ACS Synthetic Biology

pubs.acs.org/synthbio Research Article

Machine Learning-Enabled Genome Mining and Bioactivity Prediction of Natural Products

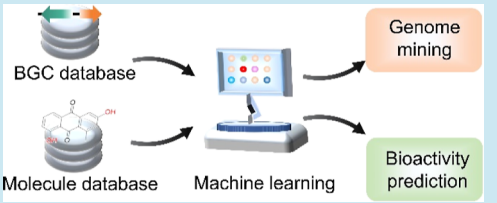
Published as part of the ACS Synthetic Biology virtual special issue "AI for Synthetic Biology".

Yujie Yuan, Chengyou Shi, and Huimin Zhao*

Cite This: *ACS Synth. Biol.* 2023, 12, 2650–2662 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Natural products (NPs) produced by microorganisms and plants are a major source of drugs, herbicides, and fungicides. Thanks to recent advances in DNA sequencing, bioinformatics, and genome mining tools, a vast amount of data on NP biosynthesis has been generated over the years, which has been increasingly exploited to develop machine learning (ML) tools for NP discovery. In this review, we discuss the latest advances in developing and applying ML tools for exploring the potential NPs that can be encoded by genomic language and predicting the types of bioactivities of NPs. We also examine the technical challenges associated with the development and application of ML tools for NP research.



KEYWORDS: *machine learning, natural product, genome mining, biosynthetic gene cluster, bioactivity prediction, model construction*

INTRODUCTION

For thousands of years, natural products (NPs) have been crucial to human health and well-being.[1] Recent advances in DNA sequencing, bioinformatics, and genome mining have made the discovery of NPs more efficient. However, with more compounds being discovered, it has become increasingly challenging to avoid discovery of previously characterized NPs. Additionally, exploring the biological functions of NPs remains difficult, particularly as some NPs exist in very small quantities, preventing extensive screening of their bioactivity. To aid in the discovery of NPs and characterization of their bioactivity, researchers have developed various strategies such as high-throughput biosynthetic gene cluster (BGC) discovery,[2,3] BGC activation by CRISPR/Cas9-mediated genome editing,[4,5] elicitor application,[6] and manipulation of global or pathway-specific regulators.[7,8] Over the last twenty years, NP research has been revolutionized by the development of computational tools for every aspect of NP discovery, ranging from BGC identification to structure prediction to linking genes to compounds.[9,10]

Machine learning (ML) is a subset of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable computers to learn from data without being explicitly programmed. Over the past few decades, the concept and tools of ML have permeated into various research fields. In NP research, ML tools have played a crucial role in improving our understanding of NPs, including detecting BGCs, predicting chemical structures, and profiling activity, as summarized in a number of recent reviews.[9,11−13] With the exceptional prediction power of these ML tools, it is possible to process a vast amount of genomic and molecular data in a high-throughput manner, which aids in selecting an experimentally feasible set for functional validation. A key to comprehending NP chemistry and biology is by understanding the genomes in which their biosynthetic pathways are encoded. Given the increasing availability of microbial genomes, ML-based genome mining approaches offer a profound opportunity to decipher the genomic language of BGCs and better understand NP chemistry and diversity.[13] Once BGC-derived NP structures are elucidated, various ML tools can provide further information on bioactivity such as antibacterial, anticancer, and anti-inflammatory activity and target prediction as well as other features.[11−13]

The workflow for building an ML model consists of four main parts: dataset preparation, molecular representations and descriptors, model training, and model evaluation (Figure 1). Dataset preparation is crucial to generate a successful ML model. A high-quality NP dataset is a prerequisite and leads to better model performance. Zhang et al. identified specific aspects that need to be considered when preparing a dataset for ML model training, such as balanced positive and negative instances, applicability domain, data consistency, inevitable data errors, and database structure.[11] Featurization plays a crucial role in translating genomic language and chemical

