ACCOUNTS of chemical research

pubs.acs.org/accounts

Article

AI-Driven Antimicrobial Peptide Discovery: Mining and Generation

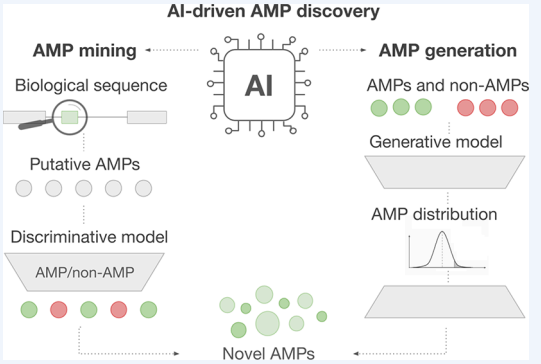
Paulina Szymczak, Wojciech Zarzecki, Jiejing Wang, Yiqian Duan, Jun Wang, Luis Pedro Coelho, Cesar de la Fuente-Nunez,* and Ewa Szczurek*

Cite This: https://doi.org/10.1021/acs.accounts.0c00594

Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: The escalating threat of antimicrobial resistance (AMR) poses a significant global health crisis, potentially surpassing cancer as a leading cause of death by 2050. Traditional antibiotic discovery methods have not kept pace with the rapidly evolving resistance mechanisms of pathogens, highlighting the urgent need for novel therapeutic strategies. In this context, antimicrobial peptides (AMPs) represent a promising class of therapeutics due to their selectivity toward bacteria and slower induction of resistance compared to classical, small molecule antibiotics. However, designing effective AMPs remains challenging because of the vast combinatorial sequence space and the need to balance efficacy with low toxicity. Addressing this issue is of paramount importance for chemists and researchers dedicated to developing next-generation antimicrobial agents.



Artificial intelligence (AI) presents a powerful tool to revolutionize AMP discovery. By leveraging AI, we can navigate the immense sequence space more efficiently, identifying peptides with optimal therapeutic properties. This Account explores the emerging application of AI in AMP discovery, focusing on two primary strategies: AMP mining, and AMP generation, as well as the use of discriminative methods as a valuable toolbox.

AMP mining involves scanning biological sequences to identify potential AMPs. Discriminative models are then used to predict the activity and toxicity of these peptides. This approach has successfully identified numerous promising candidates, which were subsequently validated experimentally, demonstrating the potential of AI in AMP design and discovery.

AMP generation, on the other hand, creates novel peptide sequences by learning from existing data through generative modeling. This class of models optimizes for desired properties, such as increased activity and reduced toxicity, potentially producing synthetic peptides that surpass naturally occurring ones. Despite the risk of generating unrealistic sequences, generative models hold the promise of accelerating the discovery of highly effective and highly novel and diverse AMPs.

In this Account, we describe the technical challenges and advancements in these AI-based approaches. We discuss the importance of integrating various data sources and the role of advanced algorithms in refining peptide predictions. Additionally, we highlight the future potential of AI to not only expedite the discovery process but also to uncover peptides with unprecedented properties, paving the way for next-generation antimicrobial therapies.

In conclusion, the synergy between AI and AMP discovery opens new frontiers in the fight against AMR. By harnessing the power of AI, we can design novel peptides that are both highly effective and safe, offering hope for a future where AMR is no longer a looming threat. Our paper underscores the transformative potential of AI in drug discovery, advocating for its continued integration into biomedical research.

KEY REFERENCES

- Szymczak, P.; Możejko, M.; Grzegorzek, T. et al. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.* **2023**, *14*, 1453.[1] *Extended conditional variational autoencoder, trained for unconstrained and analogue generation of novel active AMPs, with parametrized creativity.*
- Wan, F.; Torres, M. D. T.; Peng, J.; de la Fuente-Nunez, C. Deep-learning-enabled antibiotic discovery through molecular de-extinction. *Nat. Biomed. Eng.* **2024**, *8*, 854−871.[2] *This paper introduced APEX, a deep learning model used to mine all extinct organisms known to science as a source of antibiotic molecules, leading to the identification of preclinical candidates such as mammuthusin and elephasin.*
- Santos-Júnior, C. D.; Torres, M. D. T.; Duan, Y.; Rodríguez Del Río, Á.; Schmidt, T. S. B.; Chong, H.;

Received: November 7, 2024
Revised: April 25, 2025
Accepted: April 28, 2025

Fullam, A.; Kuhn, M.; Zhu, C.; Houseman, A.; Somborski, J.; Vines, A.; Zhao, X. M.; Bork, P.; Huerta-Cepas, J.; de la Fuente-Nunez, C.; Coelho, L. P. Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell* **2024**, *187*, P3761-3778.E16.[3] *Nearly one million candidate peptide antibiotics identified in the global microbiome with discriminative machine learning.*

- Torres, M. D. T.; Melo, M. C. R.; Flowers, L.; Crescenzi, O.; Notomista, E.; de la Fuente-Nunez, C. Mining for encrypted peptide antibiotics in the human proteome. *Nat. Biomed. Eng.* **2022**, *6*, 67−75.[4] *First exploration of the human proteome as a source of antibiotics, revealing thousands of previously unrecognized peptides, many of which may contribute to host immunity and other physiological processes.*
- Ma, Y.; Guo, Z.; Xia, B.; Zhang, Y.; Liu, X.; Yu, Y.; Tang, N.; Tong, X.; Wang, M.; Ye, X.; Feng, J.; Chen, Y.; Wang, J. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **2022**, *40*, 921−931.[5] *Expanding the deep learning methods for AMP discovery to anticancer peptides.*

INTRODUCTION

Following the boom of antibiotic discoveries in the 20th century, the past three decades have experienced a discovery void, with no novel antibiotic classes reaching the market.[6] Concurrently, resistance to existing antibiotics has escalated.[7] Antimicrobial resistance has become a significant global health and economic issue, projected to surpass cancer as a leading cause of death by 2050.[8] The antibiotic discovery process remains cumbersome, often requiring many years to identify preclinical candidates, resulting in the lack of antibiotic innovation.[9−11]

A promising strategy to combat antimicrobial resistance involves designing novel antimicrobial peptides (AMPs). These peptides are typically short (10−100 amino acids), with a net positive charge (commonly +2 to +9) and a high proportion (≥30%) of hydrophobic amino acids. Positively charged AMPs show distinct selectivity for negatively charged microbial membranes and tend not to target the neutral membranes of eukaryotic cells. Other than membrane targeting, modes of AMP action include inhibition of key processes such as protein or nucleic acid synthesis, protease activity or cell division. Microbial targets and modes of action of AMPs are determined by their amino acid composition and structure, with distinguished AMP classes including proline-rich, tryptophan- and arginine, histidine or glycine-rich peptides, majority of which are alpha helical but may also adopt beta-sheet, linear extension or mixed alpha-beta conformations. AMPs play important roles in host response to pathogen infection in a wide range of organisms, and are found in mammals, amphibia, insects, and microorganisms.[12] Importantly, microbes develop resistance to AMPs more slowly than to traditional antibiotics.[13] However, many AMPs exhibit toxicity to mammalian cells, often assessed through cytotoxic and hemolytic activities.[14] While multiple characteristics such as solubility and stability are important for AMPs,[15] the primary challenge remains in designing peptides that are highly active and minimally toxic. Given the limited success of known peptides in the clinic, innovative methods for designing AMP are essential to enhance their properties beyond those of existing peptides.[15,16]

The design of AMPs could be defined as an optimization problem, namely as searching for the most active and least toxic peptides within a vast space of potential sequences, with only scarce data on their properties. Considering sequences of up to 25 amino acids in length, a brute-force search algorithm would need to evaluate on the order of 10^{32} sequences, a task that is computationally infeasible.[17] In contrast, the number of experimentally verified AMPs (on the order of 10^4, based on the DBAASP database[18]) is minuscule. Peptides with documented activity against specific bacterial species like *Escherichia coli* (*E. coli*) are even scarcer (around 10^3).

To navigate this vast search space effectively, sophisticated algorithms are necessary. These algorithms must strike a balance between selecting *realistic* peptide sequences, i.e. those that resemble existing AMPs, and designing *idealistic* peptides, i.e., optimizing peptides for heightened activity and low toxicity, resulting in a *realism-idealism trade-off*. Advances in AI-driven AMP design revealed two primary strategies: (i) biological sequence mining, and (ii) generative AI. AMP mining identifies peptides by exploring genomes and proteomes and evaluating potential candidates using discriminative models that predict their activity or toxicity. This approach has successfully yielded peptides that are likely to be naturally produced, aligning with the realism aspect of the design. Conversely, generative AI models learn the distribution of peptide data and generate novel sequences, often optimizing them with predictive models to enhance activity and reduce toxicity. This strategy is capable of creating idealistic, synthetic peptides, potentially exceeding those found in nature, though it risks producing sequences that may not be sufficiently realistic. Both genome mining and generative AI approaches have the potential to dramatically accelerate antibiotic discovery, enabling the identification of hundreds of thousands of potential candidate molecules. Indeed, the AI-based algorithms developed so far have successfully designed and discovered peptides, some with proven efficacy in preclinical mouse models.[2−5,19,20]

Here, we describe recent advancements and the current state of the art in discriminative methods for assessing activity and toxicity, which are often instrumental for AI-driven AMP design. This is followed by a systematic review on the emerging areas of AMP mining and generative AI-based strategies for AMP discovery. Finally, we discuss remaining challenges in AMP design and outline promising research directions.

DISCRIMINATIVE METHODS

Discriminative methods serve for both AMP mining and AMP generation, and are crucial for the selection of promising active and nontoxic candidates. The majority of models broadly distinguish AMPs from non-AMPs (e.g., sAMP-pred-GAT,[21] AMPlify,[22] and AMPpredMFA[23]). More elaborate approaches focus on identifying highly potent peptides either via classification or regression by incorporating information on MIC measurements into the model.[24,25] Strain- or species-specific discriminators attempt to select peptides with an activity profile specific to a given microbe, such as AMP-META,[26] or MBC-attention.[24] While much less popular due to data scarcity, approaches for AMP toxicity exist as well, such as EnDL-HemoLyt,[27] AMP-META,[26] Macrel[28] and others.[29−32] Strikingly, only few of the discriminative models are evaluated experimentally through microbiological assays, and even less

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Table 1. Discriminative Methods for AMP Discovery[a]

method	framework	feature type	task	experimental validation	approach type
sAMPpred-GAT[21]	GNN, ATT; MLP	sequence-derived descriptors, structure	AMP		ML-based
AMPlify[22]	LSTM, ATT; MLP	sequence	AMP	microbiological assays	
AMPpredMFA[23]	LSTM, CNN, ATT; MLP	sequence	AMP		
MBC-attention[24]	CNN, ATT; MLP	sequence derived	activity		
AMP-META[26]	LGBM	structure, sequence-derived descriptors	AMP, activity, toxicity	microbiological assays	
EnDL-HemoLyt[27]	LSMT, CNN; MLP	sequence	toxicity		
Macrel[28]	RF	sequence-derived descriptors	AMP, toxicity		
Pandi et al[24]	CNN, RNN; MLP	sequence	activity	microbiological assays, hemolysis assays, cytotoxicity assays	
APEX[2]	RNN, ATT; MLP	sequence	activity	microbiological assays, in vivo animal models, cytotoxicity assays	
Capecchi et al.[29]	RNN, GRU, SVM; MLP	sequence	activity, toxicity	microbiological assays, hemolysis assays	
Ansari and White[32]	RNN, LSTM	sequence	toxicity, solubility		
ESKAPEE-MICpred[31]	LSTM, CNN; MLP	sequence, sequence-derived descriptors	activity	microbiological assays	
Ansari and White[30]	LSTM; MLP	sequence	toxicity, non-fouling activity, SHP-2		
Zhuang and Shengxin[38]	QSVM	sequence-derived descriptors	toxicity		
AmPEPpy[34]	RF	sequence	AMP		
Orsi and Reymond[46]	GPT-3; MLP	sequence	toxicity, solubility		LLM-based
iAMP-Attenpred[40]	BERT; MLP	pLM embedding	AMP		
PepHarmony[41]	ESM, GearNet; MLP	sequence, structure	solubility, affinity, self-contaction		
SenseXAMP[42]	ESM-1b; MLP	pLM embedding	activity		
HDM-AMP[43]	ESM-1b; DF	pLM embedding	activity	microbiological assays	
AMPFinder[51]	ProtTrans, OntoProtein; MLP	pLM embedding	activity		
LMPred[52]	ProtTrans; MLP	pLM embedding	activity		
PHAT[49]	ProtTrans; MLP	pLM embedding	secondary structure		
PeptideBERT[47]	BERT (ProtBert); MLP	pLM embedding	toxicity, solubility, non-fouling activity		
TransImbAMP[53]	BERT; MLP	pLM embedding	activity		
AMPDeep[45]	BERT (ProtBert); MLP	pLM embedding	toxicity		
Zhang et al.[48]	BERT; MLP	pLM embedding	activity		
Ma, Yue, et al[5]	BERT, ATT, LSTM; MLP	sequence	AMP	microbiological assays, in vivo animal models, hemolysis assays, cytotoxicity assays	
iAMP-CA2L[39]	CNN, Bi-LSTM, MLP; SVM	structure	AMP		structure-based
sAMP-VGG16[55]	CNN; MLP	sequence-derived descriptors	AMP		
AMPredictor[56]	ESM; MLP	sequence-derived descriptors, structure	activity	microbiological assays, in vivo animal models, hemolysis assays	

