023 for all ARG groups (Dataset S4). We
238 used UMAP to visualize FunRep model's learning process. UMAP was used to visualize
239 the characteristics of the final layer of all classes except the Fusidic acid class (21
240 sequences, Dataset S3). UMAP showed that FunRep can cluster the features of each
241 group in the main ARG classes, including beta-lactams (5909 sequences), Macrolides-
242 Lincosamides-Streptogramines (MLS, 2317 sequences), aminoglycosides (3483
243 sequences), and glycopeptides (2037 sequences) (Supplementary Figure 3).

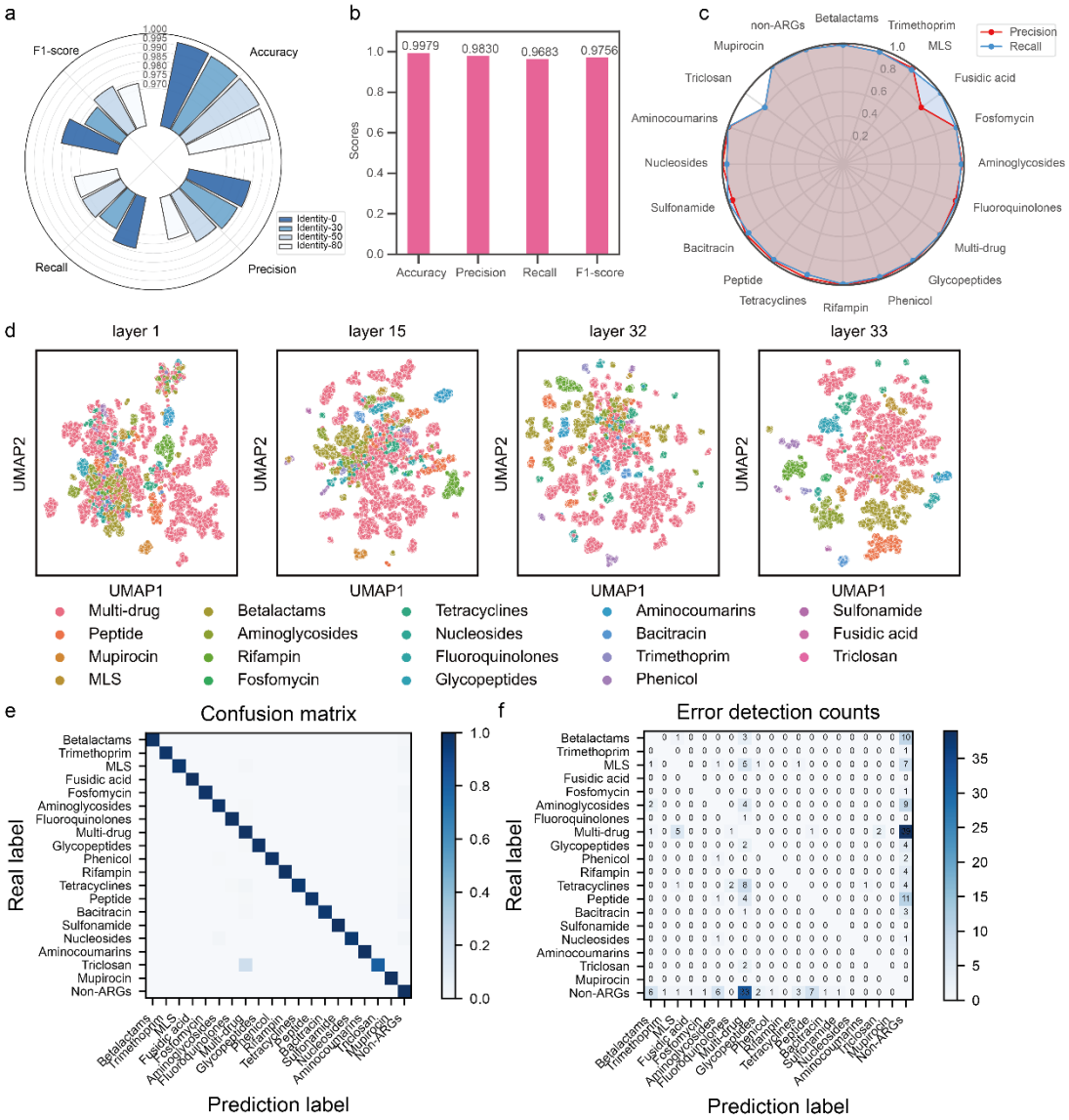


Fig. 2. Performance evaluation of deep-learning FunGeneTyper models with structured Antibiotic Resistance Gene Database (SARD) for classification of ARGs. a, Evaluation of the influence of identity threshold used for selecting the negative dataset on model performance in the classification of ARGs. **b**, Performance metrics of ARGTyper developed based on FunGeneTyper models and SARD. **c**, Performance of all 19 classes as indicated by precision and recall of ARGs and non-ARG classes. **d**, Visualization of feature learning at different layers during the ARGTyper training process. **e**, Confusion matrix for ARG class classification,

253 confusion between true (y-axis) and predicted (x-axis) ARGs. **f**, Number of ARG
254 protein sequences annotated incorrectly. MLS: Macrolides, Lincosamides and
255 Streptogramines.

256 In summary, we demonstrated the application of the FunGeneTyper framework to
257 develop ARGTyper as the first transformer-based ARG classifier trained from a
258 customized structured ARG database (SARD). The performance metrics of the testing
259 set show that FunTrans and FunRep can achieve highly accurate (accuracy=0.998) and
260 robust (F1-score=0.976) identification of all known types (classes) and subtypes
261 (groups) of ARGs in the authoritative CARD. Both the accuracy and robustness of
262 FunGeneTyper models outperform previously published results from DeepARG
263 (accuracy>0.97, F1-score>0.93 [9]) and HMD-ARG (accuracy=0.935, F1-score=0.893 [31])
264 on their own testing sets of ARGs.