Received: April 16, 2023
Published: August 22, 2023

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

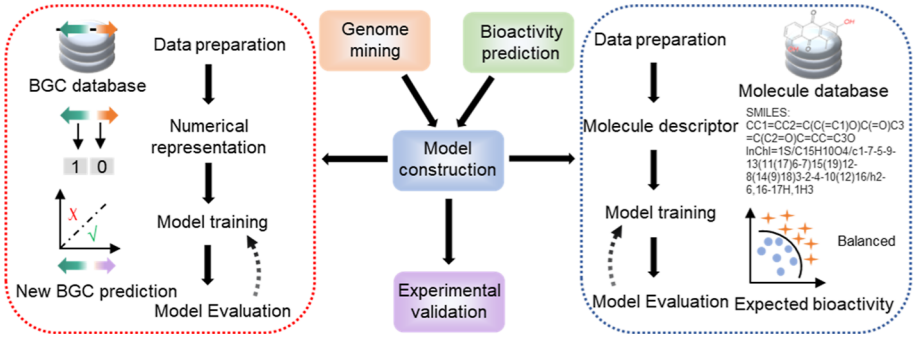


Figure 1. Overview of an ML-enabled workflow for discovery of NPs. The general workflow consists of model construction and experimental validation. Model construction involves four main parts: data preparation, molecular representation and descriptor, model training, and model evaluation. The red frame denotes model construction for BGC prediction. The blue frame represents model construction for bioactivity prediction. BGC: biosynthetic gene cluster; SMILES: simplified input line entry system; InChI: international chemical identifier.

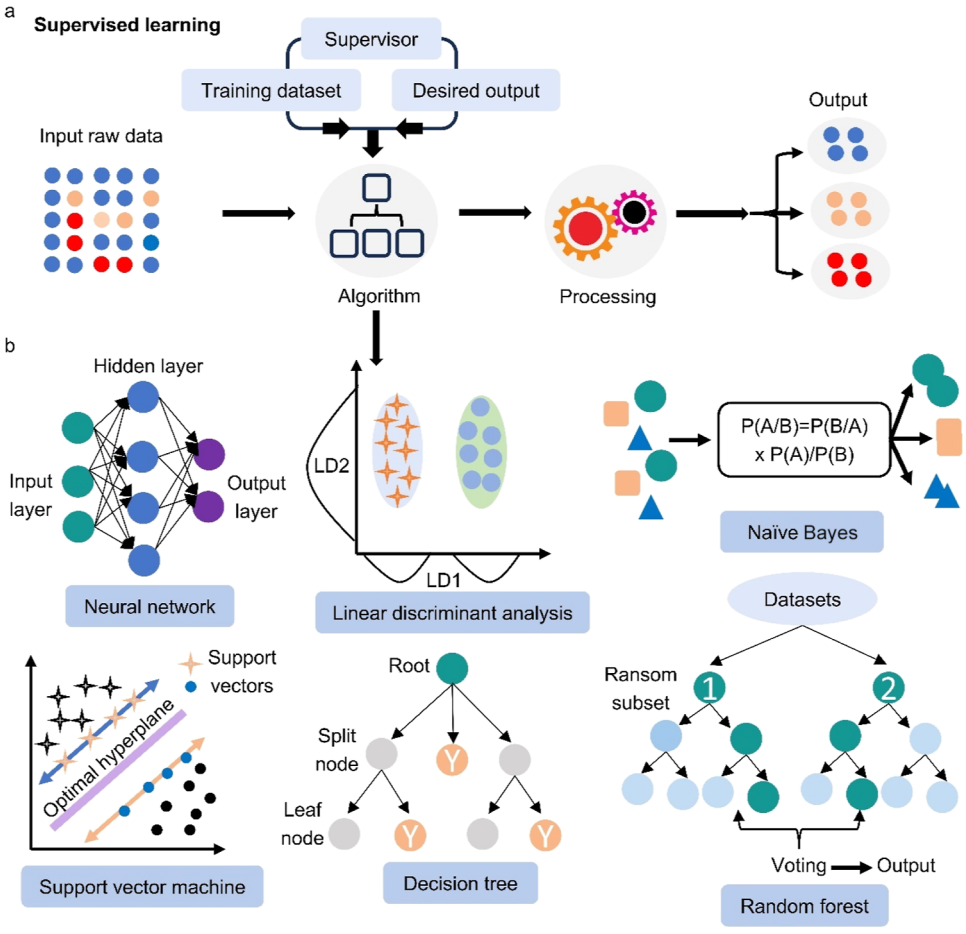


Figure 2. Supervised learning. (a) Basic architecture of supervised learning. (b) Examples present the commonly used supervised algorithms for NP discovery: neural networks, LDA, NB, SVM, DT, and RF.