[a]GNN: Graph Neural Network; ATT: attention mechanism, MLP: Multi-layer perceptron, LSTM: Long Short-Term Memory, CNN: Convolutional Neural Network, LGBM: Light Gradient-Boosting Machine, RF: Random Forest, RNN: Recurrent Neural Network, GRU: Gated Recurrent Unit, SVM: Supporting Vector Machine, QSVM: Quantum Supporting Vector Machine, GPT-3: Generative Pre-trained Transformer 3, BERT: Bidirectional Encoder Representations from Transformers, ESM: Evolutionary Scale Modeling, DF: Deep Forest, Bi-LSTM: Bi-directional Long Short-Term Memory.

frequently through hemolytic activity and cytotoxicity assays. An overview of recent discriminative methods summarizing employed frameworks, feature types, performed tasks, and experimental validation is provided in Table 1.

Models and Architectures Applied for the Development of Discriminative Methods

Traditional ML methods, such as decision trees, Support Vector Machines (SVMs), and random forest (RF), rely entirely on sequence-derived descriptors for AMP prediction.[33,34] Importantly, because of their simplicity, these

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

methods can be employed to infer biological insights, such as analysis of Shapley Additive exPlanations to reveal differences in AMP mechanisms of action between Gram-negative and Gram-positive bacteria.[26] Another example of a traditional ML discriminator for AMPs is Macrel,[28] a random forest model, trained on an unbalanced data set containing a ratio of approximately 1:50 of AMPs vs non-AMPs to more closely mimic the expected distribution in a genome mining task than a balanced data set. Macrel was successfully applied in a recent large AMP mining study, AMPSphere.[3] The relative simplicity and on-par or, for some tasks, better performance than more sophisticated, deep-learning (DL) based approaches make traditional ML methods a recommendable choice for AMP identification.[35]

Still, compared to traditional machine learning methods, DL models have the potential to increase effectiveness in addressing more complex challenges and improve prediction accuracy for AMP discrimination.[36,37] The most ubiquitous DL methods in AMP prediction are derived from models originally devised for natural languages, such as recurrent neural networks (RNNs),[19,25,29] or long short-term memory (LSTM) architectures.[25,30−32] Furthermore, attention mechanisms have emerged as a pivotal component of many recent architectures. Through examination of features related to sequence composition, such as frequency of occurrence of each amino acid, and backward and forward relationships, these models gain a profound comprehension of the "semantics" inherent in biological sequences. Models such as AMPlify[22] achieve that by incorporating autoregressive models like Bi-LSTM and attention layers. Although convolutional neural networks (CNNs) were originally introduced for vision-related tasks, they are also used for AMP prediction, based on sequence-derived features. One example of a CNN-based model is MBC-Attention,[24] which combines multibranch CNN with attention mechanism to regress the minimum inhibitory concentration of AMPs against *E. coli*. An approach combining the strengths of both autoregressive and CNN-based methods is AMPpred-MFA,[23] which uses both bidirectional-LSTMs and CNNs, followed by multihead attention mechanism to extract context dependencies of peptide sequences. Finally, a quantum supporting vector machine was proposed by Zhuang and Shengxin to detect toxicity of peptides based on sequence derived descriptors.[38]

Large Language Models Applied in Discriminative Methods

While deep learning networks such as RNNs, LSTMs or CNNs do take into account the context relationships within amino acids, the recent large language models (LLMs) based on the transformer architecture offer novel opportunities for analysis of large corpuses of sequence data, in particular by efficiently leveraging the attention mechanism. In particular, LLMs have been successfully applied to protein sequences, resulting in so-called protein language models (PLMs).[39] The process of training such models usually follows two steps. First, a transformer model is pretrained on a large corpus of proteins in a generative task, and then fine-tuned to a specific downstream task, such as function, property, or structure prediction. Similarly to ML-based approached, PLMs have been applied to predict antimicrobial activity[5,40−45] and nontoxicity,[46−48] but other properties such as solubility or secondary structure[41,46,47,49] have been tackled as well (Table 1).

Compared to typical proteins, peptides are shorter in length and have relatively less complex tertiary structure. Moreover, the number of known bioactive peptides is much smaller than the number of known proteins, and the number of AMPs validated by experimental methods is limited. Therefore, direct application of PLMs without additional fine-tuning to AMP data would result in models biased toward more protein-like properties. Indeed, models trained on proteins, 'chopped proteins' (short subsequences of proteins) and peptides result in different embeddings of the input sequences and the models trained on shorter sequences have more generalized embeddings and perform better in downstream tasks.[50] While attempts to directly use text-based pretrained LLMs without additional pretraining on protein corpus have been made,[40,46] this approach was shown to result in inferior performance to RNNs[46] as the embeddings trained on natural text are likely not suitable for the domain of peptide sequences.

The most prevalently used LLM architectures are bidirectional encoder representation from transformers (BERT), which are effective in dealing with long-distance dependencies and thus learn the global context information on input sequences. Apart from such BERT-based models, other encoder-only architectures are successfully employed for AMP classification, in particular Evolutionary Scale Modeling (ESM) encoders, built upon the concept of integrating both sequence and evolutionary information.[41−43] OntoProtein, a BERT-like model based on both protein sequences and the gene ontology (GO), was incorporated into AMPFinder[51] to predict the functional types of AMP. However, in a recent evaluation by Dee,[52] full encoder-decoder transformer architectures[49,51,52] were proven to outperform the encoder-only models, confirming the results of Elnaggar et al., who performed similar benchmarking for proteins.[44]

Apart from architecture, the PLM models in AMP prediction differ also by their pretraining corpus, with most methods using UniRef50,[41−43,49,51,52] fewer using UniRef100,[47,51] and individual cases pretraining on Pfam,[53] BFD,[45,52] and UniProt,[48] or merging corpuses.[51] The selection of the pretraining corpus has a significant influence on model performance, as more diverse corpuses, such as UniRef50 having lower between-sequence similarities than UniRef100, were shown to improve results without any changes to the architecture.[52]

While most models directly proceed with training by adding prediction heads to the pretrained models and fine-tuning for discriminative AMP tasks, some approaches incorporate an additional phase of fine-tuning beforehand, for example using secretory data as an additional corpus[45] for toxicity prediction or data for sequences shorter than 50 amino acids.[41] Such additional pretraining phases may shift the pretrained model's focus toward distributions of sequences that are more peptide- or AMP-like. Indeed, as peptides are much shorter and with simpler structure than proteins, LLMs pretrained on proteins may not adequately represent the peptide distribution.[50]

Representations of Peptides Used by Discriminative Methods

Various discriminative models differ by the representation of peptides as their input features. The most prevalent representation is the amino acid sequence. While it can serve as a primary input to the model, it is also used to obtain sequence-derived descriptors or embeddings from pretrained models. Feature encoding using PLMs outperformed human-

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Table 2. Mining approaches for AMP discovery

approach	tool applied	model	biological sequence type	biological sequence source	experimental validation	mining criteria
Torres et al.[4]	Pane *et*[77]	fitness function	proteome	*Homo sapiens*	microbiological assays, *in vivo* animal models	activity
Cesaro et al.[66]	Pane et al.[77]	fitness function	proteome	*Homo sapiens*	microbiological assays, *in vivo* animal models, cytotoxicity assays	activity
APEX[2]	APEX[2]	RNN, ATT; MLP	proteome	Extinct species	microbiological assays, *in vivo* animal models, cytotoxicity assays	activity
Torres et al.[68]	Pane et al.[77]	fitness function	proteome	*Homo sapiens*	microbiological assays, *in vivo* animal models, cytotoxicity assays	activity
panCleave[19]	panCleave[19]	random forest	proteome	extinct species	microbiological assays, *in vivo* animal models	AMP
Monsalve et al.[67]	PepMultiFinder, CAMPR3, AMP Scanner and AmpClass 1.0		proteome	eight species of bacteria, plants, a protist, and a nematode	microbiological assays, cytotoxicity assays, hemolysis assays	AMP
Klimovich and Bosch[73]	HMMER, BLAST		genome, proteome, transcriptome	*Hydra* microbiome		AMP orthologs
Ma et al.[5]	Ma et al.[5]	Ensemble of LSTM, attention and BERT	genome, proteome	*Homo sapiens* microbiome	microbiological assays, *in vivo* animal models, cytotoxicity assays, hemolysis assays	AMP
Ma et al.[72]	Ma et al.[5]	Ensemble of LSTM, attention and BERT	genome	*Homo sapiens* microbiome	microbiological assays, *in vivo* animal models, cytotoxicity assays, anticancer activity assays	AMP, ACP
Torres et al.[34]	SmORFinder, AmPEPpy		genome	*Homo sapiens* microbiome	microbiological assays, *in vivo* animal models, cytotoxicity assays	AMP/activity
Santos-Júnior et al.[28]	Macrel[28]	random forest	genome, proteome, transcriptome	global microbiome		AMP/activity, nontoxicity
GMSC-mapper[60]	GMSC-mapper[60]		genome	global microbiome		smORFs
Wu et al.[71]	Ma et al.[5]	ensemble of LSTM, attention and BERT	genome	ESKAPE + ESKAPE phages		AMP/activity
SMEP[75]	SMEP[75]	XGBoost, LSTM	hexapeptide, heptapeptide, octapeptide and nonapeptide libraries		microbiological assays, *in vivo* animal models, cytotoxicity assays, hemolysis assays	activity
FSLSMEP[78]	FSLSMEP[78]	ESM-1, MLP	hexapeptides, heptapeptides and octapeptides libraries		microbiological assays, *in vivo* animal models, cytotoxicity assays, hemolysis assays	activity

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

engineered features in a benchmarking study of García-Jacas et al.[54] Still, Zhang et al. improved the performance of their model SenseXAMP[42] in AMP prediction by fusing the embedding of a pretrained protein model with traditional protein descriptors (PD). SenseXAMP performed better than simply fine-tuning pretrained models, indicating that traditional PDs continue to play a crucial role in AMP screening tasks. Other approaches convert a sequence to an image, either using cellular automata[40] or atom connectivity information,[55] and then apply CNNs as the architecture for discriminative models.

Apart from methods that focus on amino acid sequence as the primary peptide representation, some methods try to incorporate structural information as an additional, complementary view. Of particular interest are methods that leverage graph-based approaches to encode structural information about peptides. For example, sAMP-pred-GAT[21] integrates structural, sequence, and evolutionary information on peptides to construct a graph attention network (GAT) used to identify AMPs. Similarly, AMPredictor[56] is a Graph Convolutional Net that incorporates Morgan fingerprints, peptide contact maps, and embedding from ESM to regress MIC values. Graph encoding was also combined with pLM embedding in PepHarmony,[41] which merges sequence-level encoding from ESM with structure-level embedding from GearNet in multiview contrastive learning.