265 **Model performance in the discovery of new genes**

266 The 'twilight zone' of protein sequence alignment is a complex, long-standing problem
267 plaguing protein function prediction[32,33], limiting the discovery of PCGs in the largely
268 uncultured microbial world. In contrast to classic SA-based tools, DL-based models
269 (FunRep and FunTrans) of the FunGeneTyper framework are designed with unique
270 features and intrinsic advantages for predicting remote homologues of protein
271 sequences with guaranteed accuracy and robustness, as previously demonstrated for
272 ARG classification.

273 To compare FunGeneTyper's ability to identify new PCGs with those of existing

274 methodologies, we evaluated its ability of its DL-based models to discover remote
275 homologues by predicting experimentally-confirmed protein sequences of new ARGs
276 discovered from three representative habitats: human gut (n = 168)[34], wastewater
277 treatment plants (n = 77)[35], and soil (n = 52)[36-39]. We computed the predictive
278 performance of FunGeneTyper classifier for ARGs (ARGTyper) and compared it with
279 that of three state-of-the-art tools: DL-based tools (HMD-ARG[31] and DeepARG[9]), SA-
280 based tools (RGI[7]), and HMM-based tools (Resfams[18]) (Table 1). FunGeneTyper had
281 higher accuracy, precision, recall, and F1-score for predicting new ARGs compared
282 with HMD-ARG[31] and DeepARG[9]. The significant improvement was primarily
283 attributed to our implementation of the protein semantic models (i.e., FunTrans and
284 FunRep) in FunGeneTyper, which can learn more hidden features of protein sequences,
285 especially the context information[19,21], compared with the traditional one-hot encoding
286 algorithm and the convolutional neural network used by HMD-ARG[31] and the
287 multilayer perceptron used by DeepARG[9]. Moreover, the overall classification
288 performance of FunGeneTyper, as benchmarked by the F1-score (0.5445 to 0.6948),
289 was much higher than that of the classic SA-based methods (0.0556 to 0.6598) and
290 HMM-based methods (0.2630 to 0.5224) (Table 1). Although RGI also achieved high
291 accuracy (0.8830) in human intestinal data, its precision (0.4545), recall (0.3968), and
292 F1-score (0.4195) were much lower than those of the FunTrans model (0.7500, 0.6642,
293 and 0.6948, respectively) because many of the new ARG sequences tested here fell
294 below the commonly applied stringent identity cutoffs (> 95% RGI). It is expected that

295 when a strict one-size-fits-all filter cutoff is applied to the alignment results, many false-
296 negatives would result, limiting the discovery of ARGs that show a more remote
297 homology to database sequences. The superior performance of the FunGeneTyper
298 classifier over existing tools in identifying new ARGs was further evident when
299 comparative tests were performed using wastewater treatment plant (WWTP) or soil
300 samples compared with human gut samples (Table 1). This indicates that
301 FunGeneTyper has a greater capacity to predict functional genes in complex
302 environmental samples. To further resolve the superior predictive performance of
303 FunGeneTyper for remote homologues of functional genes over existing tools, we
304 divided the ARGs data into lower homology (≤50% identity) and higher homology (≥
305 50% identity) datasets (Supplementary Figure 4). FunGeneTyper not only consistently
306 achieved better classification performance of higher homology ARGs in all three
307 sample groups (WWTP, soil, and human gut), it also showed outstanding performance
308 at accurately and sensitively predicting the function of remote homologous sequences
309 (Dataset S5).

310 Taken together, our results exemplify the discovery and classification of novel
311 ARGs, especially among relatively remote homologues (<50% identity), and
312 demonstrate that FunGeneTyper is best at predicting new ARG protein sequences,
313 exhibiting unprecedented capacity to identify new genes with high accuracy, sensitivity,
314 and robustness.

315 **Evaluating the generalizability of FunGeneTyper**

316 To demonstrate the generalizability of FunTrans and FunRep in classifying other gene
317 categories, we trained the models using a calibrated and professionally expanded
318 bacterial virulence factor database, VFNet[40] and utilized them to develop a new
319 transformer-based classifier of VFG, named VFGTyper. Before training the model, we
320 built a two-level expert-curated structured database based on the virulence ontology
321 and reference sequences in the VFNet database. Semantic and categorically ambiguous
322 data were cleaned (Methods). The final structured virulence factor database (SVFD)
323 consisted of 160484 VFG sequences distributed into 2837 classes in 45 families
324 (Dataset S6).

325 We followed a strategy similar to that mentioned above to train the model,
326 collecting a non-target dataset with 551,783 sequences (excluding VFGs) from Swiss-
327 Prot, as the negative dataset (see Methods). With the merit of the proposed adapter
328 module, we only need to re-train a new adapter when building a VFGTyper. The design
329 of adapter allows us to train only a new classifier and an adapter when predicting new
330 functions. All the parameters in the backbone network can be reused. Therefore,
331 VFGTyper can be regarded as a new task branch in the FunGeneTyper framework,
332 where only the adapter and classifier differ. We verified the VFGTyper using the testing
333 set to provide evidence of its generalizability in the functional genotyping process.
334 VFGTyper achieved an accuracy of 0.9907 (Fig. 3a) in the family level prediction task.
335 The obfuscation matrix results also showed that FunTrans achieved excellent

classification performance for each VFG at the family level (Fig. 3b, Supplementary Figure 5). In addition, FunRep was 0.9499 accurate at predicting different VFG classes in the second-stage prediction.