structure information into computer-readable formats. It is an essential step in modeling and predicting new BGCs as well as the properties of NPs and other compounds. One common example of featurization is the generation of molecular representations and descriptors. These enable the conversion of complex molecular structures into meaningful numerical features that can be utilized in various computational analyses and predictive models. Early molecular representations, such as SMILES (simplified input line entry system),[14] SMARTS (SMILES arbitrary target specification),[15] Daylight sCIS,[16] OpenEye Scientific Software,[17] and InChI (international chemical identifier),[18] were created to store and retrieve molecular information and identify shared molecular features or substructures from databases. Novel molecular representations, such as DeepSMILES[19] and SELFIES,[20] have emerged for practical use in ML tools. Molecular fingerprints, such as ECFP (extended connectivity fingerprints)[21] and MACCS (molecular access system) keys,[22] have been developed for efficient substructure searching in growing chemical databases and reduced storage space. Additionally, unlike chemoinformaticians, computational chemists usually use molecular representations to compute molecular descriptors that describe

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

Table 1. ML Tools for Genome Mining of NPs

name	scope of application	ML algorithms	data source	training dataset size/feature	refs
NRPSpredictor2	predict NRPS adenylation domain specificity	SVM, transductive SVM	manually curated dataset	576 (labeled data for SVM), 5096 (unlabeled data for TSVM); 12 AAindex and z-scales descriptors	34
SANDPUMA	predict NRPS adenylation domain specificity	DT	MIBiG[57] and manually curated dataset	928	31
RiPPMiner	predict RiPP BGC subclasses, the leader cleavage site of precursor peptide, cross-links, and posttranslationally modified residues in the core peptide	SVM, RF	RiPPDB (manually curated database)	513	36
RODEO	identify and rank RiPP precursor peptides belonging to specific subclasses and evaluate the genomic neighborhood	SVM	manually curated dataset	350	38
decRiPPter	predict RiPP precursor peptides in a class-independent manner and identify corresponding BGCs using pan-genomics	SVM	MIBiG2.0[58]	175 (positive dataset), 20,000 (negative dataset)	44
NeuRiPP	identify RiPP precursor peptides belonging to known subclasses	parallel CNN	RiPP-PRISM,[59] Thiofinder,[60] high-confident RODEO predictions	2726 (positive dataset), 19,224 (negative datset); matrix of hot vectors	45
NLPPrecursor	identify RiPP precursor peptides belonging to known subclasses	NLP	RiPPs identified by RiPP-PRISM[59]	~3000 (token vectors of length *bptt*)	46
DeepBGC	identify BGCs for all major NP classes and predict the molecular activity of the NPs	BiLSTM, RNN, skip-gram neural network, RF	ClusterFinder[53] training set	617 (positive dataset), 10,128 (negative dataset); 102-dimensional vectors and two binary flags	51
Deep-BGCpred	identify BGCs for all major NP classes	CNN, stacked BiLSTM, RF	MiBiG1.5, ClusterFinder[53] training set	1984 (positive dataset), 10,128 (negative dataset); pfam2vec embedding vectors and continuous vector representations	52
BiGCARP	identify BGCs for all major NP classes	ESM-1b, BERT, CNN	antiSMASH	127,000 (mask and corrupt tokens)	54
AniAMPpred	identify AMPs from the animal kingdom.	SVM, CNN	NCBI, StarPepDB[61]	16,096 (positive dataset), 15,747 (negative dataset)	48
AMP prediction	identify AMPs from peptide sequence features	LSTM, attention, BERT	ADAM,[62] APD,[63] CAMP,[64] LAMP[65]	10,321 (positive dataset), 3,030,124 (negative dataset); transformer-based bidirectional encoder representations	49

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

the structure and low-dimensional meaningful features of a compound. Model training involves selecting an appropriate ML algorithm for the data and learning task. Supervised algorithms, such as neural networks [e.g., graph neural networks (GNNs), convolutional neural networks (CNNs), and deep neural networks (DNNs)],[23−25] linear discriminant analysis (LDA),[26] naive Bayes (NB),[27] support vector machine (SVM),[28] decision tree (DT),[29] and random forest (RF),[30] are commonly used for NP prediction (Figure 2). The choice of algorithm depends on factors such as data quantity and quality, type of learning task, and interpretability of results. In ML, the ability of a model to make accurate predictions on new, unseen data is referred to as its generalization ability. To evaluate this ability, the dataset is typically split into training data (the portion of the dataset used to train the models), validation data (the subset employed to tune model hyperparameters and compare different models during cross-validation), and testing data (the held-out set utilized to evaluate the final performance of the selected model). The model is trained on the training set, and its performance is evaluated on the testing set using various evaluation metrics depending on the type of problem being solved. Common metrics for classification tasks include accuracy, precision, recall, and F1 score, while regression tasks use mean squared error, mean absolute error, and R-squared. To ensure the model's performance consistency across various dataset partitions, cross-validation is employed. This process includes random partitioning of the dataset into multiple training and validation sets. By utilizing cross-validation, one can effectively compare different models, select the best model and hyperparameters, and subsequently employ a held-out test set to obtain a more accurate measure of the optimal model's real-world performance. This approach enhances the reliability and robustness of the model's evaluation, leading to more meaningful and dependable results in practical applications. A model that performs well on both the testing set and cross-validation is considered to have good generalization ability and can be used to make predictions on new, unseen data.