AMP MINING

The availability of biological sequence data has seen an unprecedented expansion in recent years sparking efforts to discover new AMPs using mining strategies. AMP mining involves applying the previously discussed discriminative methods to biological sequence data, including genomes, proteomes, and metagenomes. Historically, AMPs were oftentimes found in the skin secretions from amphibians.[57] While AMP mining requires careful handling to minimize false positives, particularly given the very large size of input data sets, it can yield high-quality predictions that have been substantiated through both *in vitro* and *in vivo* validations.[3,5,58] The AMP mining approaches primarily target antimicrobial properties and they differ by the type and source of analyzed biological sequence collections (Table 2). Majority of AMP mining methods do not rely on toxicity predictors, likely due to their low reliability.

Biological Sequence Collections Amenable for AMP Mining

Today, millions of genomes are accessible, and, together with metagenomes (which contain the genomic material of multiple organisms in a microbial community) or proteomes, they are collected in public databases.[59−63] An example of rich data is the nonredundant Global Microbial Gene Catalogue (GMGCv1), created from thousands of metagenomes across numerous habitats. This resource includes billions of open reading frames (ORFs) clustered at a high nucleotide identity level, resulting in hundreds of millions of species-level unigenes.[64] The GMGCv1 also contains tens of thousands of AMR genes, identified through homology-based searches against the Comprehensive Antibiotic Resistance Database (CARD)[65] and alignment with known resistance gene sequences. In another study, Duan and colleagues constructed a global microbial catalog of small open reading frames (smORFs), which encode small proteins. The catalog, named GMSC, was derived from thousands of publicly available metagenomes across multiple distinct habitats and thousands of high-quality isolate genomes. GMSC contains close to a million nonredundant smORFs with comprehensive annotations and provides a tool called GMSC-mapper to identify and annotate small proteins from microbial (meta)genomes.[60]

AMP Mining of Genomes and Proteomes

Recently, the human proteome was explored as a source of antibiotics.[4,66−68] The landmark study of Torres et al. employed an algorithm that utilized key physicochemical properties such as sequence length, net charge, and average hydrophobicity to predict antimicrobial activity. Building on the work of Pane et al.,[77] this algorithm models antimicrobial potency as being linearly dependent on physicochemical properties raised to exponents - model parameters that were fitted using known AMPs. Specifically, the algorithm scanned 42,361 protein sequences from the human proteome and identified 2,603 potential AMP candidates, many of which were previously unrecognized as antimicrobials or to play a role in host immunity. By avoiding known AMP motifs and focusing on physicochemical characteristics, these explorations led to the discovery of novel antimicrobials, several of which were synthesized, validated experimentally, and showed efficacy in animal models.

AI has also enabled biological mining efforts[69] to explore proteins from extinct species, such as Neanderthals and Denisovans, revealing a new set of antimicrobial sequences and launching the field of molecular de-extinction.[19] In this work, the authors introduced panCleave - a random forest model for proteome-wide cleavage site prediction. For selection of candidate AMPs, apart from expert curation, they used a consensus of six publicly available traditional ML-based AMP models, including Macrel.[28] Another study mined the proteomes of all available extinct organisms, including the woolly mammoth. Using a more powerful deep learning model called APEX,[2] this study led to the discovery of novel AMPs, such as neanderthalin-1, mammuthusin-2, and elephasin-2, which now represent preclinical candidates. These computational efforts have drastically accelerated our ability to discover new antibiotics, transforming a process that once took years into one that can be completed in hours.[70]

An alternative approach for mining for phage peptidoglycan hydrolases (PGHs)−derived antimicrobial peptides was proposed by Wu et al.[71] The study introduced a computational pipeline to mine AMPs derived from ESKAPE microbes (a group of clinically dangerous pathogens comprising *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa* and *Enterobacter* spp) and their associated phages. To evaluate the antibacterial activity of the extracted peptides, the authors trained a model with CNNs and LSTM layers, basing on the model used by Ma et al.[5] The result is a database, ESKtides, containing over 12 million peptides with predicted high antibacterial activity.

AMP Mining of the Microbiome

Using the human gut microbiome as the biological sequence resource, Ma et al.[5] mined for AMPs using deep learning techniques, including LSTM, attention, and BERT. The study identified 181 peptides showing antimicrobial activity, many of which had less than 40% sequence homology to known AMPs, demonstrated significant efficacy against antibiotic-resistant,

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Table 3. Generation Approaches to AMP Discovery

method	generation mode	controlled generation	aimed properties	generation framework	experimental validation	MD
AMP-GAN[93]	unconstrained	conditional generation	sequence length, microbial target, target mechanism, activity	cGAN	microbiological assays, cytotoxicity assays	yes
MMCD[102]	unconstrained	conditional generation, contrastive learning	AMP, ACP	diffusion		
CLaSS[80]	unconstrained	discriminator-guided filtering	AMP, activity, nontoxicity, structure	WAE	microbiological assays, *in vivo* animal models, cytotoxicity assays, hemolysis assays	yes
LSSAMP[83]	unconstrained	latent space sampling	secondary structure	vector quantized VAE	microbiological assays, *in vivo* animal models, cytotoxicity assays, hemolysis assays	
AMP-Diffusion[101]	unconstrained	positive-only learning	AMP	PLM + diffusion	microbiological assays, *in vivo* animal models, cytotoxicity assays	
AMPGAN v2[94]	unconstrained	conditional generation	sequence length, microbial target, target mechanism, activity	cGAN		
AMPTrans-LSTM[82]	unconstrained	discriminator-guided filtering	AMP	LSTM + transformer		
Zeng et al.[99]	unconstrained	discriminator-guided filtering	AMP	PLM	microbiological assays	
Jain et al.[96]	unconstrained	active learning	AMP	GFlowNets + active learning		
Pandi et al.[24]	unconstrained	discriminator-guided filtering	activity	VAE	microbiological assays, cytotoxicity assays, hemolysis assays	yes
M3-CAD[85]	unconstrained	conditional generation, discriminator-guided filtering	microbial target, nontoxicity, mode of action	cVAE	microbiological assays and *in vivo*, cytotoxicity assays, hemolysis assays	
Ghorbani et al.[88]	unconstrained		AMP	VAE		
MODAN[97]	optimized	bayesian optimization	Activity and nontoxicity	Gaussian process	microbiological assays, hemolysis assays	
Cao et al.[92]	unconstrained	discriminator-guided filtering	AMP	GAN	microbiological assays	yes
Diff-AMP[100]	unconstrained	discriminator-guided filtering	AMP	Diffusion		
HydrAMP[1]	unconstrained, analogue	conditional generation	AMP, activity	cVAE	microbiological assays, hemolysis assays	yes
AMPEMO[98]	optimized	discriminator-guided filtering	AMP, diversity	Genetic algorithm		
Buehler et al.[103]	unconstrained	conditional generation	secondary structure, solubility	GNN		
Renaud and Mansbach[84]	unconstrained, analogue	latent space sampling	AMP, hydrophobicity	VAE		
Capecchi et al.[29]	unconstrained	discriminator-guided filtering, positive-only learning	activity, nontoxicity	RNN	microbiological assays, hemolysis assays	
Multi-CGAN[90]	unconstrained	conditional generation	activity, nontoxicity, structure	cGAN		
QMO[89]	optimized	zeroth-order optimization, gradient descent	activity, nontoxicity	WAE		
PandoraGAN[91]	unconstrained	positive-only learning	antiviral activity	GAN		
PepVAE[86]	unconstrained	latent space sampling	activity	VAE	microbiological assays	
ProT-Diff[104]	unconstrained	discriminator-guided filtering, positive-only learning	AMP, activity	PLM + diffusion	microbiological assays and *in vivo*, cytotoxicity assays, hemolysis assays	
MOQA[87]	optimized	D-wave quantum annealer	activity, nontoxicity	binary VAE	microbiological assays, hemolysis assays	

Gram-negative bacteria, and in reducing bacterial load in a mouse model of lung infection.

In another study, deep learning discriminator methods were applied for the task of anticancer peptide (ACP) prediction, leveraging the overlap between ACPs and AMPs.[72] The study employed a high-throughput mining process to identify 40 potential ACPs from the gut microbiome metagenomic data. Out of these, 39 peptides showed significant anticancer activity in various cancer cell lines. Two peptides, in particular, demonstrated exceptional efficacy in reducing tumor size in a mouse model without causing toxicity.

Furthermore, nearly a million new potential AMPs were recently discovered by computational analysis of the global microbiome.[3] Using machine learning, the authors explored the vast diversity of the microbial world, analyzing 63,410 metagenomes and 87,920 microbial genomes. Additionally, identification of the peptides in proteomics and transcriptomics data was used as a filtering step after identification based on genomic sequence, The study resulted in the computational prediction of nearly a million candidates for new AMP, which were deposited in the AMPSphere database.

Another study integrated metagenomes from four different body sites to identify smORF-encoded peptides.[58] To evaluate smORFs that were likely to encode AMPs, the authors used a random-forest based discriminator model. The study identified 323 candidate antibiotic peptides, showing activity against clinically relevant pathogens, both *in vitro* and *in vivo*.

Finally, several recent studies turned to mining microbiomes other than human gut, focusing on particularly promising species that are known to effectively maintain microbiome homeostasis.[73,74] For example, Klimovich et al. performed high-throughput transcriptome and genome sequencing, followed by machine learning-based analysis of freshwater polyp *Hydra*'s microbiome. The study revealed that AMP-

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

encoding genes underwent a recent, rapid evolution in the Hydra species, that AMPs are selectively expressed in certain cell types, and finally, that the AMP activity follows a spatial pattern, suggesting that depending on the microhabitat different cocktails of AMPs targeting different bacterial species are secreted to generate a specific chemical landscape to locally control the density and shape the composition of the microbiome of *Hydra*.[73] Another study, based on a deep learning model with Dense-Net blocks and a self-attention module,[74] focused on the gut microbiome of cockroaches, which harbors harmful species without occurring pathogenesis.

Exhaustive Mining of Combinatorial AMP Sequence Spaces for Short Peptides

Instead of mining natural biological sequence resources, recent efforts focus on evaluating all possible sequences of up to a fixed, short length. Huang et al. developed a machine-learning-based pipeline to systematically identify effective AMPs from a vast virtual library of peptides made of 6−9 amino acids.[75] Their pipeline consists of multiple sequential machine-learning modules designed to filter, classify, rank, and predict the efficacy of potential AMPs. Since their discriminator models were trained on the GRAMPA data set,[76] a compiled collection of MIC measurements that could suffer from lab-specific biases, the authors adopted a two-step experimental validation strategy, refining their discriminator after the initial stage to mitigate possible biases in the training data. The results of the study include the identification of three potent hexapeptides that exhibit strong antimicrobial activities against multidrug-resistant pathogens, comparable efficacy to penicillin in treating bacterial infections in mice, and low toxicity. Another study focused on developing AMPs against *Acinetobacter baumannii*, scanning through the entire libraries of hexapeptides, heptapeptides, and octapeptides, encompassing tens of billions of candidates. Their pipeline included classifiers for *A. baumannii*-specific AMPs, trained on an extremely scarce training data set of only 148 sequences using a few-shot learning strategy involving pretraining and multiple fine-tuning steps.

AMP GENERATION

Generative AI holds the promise to transform the discovery of novel drug candidates.[79] By learning and modeling underlying data distributions, generative AI may become indispensable tools for peptide generation in the future. Several generative AI methods have already been applied for this purpose, showing promising results and paving the way for future advancements in peptide-based antimicrobial drug discovery.[1,15,80,81]

Modeling Frameworks Employed in AMP Generation

The generative AI approaches applied so far differ by the specific modeling framework that they use (Table 3). While the use of autoregressive models, such as LSTMs and more generally RNNs,[29,82] has been explored, they are currently less frequently used compared to other methods. The majority of research thus far has focused on the implementation of variational autoencoders (VAEs)[1,25,83−88] or Wasserstein autoencoders.[80,89] Generative adversarial networks (GANs)[90−94] have also seen significant application, demonstrating their ability to generate new AMP sequences. Most AMP generation methods focus on generating candidates with promising antimicrobial candidates, with a smaller number addressing hemolytic or cytotoxic properties as well. While some generative AI studies included microbiological testing of generated AMP candidates, relatively few were further tested in animal models (Table 3). The different model architectures and their training processes have been discussed in detail in our recent review.[95]

Controlled AMP Generation

Generative AI methods can efficiently produce thousands of plausible peptide candidates, making it crucial to direct the generation process toward acquiring desired properties to increase the likelihood of identifying relevant hits. One fundamental approach to controlled generation is the use of auxiliary discriminators to guide the generative process and to filter top candidates. In the CLaSS model,[80] discriminative models are trained on the latent space of a WAE, guiding the generation toward peptides with targeted activity and toxicity. Another straightforward strategy is based on positive-only learning, as demonstrated by PandoraGAN,[91] where only highly active peptides are used for training. However, both discriminator-guided filtering and positive-only learning are limited by the sparse availability of positively labeled training data, namely active and nontoxic AMPs.