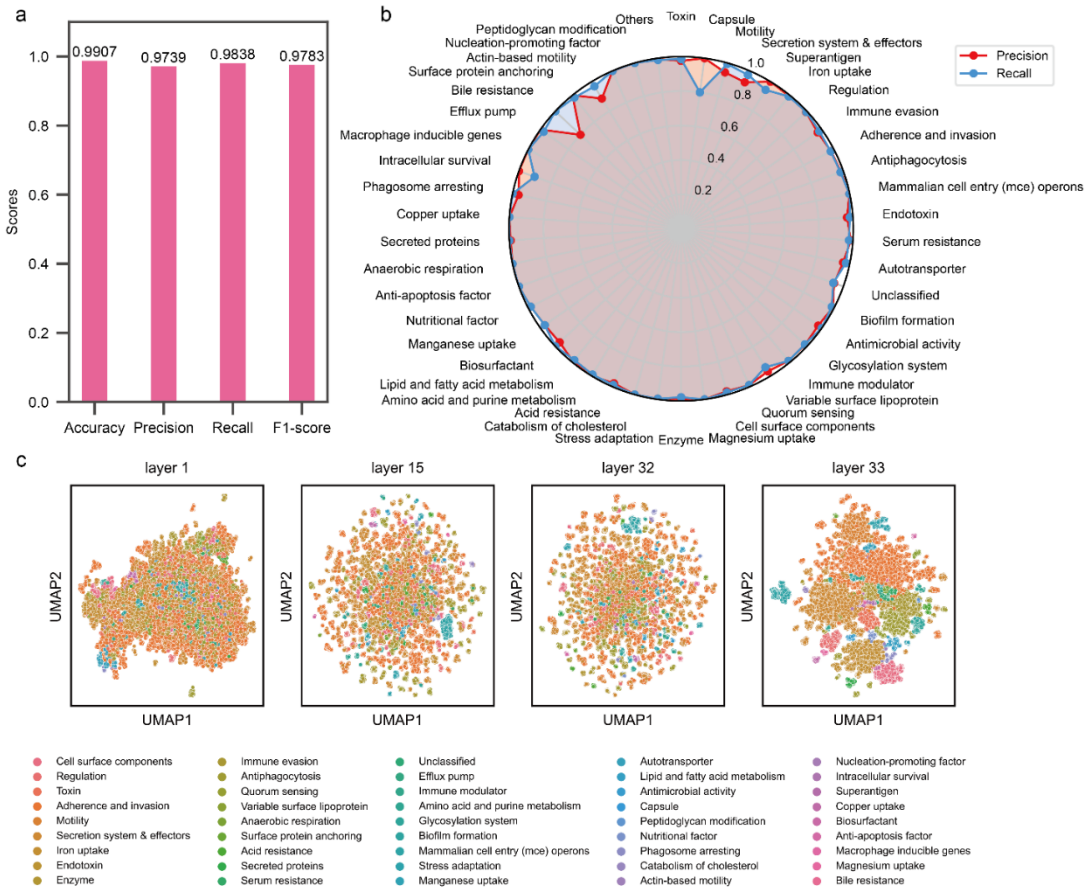


Fig. 3. Transfer learning of FunGeneTyper models on Structured Virulence Factor Gene Database (VFGD) and performance evaluation for VFG classification. a, Performance metrics of VFGTyper developed based on FunGeneTyper models and VFGD. **b**, Precision and Recall of VFGs family and non-VFGs category. **c**, Visualization of feature learning at different layers in VFGs FunTrans training. VFGs: virulence factor genes.

In conclusion, we demonstrated that FunGeneTyper can be successfully applied to VFGs by adding new adapters to increase the power of the VFGTyper. To determine

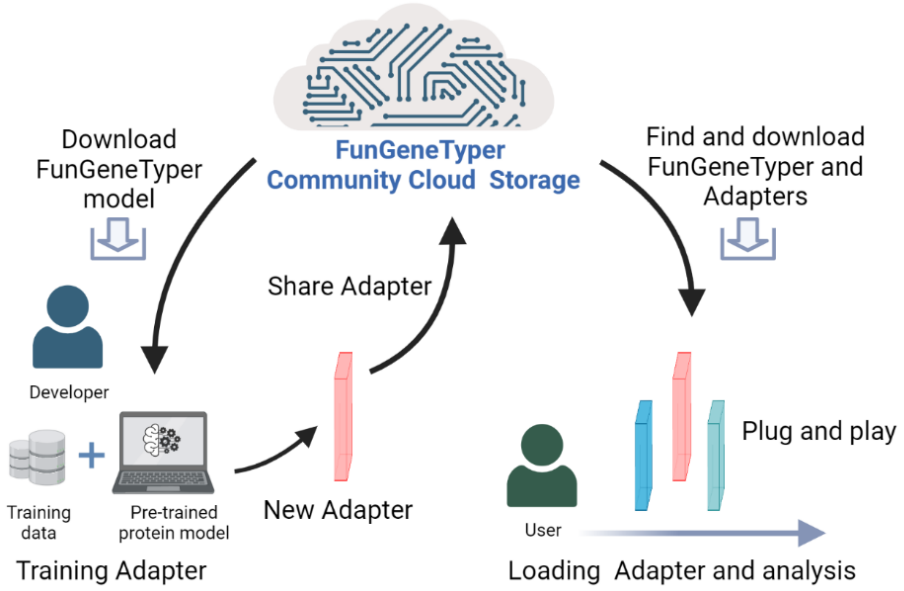
348 whether FunGeneTyper can learn valuable features from different datasets, we
349 visualized the learning process of FunTrans and FunRep models for VFG sequences at
350 the family (Fig. 3c) and class (Supplementary Figure 6) levels. Consistent with the
351 learning process for ARGs, FunTrans and FunRep conducted effective feature
352 clustering and classification of VFGs at family and class levels.

353 **Privacy-preserving global sharing of plug-and-play adapters for functional gene**
354 **discovery**

355 Benefiting from the parameter-efficient property, FunGeneTyper has two novel merits.
356 First, FunGeneTyper enables effective effort-sharing by the entire community.
357 Specifically, a researcher who has trained our FunGeneTyper model to discover a new
358 or unknown gene can submit their adapters (along with a classification layer) to the
359 adapter hub. Once the adapter has been submitted, the module can be downloaded and
360 inserted into the FunGeneTyper model.