This review will examine how ML tools have been applied in NP discovery, with a particular focus on how ML tools are leveraged to comprehend the unique "genomic language" that provides insights into NP chemistry. Additionally, we will explore the applications of ML tools in predicting the biological effects of NPs.

■ ML-ASSISTED GENOME MINING OF NPs

NPs are structurally diverse and can be grouped into many classes based on their biosynthetic principles. Numerous genome mining tools have been developed to identify BGCs directly from genome information. Most of them utilized Basic Local Alignment Search Tool (BLAST) or profile-hidden Markov models (pHMMs) to mine signature genes that are responsible for the biosynthesis of a specific class of NPs (e.g., antiSMASH[31] and PRISM[32]) and then determine the boundaries of the BGCs based on a set of predefined rules. Over the years, ML tools have been introduced to genome mining with the goal of discovering new BGCs that may be overlooked by traditional rule-based models. Here, we discuss the ML-based genome mining tools developed for different classes of NPs (Table 1).

Nonribosomally Synthesized Peptides. Nonribosomally synthesized peptides (NRPs) are synthesized by multimodular mega-enzymes named nonribosomal peptide synthetases (NRPSs). Each module minimally consists of three domains: the adenylation domain (A-domain), the peptidyl carrier domain (PCP-domain), and the condensation domain (C), responsible for the recruiting, tethering, and condensation of the substrate into the growing peptide chain.[33] The primary structure of the NRP depends on the sequential order of the modules and domain composition. To aid in the discovery of new NRPs, Rottig et al. developed NRPSpredictor2[34] to predict the specificity of the A-domain to the amino acid substrate. Built on 34 specificity-conferring active site residues in the A-domain, NRPSpredictor2 employs SVM trained on 576 labeled A-domains and transductive SVM trained on 5096 unlabeled A-domains for the prediction of substrate specificity. For bacteria, the predictor can predict both gross physicochemical properties of an A-domain's substrates and detailed single-amino acid substrates. For fungi, the predictor can only predict gross physicochemical properties of substrates due to the lack of sufficient fungal training data. In another study, Blin et al. developed SANDPUMA (Specificity of Adenylation Domain Prediction Using Multiple Algorithms)[31] for ensemble prediction of substrate specificity of the A-domain by using a DT schema that performed individual predictions and combined the results into a single prediction. With an expanded training dataset containing 928 unique A-domain sequences, the ensemble method significantly outperforms individual methods by leveraging the strengths of the active site motif (ASM), SVM, prediCAT (a phylogenetically driven algorithm), and pHMMs.

Ribosomally Synthesized and Posttranslationally Modified Peptides. Ribosomally synthesized and posttranslationally modified peptides (RiPPs) are an emerging class of NPs that are especially attractive for ML-based genome mining efforts due to the relatively small size of RiPP BGCs and the lack of universal signature biosynthetic genes across all RiPP families. Based on the type of posttranslational modification installed on the precursor peptide, RiPPs can be categorized into more than 40 subclasses.[35] In 2017, Agrawal et al. developed RiPPMiner[36] to predict chemical structures and subclasses of RiPPs directly from precursor peptide sequences based on SVM and RF classifiers trained on 513 experimentally characterized RiPPs from 13 RiPP subclasses. RiPPMiner can also predict the leader cleavage site, complex cross-links, and posttranslationally modified residues in the core peptide for the major RiPP subclasses like lanthipeptides, cyanobactins, thiopeptides, and lasso peptides that contain more than 50 entries in the training dataset. An updated version of RiPPMiner called RiPPMiner-Genome[37] can directly take genome sequences as input for automated identification of RiPP BGCs.