The advancement of GAN or VAE-based models has led to the development of their conditional variants, such as cGANs[90,93,94] and cVAEs.[1,85] These models are configured during the generation phase to produce peptides more likely to meet specific criteria. For instance, Multi-CGAN[90] optimizes the generation process to address multiple properties simultaneously, while M3-CAD,[85] a multimodal, multitask, and multilabel cVAE, targets eight feature categories including predicted 3D structure, species-specific antimicrobial activities, antimicrobial mechanisms, and toxicity.

Additionally, some methods exploit the model's latent space to guide generation. Techniques such as latent space sampling allow for the selection of peptides from regions expected to encode desirable attributes.[83,84,86] For example LSSAMP[83] discretizes the latent representation to encode both sequence and structural information, facilitating the generation of peptides with desired secondary structures.

To tackle challenges like training data deficiency, the model of Szymczak et al., called HydrAMP, introduced several key enhancements to the standard cVAE framework.[1] The model focused on generating highly active antimicrobial peptides (AMPs) by conditioning on properties like low MIC values. HydrAMP included a pretrained classifier to ensure that the generated peptides retained the desired properties. To improve training stability, the authors added terms to the loss function that made sure the generated peptides closely matched the input and that the latent representations of the input and the generated peptides matched as well. HydrAMP also featured the ability to modify an existing peptide to meet specific activity conditions, controlled by a creativity parameter. Higher creativity led to more diverse analogues. Unlike standard cVAEs, which generate peptides by sampling from the latent space, HydrAMP could improve both known AMPs and peptides experimentally proven to lack antimicrobial activity. Molecular dynamics (MD) simulations provided additional descriptors of peptide activity, which, combined with a classifier ensemble, helped rank candidates for experimental testing. Finally, the most promising peptides were synthesized, and their activity and toxicity were experimentally validated. Using HydrAMP, Szymczak et al. discovered 15 novel, highly potent AMPs, that were active against several strains of bacteria, including multi drug resistant strains.

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Analog generation, as employed by HydrAMP, is one way of lead optimization for AMPs, as it produces peptides similar to a nonactive prototype but with enhanced properties. A promising research direction for idealistic peptide design is direct optimized generation, often using tailored cost functions. In this domain, QMO[89] uses zeroth-order gradient optimization to navigate the latent space. Other approaches addressing the optimization challenge include active learning with GFlowNets,[96] quantum annealing,[87] Bayesian optimization,[97] and evolutionary algorithms.[98]

Large Language Models Applied in AMP Generation

With the current success of tools such as chatGPT, generative language modeling becomes popular also in the field of AMP generation. While pretrained LMs are widely used in discriminative tasks, their application to AMP generation has been limited. AMP generation from pLMs typically involves either decoder-like architectures such as GPT[99] or a diffusion process trained on continuous embeddings obtained from pretrained LMs.[100,101] Unfortunately, these methods have so far implemented relatively simplistic strategies for controlled design, relying mainly on either positive-only learning or discriminator-guided filtering.[99−101] A promising direction has been the use of contrastive learning, as in MMCD[102] where training of a diffusion-based model involves contrasting the embeddings of known positive AMP examples with negative ones.

CHALLENGES AND PROSPECTS FOR THE FUTURE

With the AI revolution transforming the world today, there is growing potential for AI to enhance the design of novel AMP-based antibiotics and to fight antimicrobial resistance. Indeed, many breakthroughs have been made in recent years.[1,2,4,5,15,19,20,74]

The two described approaches: AMP mining, and AMP generation, equipped and enhanced by the use of cutting edge discriminative methods, are among the most promising strategies for AI-driven AMP discovery. However, limitations of existing tools in this domain need to be addressed, and open avenues of research remain unexplored.

Challenges to Be Addressed in the Realm of Discriminative Models

Despite intensive development of discriminative models, challenges remain regarding their application in the AMP field. First, the development of predictive methods is hindered by the relatively small volume of available data. Recent approaches based on transfer learning, with particularly wide usage of pretrained and fine-tuned LLMs, promise to partially solve the problem of data scarcity. More developments for computational handling of low-data regimes, as well as initiatives enhancing data sharing and experimental validation efforts are needed as the data for multiresistant strains will remain insufficient for training strain-specific activity predictors. Finally, more emphasis should be placed on developing methods predicting peptides' toxicity, a major barrier to the clinical application of AMPs. Currently, due to limited training data, existing methods for hemolytic activity prediction perform suboptimally, while cytotoxicity prediction methods are severely lacking.

Another major challenge is the lack of experimentally validated negative examples, which is easily explained by the lack of incentive to generate negative data.[105] Moreover, a peptide may falsely appear negative if not tested against a sensitive strain or due to technical issues such as clumping in solution which can easily be mistaken for lack of activity. The lack of well-defined negatives poses a particularly difficult problem in the supervised learning paradigm. As shown by Sidorczuk et al.[106] negative data set construction heavily influences the performance of models in the task of AMP prediction. Some approaches try to address the issue by modifying the loss function, for example by using asymmetric loss that down-weights the negative samples,[53] or by adapting the data sampling procedure, for example down-sampling the negative examples.[30,48] Additionally, different studies determine both positives and negatives under different experimental conditions, such as medium and bacterial concentrations, further adding confusion to the definition of labels for model training.[15] Therefore, databases of AMPs should be populated by researchers not only with the experimentally validated positives, but also negatives, and standards for the experimental conditions should be unified.

In addition, full use of peptide structural information could be more effective for functional prediction. However, due to scarce structural information, in-depth analysis of the importance of structure formation for AMP property prediction is currently limited. Moreover, the structures available in the databases such as DBAASP are obtained without taking into consideration neither the proximity of the cell membrane nor the effect of self-association of the peptides into larger aggregates. Additional structural information such as secondary and tertiary structures and post-translational modifications could be considered in future studies, providing an opportunity to improve the performance of AMP predictions.

Moreover, it is crucial to verify the robustness and generalizability of the model on different independent data sets. However, existing deep learning models have not yet been objectively evaluated using external data sets, which should be addressed.[35]

While quite successful at classification of linear peptide data, current discriminative methods are not applicable to a wide realm of modified peptides. A significant subset of antimicrobial peptides contains noncanonical building blocks such as cycles, β-amino acids, modified cysteines, and lipid attachments, distinguishing them from purely linear peptides, however those types of peptides are not abundant enough for AI classifier training[107] Existing FDA-accepted AMPs, such as polymyxins, significantly differ from such linear peptides.[108] This limits the applicability of discriminative methods for detection of clinically relevant AMPs. Therefore, more data on complex and modified peptides should be deposited in dedicated databases to enable AI model training.

Furthermore, data on other important properties of peptides could be used in the future to train better classifiers. For example, peptides can be degraded *in vivo*, resulting in half-lives that are insufficient for translational development. Currently, relevant training data on half-lives is not available and would greatly benefit AI-driven design. Such data could be collected by conducting systematic experimental studies that measure peptide stability in various biological environments, and compiling the results into publicly accessible databases for machine learning applications. To our knowledge, there is also no available data about ADMET properties of AMPs, which could be used to train AI models. To address this challenge, a large experimental study testing ADMET properties of peptides would be required. Alternatively, as performed by

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Mishra and Muthukaliannan et al.,[109] machine learning approaches for predicting ADMET properties of small molecules could be applied to SMILES representations of peptides. However, such predictions may be unreliable due to the larger size and different physicochemical properties of peptides compared to small molecules.

Challenges to Be Addressed in AMP Mining

AMP mining in biological sequences brings the advantages of identifying highly realistic peptides as AMP candidates. Indeed, the biological peptide sequences contain only L-amino acids and thus can be easily synthesized by solid phase synthesis with low costs and limited byproduct reactions. AMPs secreted in living organisms, such as in the gut microbiome, may specifically target invading microbes while avoiding toxicities to hosts. However, the potential of biological sequence mining strategies for discovering AMP candidates is realized only under specific conditions: first, AMPs with high activity and low toxicity must be present within the biological sequences being mined; second, the discriminative methods used must be capable of accurately identifying these AMPs. While the existing mining approaches have already proven successful,[2,3,5,19,74] these conditions underline the necessity of having both a rich biological data set and robust discriminative models. Furthermore, the genomic context has long been recognized as a critical factor in functional predictions of biological sequence properties.[110] Recent advancements have seen techniques from natural language processing being innovatively applied to gene function prediction, providing new ways to interpret complex biological data.[111−113] Adapting these approaches in the context of AMP mining could allow for better use of the limited labeled data for AMPs, but testing their accuracy empirically remains an open problem.

AMP mining could also benefit from expanding the analytical framework beyond genomics and proteomics alone by integrating additional data types into the predictive models. These could include transcriptomics data, which has been shown to be predictive of protein function[114] or ribosomal sequencing data.[115] However, there is still a shortage of data sets with these data types compared to the widespread availability of genomes.

While some peptides are produced by a complex process of post-translational modifications or even nonribosomally, many peptides are encoded by single genes, which can be directly transcribed and translated, or derived from the cleavage of a single precursor protein. This simpler genetic architecture makes peptides particularly amenable to exploitation for biotechnological purposes. Exploiting the potential of biosynthetic gene clusters, which must be expressed as a group to produce the desired compound[116,117] could prove beneficial for AMP discovery.

Moreover, processing each sequence independently with a pretrained model, although practical, overlooks valuable information contained in natural variations. Where multiple sequences are available,[118] employing models that analyze multisequence alignments rather than individual sequences can offer significant benefits.

Finally, since AMP mining relies on discriminative methods, their limitations transfer naturally to this approach. In particular, as discriminative methods cannot detect complex, chemically modified peptides, currently only linear peptides can be mined. Although recent AMP mining studies did perform preclinical testing of the most promising AMP candidates in animal models, none of those candidates went further to clinical studies yet.

Challenges to Be Addressed in AMP Generation

Similarly to AMP mining, generative AI has the potential to accelerate AMP discovery, but several obstacles remain. First, evaluation and benchmarking of generative models prove difficult. Generated peptides are most often evaluated with respect to diversity, novelty, and similarity to training data, but the activity and toxicity remain unknown except for a small experimentally validated subset. Auxiliary discriminators are used to estimate those properties of interest, but the choice of such models is entirely arbitrary, making comparison between models impossible. Second, as generative methods are capable of generating thousands of candidates within a short span of time, efficient methods to rank top candidates are needed. Currently, top candidates are selected using extensive filtering and expert knowledge. The existing generative approaches are also limited in their performance due to low data availability. Additionally, searching for potent peptides can be thought of as generating examples out-of-distribution, a widely recognized problem in generative modeling.

With few exceptions,[97] generative AI methods operate only on the 20-letter amino acid alphabet, without taking into account post-translational modifications or nonstandard amino acids. Thus, they are unable to sample from the huge space of peptides with chemical modifications, thereby largely underestimating the full complexity of the peptide world. Further extensions of the generative models to account for nonstandard amino acids may result in highly potent AMP designs in the future. So far, there is not enough training data to equip generative AI with the abilities to directly design such complex peptides as those currently in use in the clinic. However, the promising linear peptide candidates can be further enhanced by rational design, increasing their stability, efficacy and safety by choosing from the repertoire of typical chemical modifications, such as cyclization, residue phosphorylation or addition of lipids.[12]

In comparison to AMP mining, relatively fewer AMPs obtained from generative AI methods were confirmed preclinically *in vivo*, which may stem from the fact that the majority of AI laboratories have limited experimental capacities. Therefore, AI-discovered AMPs are yet to be tested in clinical trials. This calls for collaborative efforts of AI, chemical and biological laboratories joining forces with industrial partners to cover the steps from discovery to the clinics.