361 Second, the adapter design helps solve data privacy issues. Where researchers have
362 not publicly released their own datasets, they can train FunGeneTyper with their private
363 datasets, submit only the adapter module (again along with a classification layer), and
364 provide functional descriptions of their FunGeneTyper. Thus, the private datasets are
365 protected, while the uploaded adapter models can be used without model training. As
366 the number of researchers getting involved in the development of FunGeneTyper
367 increases, the model may become a universal toolkit that can be used for predicting

368 functional genes simply by looking up related functional modules. We believe that with
369 the elegant adapter module, FunGeneTyper will facilitate adapter sharing and model
370 integration globally.



371

372 **Fig. 4. Schematic of the Adapter Sharing Community (ASC) in the framework of**
373 **FunGeneTyper.** The community developers are cyber de-centralized to train
374 customizable structured databases and develop deep learning classifiers of various
375 categories of functional genes, while users utilize the classifiers of interest to accelerate
376 the discovery of genes which, in turn, provide new experimentally-confirmed
377 sequences to expand the structured databases and improve deep-learning models.

378 **Discussion**

379 Metagenomics has provided an opportunity for identifying microbiome diversity and
380 novel functionalities. However, the speed at which high-throughput DNA sequencing

381 technologies unravel the vast genetic novelties of uncultured microbes in nature
382 outpaces our capacity to understand their function. Previous approaches for functional
383 classification of PCGs were based on sequence alignment using tools such as BLAST[4],
384 usearch[5], and Diamond[6] or conserved motifs and domains using Hidden Markov
385 Models. Selection of uniform cutoffs and thresholds usually limit the accuracy and/or
386 sensitivity of these methods for functional gene prediction. Protein semantic algorithms
387 based on NLP methods have been developed [20,24]. However, these algorithms are not
388 optimized for classifying different categories of microbial genes, and a unified thinking
389 paradigm is required to meet the needs for accelerated discovery of new genes.

390 Our study provides a scalable and expandable deep learning paradigm and
391 technical framework for efficient and robust gene function prediction to meet the
392 aforementioned urgent needs and tackle unprecedented challenges in the prediction of
393 PCGs. We propose an end-to-end FunGeneTyper framework for the classification
394 prediction of gene functions. We exemplify the framework by developing two
395 transformer-based classifiers, ARGTyper and VFGTyper, based on deep learning
396 models coupled with expert-curated structured databases (SARD and SVFD) to realize
397 robust functional classification of bacterial ARGs and VFGs, which are two categories
398 of genes key to WHO's one health approach for human, animal, and environmental
399 health protection[41].

400 A series of experimental validations, including five-fold cross-validation, testing
401 set validation, and experimentally-confirmed protein sequence validation, demonstrate

402 the effectiveness and robustness of FunGeneTyper. Using ARG as an example,
403 FunGeneTyper models are more effective than SA-based and DL-based models in
404 predicting protein sequences of new ARGs from the human gut, WWTP, and soil
405 microbiomes with relatively low homology (< 50% similarity) to known ARGs. This
406 shows that ARGTyper has an unmatched advantage in discovering ARGs, primarily
407 because of the powerful learning ability of protein semantic models. Since
408 experimentally-confirmed sequences of the major categories, types, and subtypes of
409 genes are not sufficient, expanding the database based on sequence homology is
410 common and necessary to obtain sufficient training sequence data. UMAP analysis
411 showed that the expanded sequences represent reliable datasets and support our model
412 to better learn discriminative protein semantic features to achieve satisfactory
413 performance in identifying functional genes, including ARGs and VFGs.

414 Accurately classifying target genes from the huge interference of non-target gene
415 data is a problem. Therefore, we purposefully introduced non-functional genetic
416 datasets as part of the training set. Although this operation increases training complexity,
417 it enables our model to accurately classify target genes from noisy data when used to
418 analyze large-scale metagenomic sequence datasets from environmental or animal
419 microbiome samples. Some machine learning methods rely on sequence alignment
420 tools to create a similarity score matrix of potential gene sequences and databases[9,42].
421 Such practices will inevitably be affected (and limited) by the selection of arbitrary
422 thresholds for the results. The FunGeneTyper framework proposed here can accurately

423 annotate genes via classification through discriminative features learned from multiple
424 sequences. The limited number of training sequences may prevent the models from
425 learning sufficient features. This transient issue would, however, be easily solved once
426 more experimentally-confirmed reference protein sequences of target genes are
427 available for model retraining and refinement. Meanwhile, the robustness of deep
428 learning to noise labels[43] can also help our framework models and classifiers
429 outperform existing ones in discovering new genes.

430 In particular, once large amounts of (meta)genomic data are freely available, a
431 uniform and convenient understanding of the relationship between microbial gene
432 sequences and protein function becomes a perennial challenge that can be tackled to
433 create opportunities for gene discovery. There are other gene categories, apart from
434 ARGs and VFGs, including those associated with microbially-driven global
435 biogeochemical cycling (carbon, nitrogen, phosphorus, and sulfur) or microbial
436 biodegradation (bioremediation and bio-restoration) and biosynthesis (biomedicine and
437 bioresources) (Fig. 5), such as those well established by the RDP's FunGene database[44].
438 Building a dynamic metagenomic bioinformatics community will help us better
439 understand gene function. In principle, FunGeneTyper can predict the function of any
440 gene category based on prior parameters of the pre-trained model and the adapter's
441 transfer-learning ability. The adapter module used in FunGeneTyper is a lightweight
442 plug-and-play neural network that only fine-tunes and maintains a small set of
443 parameters and is conducive for sharing and promotion. Therefore, other researchers

444 can use the framework and training parameters we provide to train their own core
445 datasets to easily develop predictive deep learning models of genes of interest.
446 Researchers can also share a trained adapter through the adapter sharing community
447 (ASC) without disclosing their private datasets. The future prosperity and collaboration
448 of the ASC under the guidance of FunGeneTyper framework provide an interactive,
449 dynamic, and continuously improving or evolving platform for functional classification
450 of various PCG sequences. More importantly, FunGeneTyper and ASC are expected to
451 contribute significantly to advances in industrial biotechnology, health and medicine,
452 food and agriculture, environmental biotechnology, and bioenergy (Fig. 5), as they are
453 increasingly applied to accelerate the discovery of new genes and enzymatic resources
454 from microbiomes.