In another study, Tietz et al. developed RODEO (Rapid ORF Description and Evaluation Online)[38] for mining RiPP BGCs. Unlike RiPPMiner that uses a whole genome or precursor peptide sequence as input, RODEO uses a single protein of interest as query and captures the neighboring genomic region to predict the function of nearby genes by analyzing their Pfam pHMMs. A tripartite procedure of heuristic scoring, SVM, and motif analysis was then utilized to predict and rank precursor peptides. The RODEO tool first demonstrated its utility by surveying the lasso peptide biosynthetic landscape, revealing over 1400 BGCs and guiding the discovery of five novel lasso peptides. It has been further developed to survey additional RiPP subclasses including thiopeptides,[39] lanthipeptides,[40] linaridins,[41] ranthipeptides,[42] and graspetides.[43]

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

Table 2. ML Tools for Bioactivity Prediction of NPs

name	scope of application	ML algorithms	data source	training dataset size	ref
			Antimicrobial		
KNIME	antimalarial	NB, SMO, RF, VP	ChEMBL, PubChem, literature, thesis	1147	72
	antifungal	ISE	FDA-approved drugs, ADG	3132	69
	anti-MRSA	RF, SVM, GP, CNN	ChEMBL, PubChem, ZINC literature	6645 (for molecular descriptors) 155 (for NMR descriptors)	67
	antimicrobial	ISE	CMC, ADG, literature	3520	68
	antibiotic discovery	DMP-DNN	FDA-approved drugs	2335	70
			Anticancer		
CDRUG	anticancer	RFW, TC, KMM	NCI-60 DTP	18,369	73
	anticancer	DT, SVM, RF, RoF	GDSC, PubChem	8420	75
CDK + PM6 Rf	anticancer, antibiotic	SVM, RF, CT	PubChem, AntiMarin	1746	76
	anticancer	ISE	CMC, NCI drug dictionary	3509	79
	anticancer	Causation analysis	experiment data	28	80
KekuleScope	anticancer	CNN, DNN, RF	ChEMBL	62,981	78
			Anti-Inflammation		
	anti-inflammation	LDA	MicroSource, literature	824	81
	anti-ulcerative colitis	LDA	MicroSource, Sigma-Aldrich databases	53	82
	anti-inflammation	ISE	AnalytiCon Discovery	3333	83
InflamNat	anti-inflammation, compound−target interaction	MTT	literature	1351	84
			Target Prediction		
STarFish	target protein	*k*NN, RF, MLP	AfroCancer, AfroDb, AfroMalaria, Analyti-Con, Carotenoids, ConMedNP, InterBioScreen (IBS), Mitishamba, NANPDB, NP Atlas, NPACT, NPASS, NuBBE, pANAPL, SANCDB, Super Natural II, TCM, TIPdb, UNPD, ZINC ChEMBL	438,258	86
DeepDTA	target protein	CNN	Davis, KIBA	30,056 (from Davis), 118,254 (from KIBA)	87
DeepAffinity	nuclear estrogen receptors, GPCR, ion chanel, receptor tyrosine kinases	GNN, RNN	BindingDB, STITCH, UniRef	489,280	88
DeepConv-DTI	target protein	CNN	DrugBank, KEGG, IUPHAR	48,193	90
	GPCR, ion chanel, transporter, receptor, enzyme, others	SVM	DrugBank	2107 (for GPCR), 502 (for ion chanel), 311 (for transporter), 199 (for receptor), 410 (for enzyme), 83 (for others)	85
DEEPScreen	target protein	CNN	ChEMBL	769,935	91
MolTrans	target protein	CNN	BIOSNAP, DAVIS, BindingDB	27,482 (from BIOSNAP), 11,103 (from DAVIS), 32,601 (from BindingDB)	92
DeepRelations	target protein	GCN, GIN, RNN	Davis, KIBA, PDBbind	25,046 (from Davis), 98,545 (from KIBA), 2921 (from PDBbind)	89
	target PKBβ	QSAR	SWMD	157	93
DeepCYP	target CYP450	DNN	PubChem BioAssay	17,143	94
	target human plasma proteins	RF, BT, MLR, KNN, SVR, MNN	Votano, PKDB, DrugBank	1209	95
	target SITR1	QSAR	PubChem	354	96
	target ERa	NB, RP	BindingDB, DUD-E	6556	97

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

Despite the progress in predicting RiPP BGCs belonging to known subclasses, genome mining of new RiPP subclasses remains a daunting challenge. In 2020, Kloosterman et al. established the Data-driven Exploratory Class-independent RiPP TrackER (decRiPPter)[44] to tackle this challenge by combining a SVM trained on 175 known RiPP precursors to identify candidate precursor genes regardless of RiPP subclasses and pangenomic analyses to identify the corresponding BGCs from those operon-like structures that are sparsely distributed among genomes. Analysis of 1295 *Streptomyces* genomes using decRiPPter led to the discovery of a new lanthipeptide subfamily, serving as an experimental validation of the approach. Geared toward novelty, this approach inevitably suffered from a higher number of false positives compared with the above-mentioned genome mining tools for RiPPs.