Finally, the emerging generative AI methods are usually benchmarked and developed for generation of text or images and are not always well-suited for the generation of peptides. Importantly, most existing generative AI models need specific modeling extensions to achieve controlled generation, which poses an important research direction that is largely unexplored for the design of AMPs.

SUMMARY

The future of antimicrobial peptide discovery is on the cusp of a transformative revolution with the integration of AI technologies. Since early pioneering work demonstrated that machines could design peptide antibiotics effective in preclinical mouse models,[20] this field has grown and matured significantly. AI-driven approaches have already dramatically accelerated our ability to identify new AMPs. By leveraging

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

large-scale genomic and proteomic data, coupled with sophisticated generative and discriminative models, AI will facilitate the design of potent AMPs specifically tailored to combat emerging resistant pathogens. This synergy between AI and biotechnology promises to accelerate the drug discovery process while also overcoming limitations associated with traditional approaches.

AUTHOR INFORMATION

Corresponding Authors

Cesar de la Fuente-Nunez − *Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, and Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States of America; Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States of America;* orcid.org/0000-0002-2005-5629; Email: cfuente@upenn.edu

Ewa Szczurek − *Institute of AI for Health, Helmholtz Zentrum Munich, Neuherberg 85764, Germany; Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw 02-097, Poland;* orcid.org/0000-0002-1320-6695; Email: em.szczurek@uw.edu.pl

Authors

Paulina Szymczak − *Institute of AI for Health, Helmholtz Zentrum Munich, Neuherberg 85764, Germany;* orcid.org/0000-0001-8419-4006

Wojciech Zarzecki − *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw 02-097, Poland; Faculty of Electronics and Information Technology, Warsaw University of Technology, Warsaw 00-661, Poland*

Jiejing Wang − *Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China*

Yiqian Duan − *Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China*

Jun Wang − *Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China*

Luis Pedro Coelho − *Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland 4102, Australia; Centre for Data Science, Queensland University of Technology, Brisbane, Queensland 4001, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00594

Author Contributions

CRediT: **Paulina Szymczak** conceptualization, data curation, visualization, writing - original draft; **Wojciech Zarzecki** data curation, writing - original draft; **Jiejing Wang** writing - original draft; **Yiqian Duan** writing - original draft; **Jun Wang** funding acquisition, writing - original draft; **Luis Pedro Coelho** funding acquisition, writing - review & editing; **Cesar de la Fuente-Nunez** funding acquisition, supervision, writing - original draft; **Ewa Szczurek** conceptualization, funding acquisition, supervision, writing - original draft.

Notes

The authors declare the following competing financial interest(s): Szczurek lab receives funding from Merck Healthcare. Cesar de la Fuente-Nunez is a co-founder of, and scientific advisor, to Peptaris, Inc., provides consulting services to Invaio Sciences, and is a member of the Scientific Advisory Boards of Nowture S.L., Peptidus, European Biotech Venture Builder, the Peptide Drug Hunting Consortium (PDHC), ePhective Therapeutics, Inc., and Phare Bio. The de la Fuente Lab has received research funding or in-kind donations from United Therapeutics, Strata Manufacturing PJSC, and Procter & Gamble, none of which were used in support of this work. The other authors declare no competing interests.

Biographies

Paulina Szymczak is a PhD student at the Institute of AI for Health at Helmholtz Munich, Germany. She obtained multidisciplinary training in both molecular biology (University of Wrocław) and Bioinformatics (University of Warsaw). Since 2020 she has actively worked on antimicrobial peptide discovery, including both discriminative methods and generative AI.

Wojciech Zarzecki is a computer science student at the Warsaw University of Technology. He is also a member of the Computational Medicine Group led by Prof. Ewa Szczurek. His main interests are the application of deep learning in antimicrobial peptide discovery and computer vision.

Jiejing Wang is a graduate student in the research group of Prof. Jun Wang at the Institute of Microbiology, Chinese Academy of Sciences (IM-CAS) in Beijing, China. She graduated from Beijing Forestry University and is currently in her second year of a master's degree program, majoring in Pathogenic Biology. Her research focuses on the analysis of microbiome and virome, with a particular emphasis on utilizing machine learning methods.

Yiqian Duan is a PhD student in BDB-Lab, at Fudan University, Shanghai China. She obtained her Bachelor's degree on Bioinformatics in Huazhong University of Science and Technology, Wuhan China (2020). Her main interest is small proteins and antibiotic resistance of the microbiome.

Jun Wang is a research group leader at the Institute of Microbiology, Chinese Academy of Science, Beijing China. He obtained PhD from Max-Planck Institute for Evolutionary Biology in Ploen, Germany, followed by Postdoc at VIB and Rega Institute in KU Leuven, Belgium. He started his own group in IM-CAS in 2017, supported by the Chinese central funding agencies and also by Max-Planck Society as a Partner group. His research focuses on employing state-of-art sequencing and analytical approaches for microbiome research, developed several bioinformatic tools to investigate structural variations in gut microbiome, as well as mycobiome and virome studies; utilizing the power of deep learning, his group has revealed the large reservoir of functional peptides including antimicrobial and anticancer peptides in human gut microbiome.

Luis Pedro Coelho is a group leader and Australian Research Council Future Fellow at the Centre for Microbiome Research (Queensland University of Technology, Brisbane). He holds a Masters degree in Computer Science (University of Lisbon) and a PhD in Computational Biology (Carnegie Mellon University). He did postdoctoral training with Peer Bork at the European Molecular Biology Laboratory (Heidelberg, Germany) and was a Junior Principal Investigator at Fudan University (Shanghai). His research group develops and applies computational methods to the study of the

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

global microbiome, with a focus on small proteins, including antimicrobial peptides.

Cesar de la Fuente-Nunez is a Presidential Associate Professor at the University of Pennsylvania, where he leads the Machine Biology Group. He completed postdoctoral research at the Massachusetts Institute of Technology (MIT) and earned a PhD from the University of British Columbia (UBC). He is best known for pioneering computational and artificial intelligence approaches to antibiotic discovery, which have drastically accelerated the time needed to identify preclinical candidates, from years to hours. These candidates show promise for therapeutic intervention against currently untreatable infections. Moreover, he spearheaded the discovery of therapeutic molecules from extinct organisms, and his lab has uncovered a myriad of novel peptide molecules across the tree of life, revealing a previously unrecognized branch of host immunity.

Ewa Szczurek is the director of the Institute of AI for Health at Helmholtz Munich, Germany, and leads joint laboratories at Helmholtz Munich and at the University of Warsaw, Poland. She holds Master degrees in computer science from the University of Warsaw, Poland and Uppsala University, Sweden. She obtained her doctoral degree from the Max Planck Institute for Molecular Genetics in Berlin, followed by a postdoctoral fellowship in Switzerland at ETH Zurich She was a visiting associate professor at Northwestern University in the United States and a visiting fellow at the Center for Interdisciplinary Research, Bielefeld, Germany. Prof. Szczurek acts as an area chair of the NeurIPS conference, a program committee member for the ISMB and RECOMB-CCB conferences, as well as associate editor for Genome Biology. Her research focuses on artificial intelligence, in particular probabilistic graphical models and deep generative models, and their applications in computational medicine. Her specific applications include oncology, pulmonology and the AI-driven design of antimicrobial peptides.

■ ACKNOWLEDGMENTS

This project has received funding from the European Research Council (ERC) under the European Funding Union's Horizon 2020 research and innovation programme (grant agreement No 810115 − DOG-AMP), by the Australian Research Council grant FT230100724 to L.P.C., and the Program of the Beijing Natural Science Foundation (JQ22017) to Jun Wang. Cesar de la Fuente-Nunez holds a Presidential Professorship at the University of Pennsylvania, is a recipient of the Langer Prize by the AIChE Foundation, and acknowledges funding from the IADR Innovation in Oral Care Award, the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, Penn Health-Tech Accelerator Award, the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania, the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM138201, and the Defense Threat Reduction Agency (DTRA; HDTRA1-22-10031, HDTRA1-21-1-0014, and HDTRA1-23-1-0001).

■ REFERENCES

(1) Szymczak, P.; Możejko, M.; Grzegorzek, T.; Jurczak, R.; Bauer, M.; Neubauer, D.; Sikora, K.; Michalski, M.; Sroka, J.; Setny, P.; Kamysz, W.; Szczurek, E. Discovering Highly Potent Antimicrobial Peptides with Deep Generative Model HydrAMP. *Nat. Commun.* **2023**, *14* (1), 1453.

(2) Wan, F.; Torres, M. D. T.; Peng, J.; de la Fuente-Nunez, C. Deep-Learning-Enabled Antibiotic Discovery through Molecular de-Extinction. *Nat. Biomed. Eng.* **2024**, *8* (7), 854−871.

(3) Santos-Junior, C. D.; Torres, M. D.T.; Duan, Y.; Rodriguez del Rio, A.; Schmidt, T. S.B.; Chong, H.; Fullam, A.; Kuhn, M.; Zhu, C.; Houseman, A.; Somborski, J.; Vines, A.; Zhao, X.-M.; Bork, P.; Huerta-Cepas, J.; de la Fuente-Nunez, C.; Coelho, L. P. Discovery of Antimicrobial Peptides in the Global Microbiome with Machine Learning. *Cell* **2024**, *187* (14), 3761.

(4) Torres, M. D. T.; Melo, M. C. R.; Flowers, L.; Crescenzi, O.; Notomista, E.; de la Fuente-Nunez, C. Mining for encrypted peptide antibiotics in the human proteome. *Nat. Biomed. Eng.* **2022**, *6* (1), 67−75.

(5) Ma, Y.; Guo, Z.; Xia, B.; Zhang, Y.; Liu, X.; Yu, Y.; Tang, N.; Tong, X.; Wang, M.; Ye, X.; Feng, J.; Chen, Y.; Wang, J. Identification of Antimicrobial Peptides from the Human Gut Microbiome Using Deep Learning. *Nat. Biotechnol.* **2022**, *40*, 921−931.

(6) Silver, L. L. Challenges of Antibacterial Discovery. *Clinical Microbiology Reviews* **2011**.

(7) Miethke, M.; Pieroni, M.; Weber, T.; Brönstrup, M.; Hammann, P.; Halby, L.; Arimondo, P. B.; Glaser, P.; Aigle, B.; Bode, H. B.; Moreira, R.; Li, Y.; Luzhetskyy, A.; Medema, M. H.; Pernodet, J.-L.; Stadler, M.; Tormo, J. R.; Genilloud, O.; Truman, A. W.; Weissman, K. J.; Takano, E.; Sabatini, S.; Stegmann, E.; Brötz-Oesterhelt, H.; Wohlleben, W.; Seemann, M.; Empting, M.; Hirsch, A. K. H.; Loretz, B.; Lehr, C.-M.; Titz, A.; Herrmann, J.; Jaeger, T.; Alt, S.; Hesterkamp, T.; Winterhalter, M.; Schiefer, A.; Pfarr, K.; Hoerauf, A.; Graz, H.; Graz, M.; Lindvall, M.; Ramurthy, S.; Karlén, A.; van Dongen, M.; Petkovic, H.; Keller, A.; Peyrane, F.; Donadio, S.; Fraisse, L.; Piddock, L. J. V.; Gilbert, I. H.; Moser, H. E.; Müller, R. Towards the Sustainable Discovery and Development of New Antibiotics. *Nat. Rev. Chem.* **2021**, *5* (10), 726−749.

(8) O'Neill, J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*. Review on Antimicrobial Resistance, 2016.

(9) Antimicrobial Resistance Division; Global Coordination and Partnership. *2021 Antibacterial agents in clinical and preclinical development: an overview and analysis*; World Health Organization, 2021. https://www.who.int/publications/i/item/9789240047655 (accessed 2024-10-18).

(10) WHO Antibacterial Pipeline Team. *Lack of innovation set to undermine antibiotic performance and health gains*. World Health Organization, 2022. https://www.who.int/news/item/22-06-2022-22-06-2022-lack-of-innovation-set-to-undermine-antibiotic-performance-and-health-gains (accessed 2024-10-18).