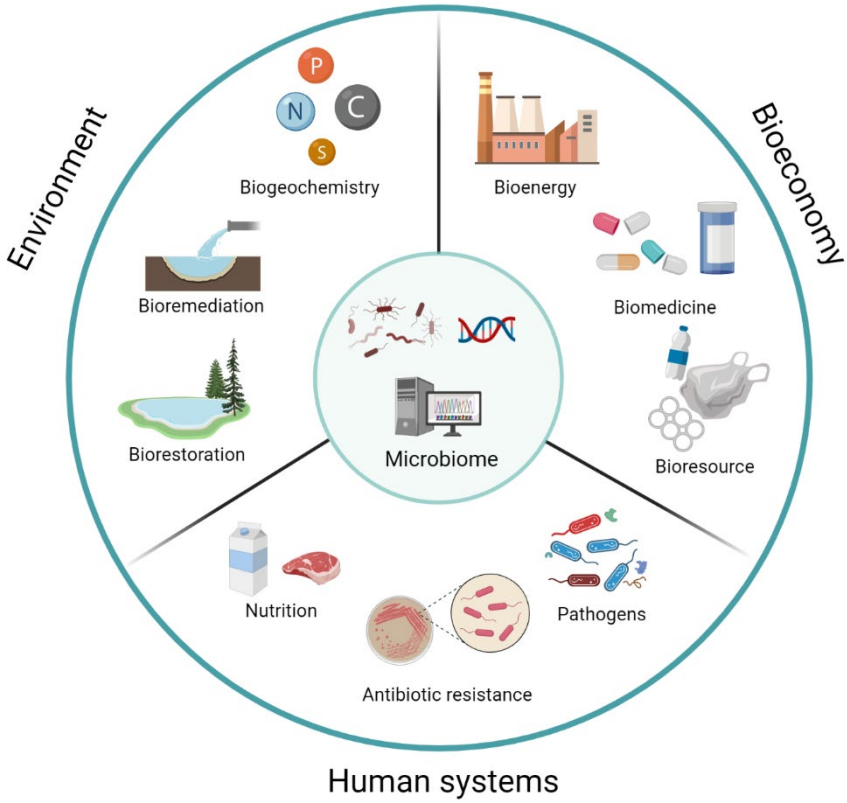


Fig. 5. Potential applications of FunGeneTyper to the discovery of microbiome resources for enhancing our environment, bioeconomy, and human systems. Metagenomic discovery of precious genetic and enzymatic resources facilitated by the Adapter Sharing Community of FunGeneTyper can contribute to follow-up microbiome, genetic and protein engineering researches for enhancing human health and eco-environment systems.

In conclusion, FunGeneTyper provides an innovative and unified framework with deep learning models (i.e., FunTrans and FunRep), expandable toolkits (e.g., ARGTyper and VFGTyper) and customizable structured databases for the classification and discovery of functional genes (e.g., ARGs and VFGs) that have scientific and

466 biotechnological significance. This framework will contribute to the continuous
467 discovery of various new genetic and enzymatic resources from diverse uncultured
468 microbes and microbiomes that are critical to the formation of the earth's biosphere and
469 the maintenance of our environment (biogeochemistry, bio-restoration, and
470 bioremediation)[14] bioeconomy (bioenergy and bioresources)[13], and human systems
471 (food and health)[20,45].

472 **Methods**

473 **Collection and expansion of the core dataset**

474 The core dataset used for FunGeneTyper model training is a set of experimentally-
475 confirmed reference sequences of target functional genes collected from literature
476 and/or expert-curated databases. Because the core dataset does not always contain a
477 sufficient number of experimentally-confirmed sequences (no more than 10
478 sequences[40]) for every type or subtype of functional gene, it is expanded to retrieve
479 more sequence data to improve and optimize the training of deep learning models. In
480 the subsequent training method, which separates the extended categories of five or more
481 sequences into the training set, verification set, and testing set at a ratio of 6:2:2, any
482 categories that are unsuitable for inclusion in these five sequences are included in the
483 training set.

484 **Construction of structured antibiotic resistance database (SARD)**

485 **Core ARGs dataset:** To ensure the professionalism and accuracy of the training dataset,
486 reference protein sequences of ARGs defined by homologs in the authoritative CARD
487 were selected as core data for downstream model training. The sequences were

488 clustered using CD-HIT[46] (*v*4.8.1) at an amino acid sequence identity of 100%, and all
489 protein sequences and their ontological information were manually checked to ensure
490 that each ARG was properly classified into class (type) and group (subtype) based on
491 their ontological information. Generally, class is equivalent to CARD's ontology terms
492 for antibiotic drug types, and group is equivalent to the specific sequence category.
493 Macrolides, lincosamides, and streptogramins were combined into the MLS class.
494 Based on the above procedures, a core dataset of 2972 non-redundant sequences
495 representing 2972 groups of ARGs from 19 classes was obtained and used to build the
496 SARD, which was used in subsequent analyses.

497 **Expanded ARGs dataset:** To ensure sufficient training data, the core dataset was
498 expanded by retrieving close homologues of its ARGs from the Uniref100 database
499 following strict screening criteria. Briefly, Diamond was used to index the ARG
500 sequences in the core dataset and to search for homologous sequences with an amino
501 acid identity and coverage greater than or equal to 80%. The extracted candidate
502 sequences were dereplicated and used as expanded datasets.