In a departure from traditional ML-based tools including SVM and DT classifiers (DT), deep learning-based genome mining methods have also been utilized to identify RiPP precursor peptides with higher accuracy. In 2019, de Los Santos developed a deep neural network (DNN) classifier, NeuRiPP,[45] which was trained on over 9454 peptide sequences for identifying known precursor peptides and new precursor peptide-like sequences, with the best parallel CNN architecture achieving over 99% accuracy. Another tool developed by Merwin et al., called NLPPrecursor,[46] employs natural language processing (NLP)[47] to identify precursor peptides in a class-independent manner but is parameterized for the detection of known RiPP subclasses. NLPPrecursor is a part of DeepRiPP, which also includes two other modules to automate the selective discovery of novel RiPPs. One module is Basic Alignment of Ribosomal Encoded Products Locally (BARLEY), which is used for prioritizing loci that encode novel products by matching the predicted RiPP to a chemical structure database of previously characterized members using a cheminformatic local alignment algorithm. The last module, Computational Library for Analysis of Mass Spectra (CLAMS), automates the identification of the corresponding product in mass spectrometry data by comparative metabolomic analysis. By integrating these three modules, DeepRiPP successfully guided the discovery of three novel RiPPs, including deepstreptin (lasso peptide) and two lanthipeptides, deepflavo and deepginsen.

Antimicrobial Peptides. Beyond RiPPs, the discovery of antimicrobial peptides (AMPs) also greatly benefited from various ML tools. In 2021, Sharma et al. developed AniAMPpred, which utilized a SVM and 1D CNN with Word2vec embedding to identify AMPs from the animal kingdom. Trained on a curated dataset consisting of 10,187 AMPs and 15,747 non-AMPs, the model can confidently classify both AMPs and non-AMPs for diverse peptides of varying lengths with an F1 score of 96% on independent datasets. They further utilized AniAMPpred to identify 436 probable antimicrobial peptides from the genome of *Helobdella robusta* but did not proceed with experimental validation.[48] In a recent work on AMP prediction, Ma et al. combined three NLP models [Long Short-Term Memory (LSTM), Attention, and Bidirectional Encoder Representations from Transformers (BERT)] for mining AMPs from the human gut microbiome.[49] The model performance was superior to that of other available AMP prediction methods using the same test dataset in terms of area under the precision-recall curve (AUPRC) and precision. Experimental results showed that 181 of the 216 identified candidate AMPs showed antimicrobial activity (positive rate of >83%). For a comprehensive review of ML-enabled AMP discovery and design, we refer readers to the review by Yan et al.[50]

Other ML-Based Genome Mining Tools. Compared with ML-based genome mining tools developed for specific classes of NPs, examples of utilizing ML for comprehensive identification of BGCs regardless of NP classes are still limited. In 2019, Hannigan et al. developed DeepBGC,[51] which employed a Bidirectional Long Short-Term Memory (BiLSTM), a recurrent neural network (RNN), and a word2vec-like word embedding skip-gram neural network (pfam2vec) trained with 617 positive and 10,128 negative samples for improved detection of BGCs belonging to known classes and showed great potential for identifying novel BGC classes. DeepBGC was supplemented with an RF classifier that enables accurate classification of BGC product classes and some degree of prediction of the corresponding biological activities. In 2022, Yang et al. reported an improved version of DeepBGC called Deep-BGCpred,[52] which combined the multisource Pfam domain encoder and the stacked BiLSTM model for predicting BGCs with improved accuracy and reduced false-positive rates. Benchmark experiments showed that Deep-BGCpred is superior to the existing NP class-independent genome mining tool, ClusterFinder,[53] which predicts BGCs via pHMMs of a sequence of Pfam annotations. Similar to other supervised algorithms, the performance of DeepBGC and Deep-BGCpred is highly reliant on the quality of the negative examples that should contain no false negatives and display similarities with true BGCs.

In a recent study, Rios-Martinez et al. pioneered the usage of a self-supervised neural network masked language model called BiGCARP[54] that contains the ByteNet encoder-dilated CNN architecture[55] with linear input embedding and output-decoding layers for predicting and classifying BGCs from microbial genomes. Trained on 127,000 BGC sequences represented as ESM-1b-pretrained embeddings of a protein family domain,[56] BiGCARP can capture meaningful patterns in BGCs with area under the receiver operating characteristic (AUROC) scores ranging from 0.936 to 0.950 and outperforms DeepBGC on classifying four out of seven product classes. This results from the relatively large training data used in BiGCARP which is 100 times larger than that used in DeepBGC. However, it is still unclear if BiGCARP can detect some truly novel BGCs that contain noncanonical biosynthetic domains from underrepresented sources.