(11) *Five reasons to care about antimicrobial resistance (AMR)*. European Council. https://www.consilium.europa.eu/en/infographics/antimicrobial-resistance/ (accessed 2024-10-18).

(12) Huan, Y.; Kong, Q.; Mou, H.; Yi, H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front. Microbiol.* **2020**, *11*, 582779.

(13) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; Cherkasov, A.; Seleem, M. N.; Pinilla, C.; de la Fuente-Nunez, C.; Lazaridis, T.; Dai, T.; Houghten, R. A.; Hancock, R. E. W.; Tegos, G. P. The Value of Antimicrobial Peptides in the Age of Resistance. *Lancet Infect. Dis.* **2020**, *20* (9), e216−e230.

(14) Greco, I.; Molchanova, N.; Holmedal, E.; Jenssen, H.; Hummel, B. D.; Watts, J. L.; Håkansson, J.; Hansen, P. R.; Svenson, J. Correlation between Hemolytic Activity, Cytotoxicity and Systemic in Vivo Toxicity of Synthetic Antimicrobial Peptides. *Sci. Rep.* **2020**, *10* (1), 13206.

(15) Szymczak, P.; Szczurek, E. Artificial Intelligence-Driven Antimicrobial Peptide Discovery. *Curr. Opin. Struct. Biol.* **2023**, *83*, 102733.

(16) Lei, J.; Sun, L.; Huang, S.; Zhu, C.; Li, P.; He, J.; Mackey, V.; Coy, D. H.; He, Q. The Antimicrobial Peptides and Their Potential Clinical Applications. *Am. J. Transl. Res.* **2019**, *11* (7), 3919.

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

(17) Torres, M. D. T.; de la Fuente-Nunez, C. Toward Computer-Made Artificial Antibiotics. *Curr. Opin. Microbiol.* **2019**, *51*, 30−38.
(18) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: Database of Antimicrobial/Cytotoxic Activity and Structure of Peptides as a Resource for Development of New Therapeutics. *Nucleic Acids Res.* **2021**, *49* (D1), D288−D297.
(19) Maasch, J. R. M. A.; Torres, M. D. T.; Melo, M. C. R.; de la Fuente-Nunez, C. Molecular De-Extinction of Ancient Antimicrobial Peptides Enabled by Machine Learning. *Cell Host Microbe* **2023**, *31* (8), 1260−1274.
(20) Porto, W. F.; Irazazabal, L.; Alves, E. S. F.; Ribeiro, S. M.; Matos, C. O.; Pires, Á. S.; Fensterseifer, I. C. M.; Miranda, V. J.; Haney, E. F.; Humblot, V.; Torres, M. D. T.; Hancock, R. E. W.; Liao, L. M.; Ladram, A.; Lu, T. K.; de la Fuente-Nunez, C.; Franco, O. L. In Silico Optimization of a Guava Antimicrobial Peptide Enables Combinatorial Exploration for Peptide Design. *Nat. Commun.* **2018**, *9* (1), 1490.
(21) Yan, K.; Lv, H.; Guo, Y.; Peng, W.; Liu, B. sAMPpred-GAT: Prediction of Antimicrobial Peptide by Graph Attention Network and Predicted Peptide Structure. *Bioinformatics* **2023**, *39* (1), btac715.
(22) Li, C.; Sutherland, D.; Hammond, S. A.; Yang, C.; Taho, F.; Bergman, L.; Houston, S.; Warren, R. L.; Wong, T.; Hoang, L. M. N.; Cameron, C. E.; Helbing, C. C.; Birol, I. AMPlify: Attentive Deep Learning Model for Discovery of Novel Antimicrobial Peptides Effective against WHO Priority Pathogens. *BMC Genomics* **2022**, *23* (1), 77.
(23) Li, C.; Zou, Q.; Jia, C.; Zheng, J. AMPpred-MFA: An Interpretable Antimicrobial Peptide Predictor with a Stacking Architecture, Multiple Features, and Multihead Attention. *J. Chem. Inf. Model.* **2024**, *64*, 2393.
(24) Yan, J.; Zhang, B.; Zhou, M.; Campbell-Valois, F.-X.; Siu, S. W. I. A Deep Learning Method for Predicting the Minimum Inhibitory Concentration of Antimicrobial Peptides against Escherichia Coli Using Multi-Branch-CNN and Attention. *msystems* **2023**, *8*, e00345-23.
(25) Pandi, A.; Adam, D.; Zare, A.; Trinh, V. T.; Schaefer, S. L.; Burt, M.; Klabunde, B.; Bobkova, E.; Kushwaha, M.; Foroughijabbari, Y.; Braun, P.; Spahn, C.; Preußer, C.; Pogge von Strandmann, E.; Bode, H. B.; von Buttlar, H.; Bertrams, W.; Jung, A. L.; Abendroth, F.; Schmeck, B.; Hummer, G.; Vázquez, O.; Erb, T. J. Cell-Free Biosynthesis Combined with Deep Learning Accelerates de Novo-Development of Antimicrobial Peptides. *Nat. Commun.* **2023**, *14*, 7197.
(26) Tsai, C.-T.; Lin, C.-W.; Ye, G.-L.; Wu, S.-C.; Yao, P.; Lin, C.-T.; Wan, L.; Tsai, H.-H. G. Accelerating Antimicrobial Peptide Discovery for WHO Priority Pathogens through Predictive and Interpretable Machine Learning Models. *ACS Omega* **2024**, *9*, 9357.
(27) Sharma, R.; Shrivastava, S.; Singh, S. K.; Kumar, A.; Singh, A. K.; Saxena, S. EnDL-HemoLyt: Ensemble Deep Learning-Based Tool for Identifying Therapeutic Peptides with Low Hemolytic Activity. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1896.
(28) Santos-Júnior, C. D.; Pan, S.; Zhao, X.-M.; Coelho, L. P. Macrel: Antimicrobial Peptide Screening in Genomes and Meta-genomes. *PeerJ.* **2020**, *8*, No. e10555.
(29) Capecchi, A.; Cai, X.; Personne, H.; Köhler, T.; van Delden, C.; Reymond, J.-L. Machine Learning Designs Non-Hemolytic Antimicrobial Peptides. *Chem. Sci.* **2021**, *12* (26), 9221−9232.
(30) Ansari, M.; White, A. D. Learning Peptide Properties with Positive Examples Only. *Digit. Discovery* **2024**, *3* (5), 977−986.
(31) Sharma, R.; Shrivastava, S.; Singh, S. K.; Kumar, A.; Singh, A. K.; Saxena, S. Artificial Intelligence-Based Model for Predicting the Minimum Inhibitory Concentration of Antibacterial Peptides against ESKAPEE Pathogens. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 1949.
(32) Ansari, M.; White, A. D. Serverless Prediction of Peptide Properties with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2023**, *63* (8), 2546−2553.
(33) Xu, J.; Li, F.; Leier, A.; Xiang, D.; Shen, H.-H.; Marquez Lago, T. T.; Li, J.; Yu, D.-J.; Song, J. Comprehensive Assessment of Machine Learning-Based Methods for Predicting Antimicrobial Peptides. *Brief. Bioinform.* **2021**, *22* (5), bbab083.
(34) Lawrence, T. J.; Carper, D. L.; Spangler, M. K.; Carrell, A. A.; Rush, T. A.; Minter, S. J.; Weston, D. J.; Labbé, J. L. amPEPpy 1.0: A Portable and Accurate Antimicrobial Peptide Prediction Tool. *Bioinforma. Oxf. Engl.* **2021**, *37* (14), 2058−2060.
(35) García-Jacas, C. R.; Pinacho-Castellanos, S. A.; García-González, L. A.; Brizuela, C. A. Do Deep Learning Models Make a Difference in the Identification of Antimicrobial Peptides? *Brief. Bioinform.* **2022**, *23* (3), bbac094.
(36) Bárcenas, O.; Pintado-Grima, C.; Sidorczuk, K.; Teufel, F.; Nielsen, H.; Ventura, S.; Burdukiewicz, M. The Dynamic Landscape of Peptide Activity Prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 6526−6533.
(37) Wan, F.; Wong, F.; Collins, J. J.; de la Fuente-Nunez, C. Machine Learning for Antimicrobial Peptide Identification and Design. *Nat. Rev. Bioeng.* **2024**, *2* (5), 392−407.
(38) Zhuang, S.; Tanner, J.; Wu, Y.; Huynh, D.; Liu, W.; Cadet, X.; Fontaine, N.; Charton, P.; Damour, C.; Cadet, F.; Wang, J. Non-Hemolytic Peptide Classification Using a Quantum Support Vector Machine. *Quantum Inf. Process.* **2024**, *23* (11), 379.
(39) Ruffolo, J. A.; Madani, A. Designing Proteins with Language Models. *Nat. Biotechnol.* **2024**, *42* (2), 200−202.
(40) Xing, W.; Zhang, J.; Li, C.; Huo, Y.; Dong, G. iAMP-Attenpred: A Novel Antimicrobial Peptide Predictor Based on BERT Feature Extraction Method and CNN-BiLSTM-Attention Combination Model. *Brief. Bioinform.* **2023**, *25* (1), bbad443.
(41) Zhang, R.; Wu, H.; Liu, C.; Li, H.; Wu, Y.; Li, K.; Wang, Y.; Deng, Y.; Chen, J.; Zhou, F.; Gao, X. PepHarmony: A Multi-View Contrastive Learning Framework for Integrated Sequence and Structure-Based Peptide Encoding. *arXiv* **2024**, 2401.11360.
(42) Zhang, W.; Xu, Y.; Wang, A.; Chen, G.; Zhao, J. Fuse Feeds as One: Cross-Modal Framework for General Identification of AMPs. *Brief. Bioinform.* **2023**, *24* (6), bbad336.
(43) Yu, Q.; Dong, Z.; Fan, X.; Zong, L.; Li, Y. HMD-AMP: Protein Language-Powered Hierarchical Multi-Label Deep Forest for Annotating Antimicrobial Peptides. *arXiv* **2021**, 2111.06023.
(44) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44* (10), 7112−7127.
(45) Salem, M.; Keshavarzi Arshadi, A.; Yuan, J. S. AMPDeep: Hemolytic Activity Prediction of Antimicrobial Peptides Using Transfer Learning. *BMC Bioinformatics* **2022**, *23* (1), 389.
(46) Orsi, M.; Reymond, J.-L. Can Large Language Models Predict Antimicrobial Peptide Activity and Toxicity? *RSC Med. Chem.* **2024**, *15* (6), 2030.
(47) Guntuboina, C.; Das, A.; Mollaei, P.; Kim, S.; Barati Farimani, A. PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction. *J. Phys. Chem. Lett.* **2023**, *14* (46), 10427−10434.
(48) Zhang, Y.; Lin, J.; Zhao, L.; Zeng, X.; Liu, X. A Novel Antibacterial Peptide Recognition Algorithm Based on BERT. *Brief. Bioinform.* **2021**, *22* (6), bbab200.
(49) Jiang, Y.; Wang, R.; Feng, J.; Jin, J.; Liang, S.; Li, Z.; Yu, Y.; Ma, A.; Su, R.; Zou, Q.; Ma, Q.; Wei, L. Explainable Deep Hypergraph Learning Modeling the Peptide Secondary Structure Prediction. *Adv. Sci.* **2023**, *10* (11), 2206151.
(50) Sadeh, G.; Wang, Z.; Grewal, J.; Rangwala, H.; Price, L. Training Self-Supervised Peptide Sequence Models on Artificially Chopped Proteins. *arXiv* **2022**, 2211.06428.
(51) Yang, S.; Yang, Z.; Ni, X. AMPFinder: A Computational Model to Identify Antimicrobial Peptides and Their Functions Based on Sequence-Derived Information. *Anal. Biochem.* **2023**, *673*, 115196.