503 ***Negative dataset*** To ensure that the model can learn sufficient features of non-target
504 gene function, which is essential for robustly predicting target function directly from
505 metagenomic data, we used the Swiss-Prot database, an expert-validated protein
506 database, to generate a negative dataset for use as a non-ARG training set. First, protein
507 sequences associated with antibiotic resistance in the Swiss-Prot database were
508 screened out using the keywords KW-0046. The remaining sequences were aligned

against the core ARGs dataset using Diamond software. Sequences with an alignment coverage greater than 80% were extracted and categorized into four negative datasets based on their sequence alignment identity (ID): identity 0 (ID ≤0%), identity 30 (ID ≤30%), identity 50 (ID ≤50%), and identity 80 (ID ≤80%).

Construction of structured virulence factor database (SVFD)

Core VFGs dataset. Virulence factor databases were collected from VFNet[40]. Zheng et al[40] performed a detailed similarity search for known and potential VFGs in the complete bacterial genome downloaded from the NCBI server using VFanalyzer[47], with Virulence Factor Database (VFDB) as the core database[47]. VFNet is an expanded virulence factor database that can be used directly in the training process.

Negative dataset. The non-VFG collection process is similar to that of the non-ARG collection process, except that KW-0800 is used to filter sequences from Swiss-Prot database (version: June 2021).

Architecture of the FunGeneTyper model

FunGeneTyper is a universal function classification framework composed of two core deep learning models, FunTrans and FunRep, which share similar structures but are designed to classify functional genes at the type and subtype levels, respectively. Both models are modular adapter-based architectures that leverage a few extra parameters to achieve efficient fine-tuning of large-scale PLMs. In detail, utilizing the state-of-the-art large-scale protein PLM esm-1b as a 33-layer transformer encoder framework as the

foundation, we plug adapters in each transformer layer of the PLM, which are individual modular units that are used as newly introduced weights to be fine-tuned for specific functional tasks. Notably, ESM-1b, through self-supervised learning on the UniRef50 dataset, was shown to have a superior capacity to infer fundamental structural and functional characteristics of proteins from gene sequences[48].

The holistic architecture is depicted in Fig. 1a and consists of three main components: a multi-headed self-attention, a feed-forward network, and an adapter layer. Each sublayer contains layer normalization and skip connections to effectively train the neural network and avoid overfitting. It is worth noting that the bottleneck-shaped adapter module consists of a down-project linear $H \in \mathbb{R}^{d\times k} H \in \mathbb{R}^{d\times k}$ where d is embedding size of the Transformer model, k is the dimension of the adapter and $d \gg k$, a ReLU activation followed by an up-projection $L \in \mathbb{R}^{k\times d} L \in \mathbb{R}^{k\times d}$. The adapter layer is formulated as follows:

$$O_l = LayerNorm(T_l)$$
$$\text{Adapter}_l = L_l(\text{ReLU}(H_l(O_l)) + T_l$$

where T_l is the hidden feature at transformer layer l, $d = 1280$, and $k = 256$ in the actual training.

Following the approach of BERT[49], hidden features from the first token of the sequence of the last layer are extracted. In contrast to FunTrans, which adds a nonlinear layer for protein function classification after the representations of the last layer, FunRep first computes the hidden features of experimentally-confirmed core sequences

and then annotates PCGs by finding the sequence's category with the closest Euclidean distance in the representation space.

FunRep model training. Here, we use a dual-tower architecture with shared parameters similar to Sentence-BERT[50] for model training in order to place sequences with the same category closer in the representation space. FunRep is trained by constructing $< A, P, N >$ triples, where A is the anchor sequence, P is a positive example possessing the same category as A, and N is a negative example whose category is different from A and the hidden representations they obtained through FunRep are a, p, and n, respectively. The loss function adopts Triplet Loss, which is defined as follows:

$$Loss(a, p, n) = max(D(a, p) - D(a, n)) + margin, 0)$$

where D is the Euclidean distance between vectors, and $margin$ is an adjustable threshold, set to 1.0 during model training. ARGTyper-FunRep and VFGTyper-FunRep are classified at the group level with the same 21.76M learnable training parameters.

Training settings

All datasets are divided into training, validation and testing in a 6:2:2 ratio, and the five-fold cross-validation is performed. Adam optimizer with default parameters is used, dropout is set to 0.2, learning rate is 1e-5, and the early stopping method is adopted to prevent overfitting. The accuracy, precision, recall and F1-score are used to evaluate

569 the performance. As a result, the micro average of the F1-score also equals that of
570 precision and recall, as well as the overall accuracy. Thus, we report only the overall
571 accuracy for the micro average metrics while reporting precision, recall and F1-score
572 for the macro average metrics.

573 **Evaluation of FunGeneTyper for new functional gene discovery**

574 To validate the capacity of FunGeneTyper models in discovering new functional genes,
575 experimentally confirmed ARGs from functional metagenomics studies were retrieved
576 from NCBI's protein database (accession numbers in Dataset S7). After removing those
577 ARGs showing perfect sequence match to the CARD database (or core dataset of
578 ARGs), 297 experimentally confirmed ARG sequences of human gut[34] (n = 168),
579 WWTPs[35] (n = 77), and soil[36-39] (n = 52) bacteria were retained for use in the
580 downstream comparisons between FunGeneTyper and the well-established SA-based
581 (RGI[7]), HMM-based (Resfams[18]), and DL-based (DeepARG[9] and HMD-ARG[31])
582 approaches in terms of classification performance of the new ARGs. In this study, four
583 evaluation metrics including the accuracy, precision, recall and F1-score were
584 computed to assess the multi-classification results performance using the following
585 equations:

586 $$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

587 $$Precision = \frac{TP}{TP + FP}$$

588 $$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = \frac{2 * \ Precision * Recall}{Precision + Recall}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Abbreviations

SFGD: Structured Functional Gene Dataset

PCG: Protein-Coding Gene

ARG: Antibiotic Resistance Gene

VFG: Virulence Factor Gene

MLS: Macrolides, Lincosamides and Streptogramines

Author Contributions

F. Ju conceived the idea of FunGeneTyper, obtained funding, and supervised the project. G. Zhang and H. Wang performed the model construction, data analysis and visualization. F. Ju and F. Yuan co-supervised G. Zhang and H. Wang on the deep-learning model construction with additional support from G. Guo. G. Zhang constructed the structured databases and improved the data presentation with the help from Z. Zhang, L. Zhang and J. Yang. F. Ju and G. Zhang co-wrote the manuscript with assistance from F. Yuan and H. Wang. J. Yang, L. Zhang and Z. Zhang helped to revise the manuscript. All authors have given approval to the final version of the manuscript.

607 **Funding**

608 This work was supported by Zhejiang Provincial Natural Science Foundation of China
609 (grant no. LR22D010001), National Natural Science Foundation of China (22241603),
610 and Research Center for Industries of the Future (RCIF) at Westlake University (grant
611 no. WU2022C030).

612 **Acknowledgement**

613 The authors thank the Key Laboratory of Coastal Environment and Resources of
614 Zhejiang Province (KLaCER), Research Center for Industries of the Future (RCIF),
615 Westlake Center of Synthetic Biology and Integrated Bioengineering (WE-SynBio),
616 and the Westlake University HPC Center for computation support.

617 **Reference**

618 1 Ju, F. *et al.* Wastewater treatment plant resistomes are shaped by bacterial
619 composition, genetic exchange, and upregulated expression in the effluent
620 microbiomes. *ISME J* **13**, 346-360 (2019). https://doi.org:10.1038/s41396-018-
621 0277-8
622 2 Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed
623 by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and
624 Lifestyle. *Cell* **176**, 649-662 e620 (2019).
625 https://doi.org:10.1016/j.cell.2019.01.001
626 3 Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**,
627 499-509 (2021). https://doi.org:10.1038/s41587-020-0718-6
628 4 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
629 alignment search tool. *Journal of Molecular Biology* **215**, 403-410 (1990).
630 https://doi.org:10.1016/s0022-2836(05)80360-2
631 5 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST.
632 *Bioinformatics* **26**, 2460-2461 (2010).
633 6 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using

634 DIAMOND. *Nat. Methods* **12**, 59-60 (2015).
635 7 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the
636 comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517-
637 D525 (2020). https://doi.org:10.1093/nar/gkz935
638 8 Yang, Y. *et al.* ARGs-OAP: online analysis pipeline for antibiotic resistance
639 genes detection from metagenomic data using an integrated structured ARG-
640 database. *Bioinformatics*, btw136 (2016).
641 9 Arango-Argoty, G. *et al.* DeepARG: A deep learning approach for predicting
642 antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
643 10 de Nies, L. *et al.* PathoFact: a pipeline for the prediction of virulence factors
644 and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 49
645 (2021). https://doi.org:10.1186/s40168-020-00993-9
646 11 Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core
647 resistome. *Nat. Commun.* **6**, 8452 (2015).
648 12 Forsberg, K. J. *et al.* Bacterial phylogeny structures soil resistomes across
649 habitats. *Nature* **509**, 612-616 (2014). https://doi.org:10.1038/nature13377
650 13 Díaz Rodríguez, C. A. *et al.* Novel bacterial taxa in a minimal lignocellulolytic
651 consortium and their potential for lignin and plastics transformation. *ISME*
652 *Communications* **2** (2022). https://doi.org:10.1038/s43705-022-00176-7
653 14 Royo-Llonch, M. *et al.* Compendium of 530 metagenome-assembled bacterial
654 and archaeal genomes from the polar Arctic Ocean. *Nat Microbiol* **6**, 1561-1574
655 (2021). https://doi.org:10.1038/s41564-021-00979-9
656 15 Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene
657 Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations
658 in a Collection of Isolates. *Antimicrob Agents Chemother* **63** (2019).
659 https://doi.org:10.1128/AAC.00483-19
660 16 Xie, G. & Fair, J. M. Hidden Markov Model: a shortest unique representative
661 approach to detect the protein toxins, virulence factors and antibiotic resistance
662 genes. *BMC Res Notes* **14**, 122 (2021). https://doi.org:10.1186/s13104-021-
663 05531-w
664 17 Boolchandani, M., D'Souza, A. W. & Dantas, G. Sequencing-based methods and
665 resources to study antimicrobial resistance. *Nat Rev Genet* **20**, 356-370 (2019).
666 https://doi.org:10.1038/s41576-019-0108-4
667 18 Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic
668 resistance determinants reveals microbial resistomes cluster by ecology. *ISME*
669 *J* **9**, 207-216 (2015). https://doi.org:10.1038/ismej.2014.106
670 19 Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N
671 protein engineering with data-efficient deep learning. *Nat Methods* **18**, 389-396
672 (2021). https://doi.org:10.1038/s41592-021-01100-y
673 20 Ma, Y. *et al.* Identification of antimicrobial peptides from the human gut
674 microbiome using deep learning. *Nat Biotechnol* (2022).
675 https://doi.org:10.1038/s41587-022-01226-0

21 Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16**, 1315-1322 (2019). https://doi.org:10.1038/s41592-019-0598-1

22 Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org:10.1073/pnas.2016239118

23 Ofer, D., Brandes, N. & Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput Struct Biotechnol J* **19**, 1750-1758 (2021). https://doi.org:10.1016/j.csbj.2021.03.022

24 Unsal, S. *et al.* Learning functional properties of proteins with language models. *Nature Machine Intelligence* **4**, 227-245 (2022). https://doi.org:10.1038/s42256-022-00457-9

25 Bileschi, M. L. *et al.* Using deep learning to annotate the protein universe. *Nat Biotechnol* (2022). https://doi.org:10.1038/s41587-021-01179-w

26 Dohan, D., Gane, A., Bileschi, M. L., Belanger, D. & Colwell, L. in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 2782–2791 (Association for Computing Machinery, Virtual Event, Singapore, 2021).