■ ML-BASED PREDICTION OF NP BIOACTIVITY

NP discovery has been greatly accelerated by the aforementioned antiSMASH, PRISM 4, and emerging ML tools for genome mining. ML has offered a unique opportunity to link molecular structures of NPs with bioactivity. In the context of bioactivity prediction for NPs, several ML tools have been developed for various types of activities such as antimicrobial, anticancer, and anti-inflammation and target prediction. In the following sections, we present examples of ML-assisted bioactivity prediction for NPs, including the ML tools used, data sources, and dataset sizes (Table 2). It is worth noting that these ML tools heavily rely on NP structural information for bioactivity prediction. This may represent a major drawback because they are limited to known NPs that have undergone structural characterization, and obtaining structures for novel NPs can be challenging. However, to address the

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

limitations of these approaches, alternative methods for predicting NP activity from the gene cluster have been reported.[32,51,66] For instance, Skinnider et al. introduced PRISM 4, a comprehensive platform capable of predicting the chemical structures of genomically encoded antibiotics, covering all classes of bacterial antibiotics currently in clinical use. The high accuracy of chemical structure prediction facilitated the development of ML tools to predict the likely biological activity of encoded molecules.[32] To gain a deeper understanding of these studies and the various methods employed, we encourage readers to review the relevant literature thoroughly. Moreover, it is worth mentioning that data sources such as ChEMBL-, PubChem-, and FDA-approved drugs encompass a combination of both NPs and synthetic compounds. In the context of this review, most of the discussed models have been trained on datasets that include both synthetic molecules and NPs. It is important to recognize that synthetic compounds occupy a distinct area of chemical space compared to NPs, which could potentially lead to reduced accuracy when these models are employed to predict the bioactivity of NPs.

Antimicrobial. The use of ML tools in predicting the bioactivity of NPs has gained significant attention in recent years. One area where ML has been extensively employed is in the prediction of antimicrobial activity. In 2018, Dias et al. developed two QSAR (quantitative structure−activity relationship) models, one using molecular descriptor (approach A) and the other using ^{1}D NMR descriptors (approach B), to discover new inhibiting agents against methicillin-resistant *Staphylococcus aureus* (MRSA) infection. They used regression models to predict 6645 molecules retrieved from various databases in approach A, achieving an R^2 of 0.68 and an RMSE of 0.59 for the test set. In approach B, a new NP drug discovery methodology was developed using ^{1}D NMR descriptors, with the best model achieving a prediction accuracy of over 77% for both training and test datasets.[67] Masalha et al. developed an ML tool using the ISE (iterative stochastic elimination) algorithm that efficiently predicts that NPs will assist in the discovery of low-cost antibacterial drugs, achieving an AUC of 0.957 and identifying 72% of the antibacterial drugs in the top 1% of a mixed set of active and inactive substances.[68] In another study, they also used the ISE algorithm to predict NPs for their antifungal activity, resulting in a predictive model with an AUC of 0.89, successfully detecting 42% of the antifungal drugs in the top 1% of the screened chemicals.[69] Unlike QSAR and ISE algorithms, in 2020, Stokes et al. developed a DNN model (Chemprop) to predict molecules with antibacterial activity, identifying a molecule called halicin that demonstrated bactericidal activity against various pathogens in murine models. Additionally, the model identified eight antibacterial compounds that were structurally different from known antibiotics, highlighting its potential for identifying novel antibacterial agents.[70] In 2023, Liu et al. employed the same algorithm (Chemprop) to train with a growth inhibition dataset for *Acinetobacter baumannii*. The authors then conducted in silico predictions for structurally novel molecules targeting *A. baumannii*, which led to the discovery of abaucin, an antibacterial compound exhibiting narrow-spectrum activity against *A. baumannii*. These notable findings showcase the remarkable potential of Chemprop in predicting multiple targets.[71] In addition to using a simplex model, Egieyeh et al. trained four different binary classifiers, NB, RF, sequential minimization optimization (SMO), and Voted Perceptron (VP), on a dataset of NPs with in vitro antimalarial activity and applied their best models against 450 NPs from the InterBioScreen chemical library, achieving consistent antiplasmodial bioactivity class prediction for 54% of the compounds in the NP library.[72]