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

(52) Dee, W. LMPred: Predicting Antimicrobial Peptides Using Pre-Trained Language Models and Deep Learning. *Bioinforma. Adv.* **2022**, *2* (1), vbac021.
(53) Pang, Y.; Yao, L.; Xu, J.; Wang, Z.; Lee, T.-Y. Integrating Transformer and Imbalanced Multi-Label Learning to Identify Antimicrobial Peptides and Their Functional Activities. *Bioinformatics* **2022**, *38* (24), 5368−5374.
(54) García-Jacas, C. R.; García-González, L. A.; Martinez-Rios, F.; Tapia-Contreras, I. P.; Brizuela, C. A. Handcrafted versus Non-Handcrafted (Self-Supervised) Features for the Classification of Antimicrobial Peptides: Complementary or Redundant? *Brief. Bioinform.* **2022**, *23* (6), bbac428.
(55) Pandey, P.; Srivastava, A. sAMP-VGG16: Drude Polarizable Force-Field Assisted Image-Based Deep Neural Network Prediction Model for Short Antimicrobial Peptides. *bioRxiv* **2023**, DOI: 10.1101/2023.06.04.543607.
(56) Dong, R.; Liu, R.; Liu, Z.; Liu, Y.; Zhao, G.; Li, H.; Hou, S.; Ma, X.; Kang, H.; Liu, J.; Guo, F.; Zhao, P.; Wang, J.; Wang, C.; Wu, X.; Ye, S.; Zhu, C. Exploring the Repository of de Novo Designed Bifunctional Antimicrobial Peptides through Deep Learning. *bioRxiv* **2024**, DOI: 10.1101/2024.02.23.581845.
(57) Conlon, J. M.; Sonnevend, A. Antimicrobial Peptides in Frog Skin Secretions. In *Antimicrobial Peptides: Methods and Protocols*; Giuliani, A.; Rinaldi, A. C., Eds.; Humana Press: Totowa, NJ, 2010; pp 3−14. DOI: 10.1007/978-1-60761-594-1_1.
(58) Torres, M. D. T.; Brooks, E. F.; Cesaro, A.; Sberro, H.; Gill, M. O.; Nicolaou, C.; Bhatt, A. S.; de la Fuente-Nunez, C. Mining Human Microbiomes Reveals an Untapped Source of Peptide Antibiotics. *Cell* **2024**, *187* (19), 5453−5467.
(59) Fullam, A.; Letunic, I.; Schmidt, T. S. B.; Ducarmon, Q. R.; Karcher, N.; Khedkar, S.; Kuhn, M.; Larralde, M.; Maistrenko, O. M.; Malfertheiner, L.; Milanese, A.; Rodrigues, J. F. M.; Sanchis-López, C.; Schudoma, C.; Szklarczyk, D.; Sunagawa, S.; Zeller, G.; Huerta-Cepas, J.; von Mering, C.; Bork, P.; Mende, D. R. proGenomes3: Approaching One Million Accurately and Consistently Annotated High-Quality Prokaryotic Genomes. *Nucleic Acids Res.* **2023**, *51* (D1), D760−D766.
(60) Duan, Y.; Santos-Júnior, C. D.; Schmidt, T. S.; Fullam, A.; de Almeida, B. L. S.; Zhu, C.; Kuhn, M.; Zhao, X.-M.; Bork, P.; Coelho, L. P. A Catalog of Small Proteins from the Global Microbiome. *Nat. Commun.* **2024**, *15* (1), 7563.
(61) The NIH HMP Working Group; Peterson, J.; Garges, S.; Giovanni, M.; McInnes, P.; Wang, L.; Schloss, J. A.; Bonazzi, V.; McEwen, J. E.; Wetterstrand, K. A.; Deal, C.; Baker, C. C.; Di Francesco, V.; Howcroft, T. K.; Karp, R. W.; Lunsford, R. D.; Wellington, C. R.; Belachew, T.; Wright, M.; Giblin, C.; David, H.; Mills, M.; Salomon, R.; Mullins, C.; Akolkar, B.; Begg, L.; Davis, C.; Grandison, L.; Humble, M.; Khalsa, J.; Little, A. R.; Peavy, H.; Pontzer, C.; Portnoy, M.; Sayre, M. H.; Starke-Reed, P.; Zakhari, S.; Read, J.; Watson, B.; Guyer, M. The NIH Human Microbiome Project. *Genome Res.* **2009**, *19* (12), 2317−2323.
(62) Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M. L.; Burdett, T.; Burgin, J.; Caballero-Pérez, J.; Cochrane, G.; Colwell, L. J.; Curtis, T.; Escobar-Zepeda, A.; Gurbich, T. A.; Kale, V.; Korobeynikov, A.; Raj, S.; Rogers, A. B.; Sakharova, E.; Sanchez, S.; Wilkinson, D. J.; Finn, R. D. MGnify: The Microbiome Sequence Data Analysis Resource in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D753−D759.
(63) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.
(64) Coelho, L. P.; Alves, R.; del Río, Á. R.; Myers, P. N.; Cantalapiedra, C. P.; Giner-Lamia, J.; Schmidt, T. S.; Mende, D. R.; Orakov, A.; Letunic, I.; Hildebrand, F.; Van Rossum, T.; Forslund, S. K.; Khedkar, S.; Maistrenko, O. M.; Pan, S.; Jia, L.; Ferretti, P.; Sunagawa, S.; Zhao, X.-M.; Nielsen, H. B.; Huerta-Cepas, J.; Bork, P. Towards the Biogeography of Prokaryotic Genes. *Nature* **2022**, *601* (7892), 252−256.
(65) Alcock, B. P.; Huynh, W.; Chalil, R.; Smith, K. W.; Raphenya, A. R.; Wlodarski, M. A.; Edalatmand, A.; Petkau, A.; Syed, S. A.; Tsang, K. K.; Baker, S. J. C.; Dave, M.; McCarthy, M. C.; Mukiri, K. M.; Nasir, J. A.; Golbon, B.; Imtiaz, H.; Jiang, X.; Kaur, K.; Kwong, M.; Liang, Z. C.; Niu, K. C.; Shan, P.; Yang, J. Y. J.; Gray, K. L.; Hoad, G. R.; Jia, B.; Bhando, T.; Carfrae, L. A.; Farha, M. A.; French, S.; Gordzevich, R.; Rachwalski, K.; Tu, M. M.; Bordeleau, E.; Dooley, D.; Griffiths, E.; Zubyk, H. L.; Brown, E. D.; Maguire, F.; Beiko, R. G.; Hsiao, W. W. L.; Brinkman, F. S. L.; Van Domselaar, G.; McArthur, A. G. CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **2023**, *51* (D1), D690−D699.
(66) Cesaro, A.; Torres, M. D. T.; Gaglione, R.; Dell'Olmo, E.; Di Girolamo, R.; Bosso, A.; Pizzo, E.; Haagsman, H. P.; Veldhuizen, E. J. A.; de la Fuente-Nunez, C.; Arciello, A. Synthetic Antibiotic Derived from Sequences Encrypted in a Protein from Human Plasma. *ACS Nano* **2022**, *16* (2), 1880−1895.
(67) Monsalve, D.; Mesa, A.; Mira, L. M.; Mera, C.; Orduz, S.; Branch-Bedoya, J. W. Antimicrobial Peptides Designed by Computational Analysis of Proteomes. *Antonie Van Leeuwenhoek* **2024**, *117* (1), 55.
(68) Torres, M. D. T.; Cesaro, A.; de la Fuente-Nunez, C. Peptides from Non-Immune Proteins Target Infections through Antimicrobial and Immunomodulatory Properties. *Trends Biotechnol.* **2025**, *43* (1), 184−205.
(69) de la Fuente-Nunez, C. Mining Biology for Antibiotic Discovery. *PLOS Biol.* **2024**, *22* (11), No. e3002946.
(70) Wong, F.; de la Fuente-Nunez, C.; Collins, J. J. Leveraging Artificial Intelligence in the Fight against Infectious Diseases. *Science* **2023**, *381* (6654), 164−170.
(71) Wu, H.; Chen, R.; Li, X.; Zhang, Y.; Zhang, J.; Yang, Y.; Wan, J.; Zhou, Y.; Chen, H.; Li, J.; Li, R.; Zou, G. ESKtides: A Comprehensive Database and Mining Method for ESKAPE Phage-Derived Antimicrobial Peptides. *Database* **2024**, *2024*, baae022.
(72) Ma, Y.; Liu, X.; Zhang, X.; Yu, Y.; Li, Y.; Song, M.; Wang, J. Efficient Mining of Anticancer Peptides from Gut Metagenome. *Adv. Sci.* **2023**, *10* (25), 2300107.
(73) Klimovich, A.; Bosch, T. C. G. Novel Technologies Uncover Novel 'Anti'-Microbial Peptides in *Hydra* Shaping the Species-Specific Microbiome. *Philos. Trans. R. Soc. B Biol. Sci.* **2024**, *379* (1901), 20230058.
(74) Chen, S.; Qi, H.; Zhu, X.; Liu, T.; Teng, Y.; Gong, Q.; Jia, C.; Liu, T.; Chen, S.; Qi, H.; Zhu, X.; Liu, T.; Teng, Y.; Gong, Q.; Jia, C.; Liu, T. The Discovery of Antimicrobial Peptides from the Gut Microbiome of Cockroach Blattella Germanica Using Deep Learning Pipeline. *bioRxiv* **2024**, DOI: 10.1101/2024.02.12.580024.
(75) Huang, J.; Xu, Y.; Xue, Y.; Huang, Y.; Li, X.; Chen, X.; Xu, Y.; Zhang, D.; Zhang, P.; Zhao, J.; Ji, J. Identification of Potent Antimicrobial Peptides via a Machine-Learning Pipeline That Mines the Entire Space of Peptide Sequences. *Nat. Biomed. Eng.* **2023**, *7*, 797.
(76) Witten, J.; Witten, Z. Deep Learning Regression Model for Antimicrobial Peptide Design. *bioRxiv* **2019**, DOI: 10.1101/692681.
(77) Pane, K.; Durante, L.; Crescenzi, O.; Cafaro, V.; Pizzo, E.; Varcamonti, M.; Zanfardino, A.; Izzo, V.; Di Donato, A.; Notomista, E. Antimicrobial Potency of Cationic Antimicrobial Peptides Can Be Predicted from Their Amino Acid Composition: Application to the Detection of "Cryptic" Antimicrobial Peptides. *J. Theor. Biol.* **2017**, *419*, 254−265.
(78) Ji, J.; Huang, J.; Zhang, W.; Wang, A.; Lai, Y.; Xu, Y.; Wang, C.; Zhao, J.; Zhang, P. Discovery of Antimicrobial Peptides Targeting Acinetobacter Baumannii via a Pre-Trained and Fine-Tuned Few-Shot Learning-Based Pipeline. *Research Square* **2024**, DOI: 10.21203/rs.3.rs-3789296/v1.
(79) Martinelli, D. D. Generative Machine Learning for de Novo Drug Discovery: A Systematic Review. *Comput. Biol. Med.* **2022**, *145*, 105403.
(80) Das, P.; Sercu, T.; Wadhawan, K.; Padhi, I.; Gehrmann, S.; Cipcigan, F.; Chenthamarakshan, V.; Strobelt, H.; Dos Santos, C.;