27 Yuan, F., He, X., Karatzoglou, A. & Zhang, L. in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* 1469-1478 (2020).

28 Murray, C. J. L. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629-655 (2022). https://doi.org:10.1016/s0140-6736(21)02724-0

29 Du, D. *et al.* Multidrug efflux pumps: structure, function and regulation. *Nat Rev Microbiol* **16**, 523-539 (2018). https://doi.org:10.1038/s41579-018-0048-6

30 Piddock, L. J. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clin Microbiol Rev* **19**, 382-402 (2006). https://doi.org:10.1128/CMR.19.2.382-402.2006

31 Li, Y. *et al.* HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. *Microbiome* **9**, 40 (2021). https://doi.org:10.1186/s40168-021-01002-3

32 Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85-94 (1999). https://doi.org:10.1093/protein/12.2.85

33 Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst* **12**, 654-669 e653 (2021). https://doi.org:10.1016/j.cels.2021.05.017

34 Sommer, M. O. A., Dantas, G. & Church, G. M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128-1131 (2009). https://doi.org:10.1126/science.1176950

35 Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core

resistome. *Nat Commun* **6**, 8452 (2015). https://doi.org:10.1038/ncomms9452

36 Willms, I. M. *et al.* Novel Soil-Derived Beta-Lactam, Chloramphenicol, Fosfomycin and Trimethoprim Resistance Genes Revealed by Functional Metagenomics. *Antibiotics (Basel)* **10** (2021). https://doi.org:10.3390/antibiotics10040378

37 Wang, S. *et al.* Tetracycline Resistance Genes Identified from Distinct Soil Environments in China by Functional Metagenomics. *Front Microbiol* **8**, 1406 (2017). https://doi.org:10.3389/fmicb.2017.01406

38 Allen, H. K., Moe, L. A., Rodbumrer, J., Gaarder, A. & Handelsman, J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J* **3**, 243-251 (2009). https://doi.org:10.1038/ismej.2008.86

39 Donato, J. J. *et al.* Metagenomic analysis of apple orchard soil reveals antibiotic resistance genes encoding predicted bifunctional proteins. *Appl Environ Microbiol* **76**, 4396-4401 (2010). https://doi.org:10.1128/AEM.01763-09

40 Zheng, D., Pang, G., Liu, B., Chen, L. & Yang, J. Learning transferable deep convolutional neural networks for the classification of bacterial virulence factors. *Bioinformatics* **36**, 3693-3702 (2020). https://doi.org:10.1093/bioinformatics/btaa230

41 WHO, O. One health. *World Health Organization* (2017).

42 Wang, Z. *et al.* ARG-SHINE: improve antibiotic resistance class prediction by integrating sequence homology, functional information and deep convolutional neural network. *NAR Genom Bioinform* **3**, lqab066 (2021). https://doi.org:10.1093/nargab/lqab066

43 Chen, P., Ye, J., Chen, G., Zhao, J. & Heng, P.-A. in *Proceedings of the AAAI Conference on Artificial Intelligence.* 11451-11461.

44 Fish, J. A. *et al.* FunGene: the functional gene pipeline and repository. *Front Microbiol* **4**, 291 (2013). https://doi.org:10.3389/fmicb.2013.00291

45 Lee, E. D., Aurand, E. R., Friedman, D. C. & Engineering Biology Research Consortium Microbiomes Roadmapping Working, G. Engineering Microbiomes-Looking Ahead. *ACS Synth Biol* **9**, 3181-3183 (2020). https://doi.org:10.1021/acssynbio.0c00558

46 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006). https://doi.org:10.1093/bioinformatics/btl158

47 Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* **47**, D687-D692 (2019). https://doi.org:10.1093/nar/gky1080

48 Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.2012.2015.422761 (2020). https://doi.org:10.1101/2020.12.15.422761

49 Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 4171-4186 (2019).

https://doi.org:10.18653/v1/N19-1423

50 Reimers, N. & Gurevych, I. in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* 3982-3992.

767 Table 1 Comparative performance between FunGeneTyper and other alternative classification tools for the discovery of experimentally confirmed

768 new ARGs. In total, 297 experimentally confirmed ARGs sequences of human gut[34] (n = 168), WWTPs[35] (n = 77), and soil[36-39] (n = 52) bacteria

769 were included in the analysis which was performed based on default parameters of each tools recommended by the developers.

Tools	Human gut (n=168)				WWTP (n=77)				Soil (n=52)			
	Accuracy	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
DL-based tools												
FunGeneTyper	0.8512	**0.7500**	**0.6642**	**0.6948**	**0.7273**	**0.7500**	**0.5403**	**0.6072**	**0.8269**	0.5926	**0.5529**	**0.5445**
HMD-ARG	0.8452	0.6000	0.5230	0.5486	0.5714	0.7161	0.3877	0.4589	0.8077	**0.6000**	0.4560	0.5119
DeepARG	0.3512	0.6250	0.4720	0.5149	0.1688	0.5714	0.1682	0.2591	0.2885	0.3750	0.1057	0.1607
SA-based tools												
RGI	0.3452	0.6250	0.4596	0.5065	0.0390	0.3750	0.0349	0.0632	0.1538	0.1250	0.0357	0.0556
HMM-based tools												
Resfams	**0.8830**	0.4545	0.3968	0.4195	0.6234	0.6250	0.4736	0.5224	0.8088	0.2727	0.2545	0.2630

770

771

772