Chen, P.-Y.; Yang, Y. Y.; Tan, J. P. K.; Hedrick, J.; Crain, J.; Mojsilovic, A. Accelerated Antimicrobial Discovery via Deep Generative Models and Molecular Dynamics Simulations. *Nat. Biomed. Eng.* **2021**, *5* (6), 613−623.
(81) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688−702.
(82) Mao, J.; Guan, S.; Chen, Y.; Zeb, A.; Sun, Q.; Lu, R.; Dong, J.; Wang, J.; Cao, D. Application of a Deep Generative Model Produces Novel and Diverse Functional Peptides against Microbial Resistance. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 463−471.
(83) Wang, D.; Wen, Z.; Ye, F.; Zhou, H.; Li, L. Accelerating Antimicrobial Peptide Discovery with Latent Sequence-Structure Model. *arXiv* **2022**, 2212.09450.
(84) Renaud, S.; Mansbach, R. A. Latent Spaces for Antimicrobial Peptide Design. *Digit. Discovery* **2023**, *2* (2), 441−458.
(85) Wang, Y.; Gong, H.; Li, X.; Li, L.; Zhao, Y.; Bao, P.; Kong, Q.; Wan, B.; Zhang, Y.; Zhang, J.; Ni, J.; Han, Z.; Nan, X.; Ju, K.; Sun, L.; Chang, H.; Zheng, M.; Yu, Y.; Yang, X.; Zuo, X.; Li, Y. De Novo Multi-Mechanism Antimicrobial Peptide Design via Multimodal Deep Learning. *bioRxiv* **2024**, DOI: 10.1101/2024.01.02.573846.
(86) Dean, S. N.; Alvarez, J. A. E.; Zabetakis, D.; Walper, S. A.; Malanoski, A. P. PepVAE: Variational Autoencoder Framework for Antimicrobial Peptide Generation and Activity Prediction. *Front. Microbiol.* **2021**, *12*, 2764.
(87) Tučs, A.; Berenger, F.; Yumoto, A.; Tamura, R.; Uzawa, T.; Tsuda, K. Quantum Annealing Designs Nonhemolytic Antimicrobial Peptides in a Discrete Latent Space. *ACS Med. Chem. Lett.* **2023**, *14* (5), 577−582.
(88) Ghorbani, M.; Prasad, S.; Brooks, B. R.; Klauda, J. B. Deep Attention Based Variational Autoencoder for Antimicrobial Peptide Discovery. *bioRxiv* **2022**, DOI: 10.1101/2022.07.08.499340.
(89) Hoffman, S. C.; Chenthamarakshan, V.; Wadhawan, K.; Chen, P.-Y.; Das, P. Optimizing Molecules Using Efficient Queries from Property Evaluations. *Nat. Mach. Intell.* **2022**, *4* (1), 21−31.
(90) Yu, H.; Wang, R.; Qiao, J.; Wei, L. Multi-CGAN: Deep Generative Model-Based Multiproperty Antimicrobial Peptide Design. *J. Chem. Inf. Model.* **2024**, *64* (1), 316−326.
(91) Surana, S.; Arora, P.; Singh, D.; Sahasrabuddhe, D.; Valadi, J. PandoraGAN: Generating Antiviral Peptides Using Generative Adversarial Network. *bioRxiv* **2022**, DOI: 10.1101/2021.02.15.431193.
(92) Cao, Q.; Ge, C.; Wang, X.; Harvey, P. J.; Zhang, Z.; Ma, Y.; Wang, X.; Jia, X.; Mobli, M.; Craik, D. J.; Jiang, T.; Yang, J.; Wei, Z.; Wang, Y.; Chang, S.; Yu, R. Designing Antimicrobial Peptides Using Deep Learning and Molecular Dynamic Simulations. *Brief. Bioinform.* **2023**, *24* (2), bbad058.
(93) Ferrell, J. B.; Remington, J. M.; Oort, C. M. V.; Sharafi, M.; Aboushousha, R.; Janssen-Heininger, Y.; Schneebeli, S. T.; Wargo, M. J.; Wshah, S.; Li, J. A Generative Approach toward Precision Antimicrobial Peptide Design. *bioRxiv* **2021**, DOI: 10.1101/2020.10.02.324087.
(94) Van Oort, C. M.; Ferrell, J. B.; Remington, J. M.; Wshah, S.; Li, J. AMPGAN v2: Machine Learning-Guided Design of Antimicrobial Peptides. *J. Chem. Inf. Model.* **2021**, *61* (5), 2198−2207.
(95) Wan, F.; Wong, F.; Collins, J. J.; de la Fuente-Nunez, C. Machine Learning for Antimicrobial Peptide Identification and Design. *Nat. Rev. Bioeng.* **2024**, *2* (5), 392−407.
(96) Jain, M.; Bengio, E.; Hernandez-Garcia, A.; Rector-Brooks, J.; Dossou, B. F. P.; Ekbote, C. A.; Fu, J.; Zhang, T.; Kilgour, M.; Zhang, D.; Simine, L.; Das, P.; Bengio, Y. Biological Sequence Design with GFlowNets. In *Proceedings of the 39th International Conference on Machine Learning*; PMLR, 2022; pp 9786−9801.
(97) Murakami, Y.; Ishida, S.; Demizu, Y.; Terayama, K. Design of Antimicrobial Peptides Containing Non-Proteinogenic Amino Acids Using Multi-Objective Bayesian Optimisation. *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-cbrjc-v2.
(98) Liu, Y.; Zhang, X.; Liu, Y.; Su, Y.; Zeng, X.; Yen, G. G. Evolutionary Multi-Objective Optimization in Searching for Various Antimicrobial Peptides [Feature]. *IEEE Comput. Intell. Mag.* **2023**, *18* (2), 31−45.
(99) Zeng, Z.; Xu, R.; Guo, J.; Luo, X. Binary Discriminator Facilitates GPT-Based Protein Design. *bioRxiv* **2023**, DOI: 10.1101/2023.11.20.567789.
(100) Wang, R.; Wang, T.; Zhuo, L.; Wei, J.; Fu, X.; Zou, Q.; Yao, X. Diff-AMP: Tailored Designed Antimicrobial Peptide Framework with All-in-One Generation, Identification, Prediction and Optimization. *Brief. Bioinform.* **2024**, *25* (2), bbae078.
(101) Torres, M. D. T.; Chen, T.; Wan, F.; Chatterjee, P.; de la Fuente-Nunez, C. Generative Latent Diffusion Language Modeling Yields Anti-Infective Synthetic Peptides. *bioRxiv* **2025**, DOI: 10.1101/2025.01.31.636003.
(102) Wang, Y.; Liu, X.; Huang, F.; Xiong, Z.; Zhang, W. A Multi-Modal Contrastive Diffusion Model for Therapeutic Peptide Generation. *arXiv* **2024**, 2312.15665 DOI: 10.48550/arXiv.2312.15665.
(103) Buehler, M. J. Generative Pretrained Autoregressive Transformer Graph Neural Network Applied to the Analysis and Discovery of Novel Proteins. *arXiv* **2023**, 2305.04934 DOI: 10.48550/arXiv.2305.04934.
(104) Wang, X.-F.; Tang, J.-Y.; Liang, H.; Sun, J.; Dorje, S.; Peng, B.; Ji, X.-W.; Li, Z.; Zhang, X.-E.; Wang, D.-B. ProT-Diff: A Modularized and Efficient Approach to De Novo Generation of Antimicrobial Peptide Sequences through Integration of Protein Language Model and Diffusion Model. *bioRxiv* **2024**, DOI: 10.1101/2024.02.22.581480.
(105) de la Fuente-Nunez, C. AI in Infectious Diseases: The Role of Datasets. *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.* **2024**, *73*, 101067.
(106) Sidorczuk, K.; Gagat, P.; Pietluch, F.; Kała, J.; Rafacz, D.; Bąkała, L.; Słowik, J.; Kolenda, R.; Rödiger, S.; Fingerhut, L. C. H. W.; Cooke, I. R.; Mackiewicz, P.; Burdukiewicz, M. Benchmarks in Antimicrobial Peptide Prediction Are Biased Due to the Selection of Negative Data. *Brief. Bioinform.* **2022**, *23*, bbac343.
(107) Li, W.; Separovic, F.; O'Brien-Simpson, N. M.; Wade, J. D. Chemically Modified and Conjugated Antimicrobial Peptides against Superbugs. *Chem. Soc. Rev.* **2021**, *50* (8), 4932−4973.
(108) Trimble, M. J.; Mlynárčik, P.; Kolář, M.; Hancock, R. E. W. Polymyxin: Alternative Mechanisms of Action and Resistance. *Cold Spring Harb. Perspect. Med.* **2016**, *6* (10), a025288.
(109) Awdhesh Kumar Mishra, R.; Kodiveri Muthukaliannan, G. In-Silico and in-Vitro Study of Novel Antimicrobial Peptide AM1 from Aegle Marmelos against Drug-Resistant Staphylococcus Aureus. *Sci. Rep.* **2024**, *14* (1), 25822.
(110) Bundalovic-Torma, C.; Whitfield, G. B.; Marmont, L. S.; Howell, P. L.; Parkinson, J. A Systematic Pipeline for Classifying Bacterial Operons Reveals the Evolutionary Landscape of Biofilm Machineries. *PLoS Comput. Biol.* **2020**, *16* (4), No. e1007721.
(111) Miller, D.; Stern, A.; Burstein, D. Deciphering Microbial Gene Function Using Natural Language Processing. *Nat. Commun.* **2022**, *13* (1), 5731.
(112) Hwang, Y.; Cornman, A. L.; Kellogg, E. H.; Ovchinnikov, S.; Girguis, P. R. Genomic Language Model Predicts Protein Co-Regulation and Function. *Nat. Commun.* **2024**, *15* (1), 2880.
(113) Duan, C.; Zang, Z.; Xu, Y.; He, H.; Liu, Z.; Song, Z.; Zheng, J.-S.; Li, S. Z. FGBERT: Function-Driven Pre-Trained Gene Language Model for Metagenomics. *arXiv* **2024**, 2402.16901.
(114) Salazar, G.; Paoli, L.; Alberti, A.; Huerta-Cepas, J.; Ruscheweyh, H.-J.; Cuenca, M.; Field, C. M.; Coelho, L. P.; Cruaud, C.; Engelen, S.; Gregory, A. C.; Labadie, K.; Marec, C.; Pelletier, E.; Royo-Llonch, M.; Roux, S.; Sánchez, P.; Uehara, H.; Zayed, A. A.; Zeller, G.; Carmichael, M.; Dimier, C.; Ferland, J.; Kandels, S.; Picheral, M.; Pisarev, S.; Poulain, J.; Acinas, S. G.; Babin, M.; Bork, P.; Boss, E.; Bowler, C.; Cochrane, G.; de Vargas, C.;

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX

Follows, M.; Gorsky, G.; Grimsley, N.; Guidi, L.; Hingamp, P.; Iudicone, D.; Jaillon, O.; Kandels-Lewis, S.; Karp-Boss, L.; Karsenti, E.; Not, F.; Ogata, H.; Pesant, S.; Poulton, N.; Raes, J.; Sardet, C.; Speich, S.; Stemmann, L.; Sullivan, M. B.; Sunagawa, S.; Wincker, P.; Acinas, S. G.; Babin, M.; Bork, P.; Bowler, C.; de Vargas, C.; Guidi, L.; Hingamp, P.; Iudicone, D.; Karp-Boss, L.; Karsenti, E.; Ogata, H.; Pesant, S.; Speich, S.; Sullivan, M. B.; Wincker, P.; Sunagawa, S. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **2019**, *179* (5), 1068−1083.

(115) Fremin, B. J.; Sberro, H.; Bhatt, A. S. MetaRibo-Seq Measures Translation in Microbiomes. *Nat. Commun.* **2020**, *11* (1), 3268.

(116) Paoli, L.; Ruscheweyh, H.-J.; Forneris, C. C.; Hubrich, F.; Kautsar, S.; Bhushan, A.; Lotti, A.; Clayssen, Q.; Salazar, G.; Milanese, A.; Carlström, C. I.; Papadopoulou, C.; Gehrig, D.; Karasikov, M.; Mustafa, H.; Larralde, M.; Carroll, L. M.; Sánchez, P.; Zayed, A. A.; Cronin, D. R.; Acinas, S. G.; Bork, P.; Bowler, C.; Delmont, T. O.; Gasol, J. M.; Gossert, A. D.; Kahles, A.; Sullivan, M. B.; Wincker, P.; Zeller, G.; Robinson, S. L.; Piel, J.; Sunagawa, S. Biosynthetic Potential of the Global Ocean Microbiome. *Nature* **2022**, *607* (7917), 111−118.

(117) Carroll, L. M.; Larralde, M.; Fleck, J. S.; Ponnudurai, R.; Milanese, A.; Cappio, E.; Zeller, G. Accurate de Novo Identification of Biosynthetic Gene Clusters with GECCO. *bioRxiv* **2021**, DOI: 10.1101/2021.05.03.442509.

(118) Sberro, H.; Fremin, B. J.; Zlitni, S.; Edfors, F.; Greenfield, N.; Snyder, M. P.; Pavlopoulos, G. A.; Kyrpides, N. C.; Bhatt, A. S. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* **2019**, *178* (5), 1245−1259.

https://doi.org/10.1021/acs.accounts.0c00594
Acc. Chem. Res. XXXX, XXX, XXX−XXX