Anticancer. Several studies have utilized ML tools in anticancer drug discovery to predict the anticancer bioactivity of chemical compounds. Li and Huang developed CDRUG (Cancer Drug), a web server that uses a hybrid score (HSCORE) to predict the anticancer bioactivity of NPs. The model was trained on a dataset of 8565 active compounds and 9804 inactive compounds from the NCI-60 Developmental Therapeutics Program (DTP) project, achieving an AUC of 0.878, indicating its effectiveness in distinguishing active and inactive compounds.[73] Using CDRUG, the group predicted the anticancer bioactivity of 21,334 compounds from 2402 plants from the traditional Chinese medicine database (TCM), with 5278 compounds predicted as anticancer compounds and 346 compounds showing high potency in the 60 cancer cell lines test. Similarity analysis revealed that 75% of the 5278 compounds were highly comparable to approved anticancer drugs.[74] Another study by Yue et al. developed an ML tool to predict the sensitivity of cancer cells to NPs using various cell lines. The study designed DT, SVM, RF, and ROF for anticancer drug response prediction using both genomic characterizations (gene expression) and chemical descriptors. ROF achieved the best performance with an AUC of 0.87 with 10-fold cross-validation, and curcumin and resveratrol were evaluated to validate the model.[75] Pereira et al. utilized a QSAR model to predict the bioactivity of compounds for antitumor and antibiotic activities, identifying 25 and 4 lead compounds for antibiotic and antitumor drug design, respectively, using RF.[76] The study validated the usefulness of quantum-chemical descriptors in discriminating biologically active and inactive compounds, and the predictive performance was better than that of the previous model using only CDK descriptors.[77] Cortés-Ciriano et al. developed the Kekulescope tool, which utilizes the CNN algorithm for drug discovery using high-content screening images or 2D compound representations, demonstrating that in vitro activity of compounds on cancer cell lines and protein targets can be accurately predicted from their Kekulé structure representations alone. The results also showed that including additional fully connected layers in the CNNs increased their predictive power by up to 10%, and averaging the output of RF models and CNNs led to lower errors in prediction for multiple datasets than either model alone.[78] In other studies, Rayan et al. used the ISE algorithm to create a model to predict NPs for their anticancer activity, identifying twelve NPs as potential anticancer drug candidates.[79] Wang et al. employed a causation discovery algorithm that displayed more robust performance than stepwise regression to identify anticancer compounds from *Panax ginseng* extracts, with ginsenoside Rb1 identified as the most active compound.[80]

Anti-inflammation. Anti-inflammatory drugs are known for their undesirable side effects. To tackle this issue, Galvez-Llompart et al. used molecular topology and LDA to develop a topological−mathematical model to identify new anti-inflammatory drugs from NPs. The model was validated externally and led to the discovery of 74 compounds with actual anti-inflammatory activity, 54 of which had been previously described in the literature as anti-inflammatory.[81] In a subsequent study, the same group developed a QSAR model

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

based on molecular topology for predicting the IL-6-mediated (interleukin-6) anti-ulcerative colitis activity of compounds, which led to the discovery of four potentially bioactive compounds: alizarin-3-methylimino-N, *N*-diacetic acid (AMA), Calcein, (+)-dibenzyl-*l*-tartrate (DLT), and Ro 41-0960. In vitro testing on two cell lines demonstrated that three of these compounds were able to significantly reduce IL-6 levels, with Ro 41-0960 showing particular effectiveness. This study demonstrated the effectiveness of molecular topology as a tool for selecting potentially active compounds in the treatment of ulcerative colitis.[82] Separately, Aswad et al. developed a predictive model using the ISE algorithm to identify NPs with potential anti-inflammatory activity. The model was able to differentiate between active and inactive anti-inflammatory molecules and identified ten NPs as anti-inflammatory drug candidates, which highlights the potential of the ISE algorithm in identifying NPs with anti-inflammatory properties.[83]

InflamNat is an online tool which contains a database of 1351 NPs with their physicochemical properties, anti-inflammatory bioactivities, and molecular targets, along with two ML-based predictive tools specifically designed for NPs. The tools use a novel multitokenization transformer model (MTT) as a sequential encoder to predict the anti-inflammatory activity of NPs and the compound−target relationship. The experimental results showed that the proposed predictive tools achieved high accuracy in predicting both anti-inflammatory activity and compound−target interactions, with AUC values of 0.842 and 0.872, respectively. The study demonstrates the urgent need for well-curated databases and user-friendly predictive tools to facilitate NP-inspired drug development.[84]

Target Prediction. Validating the molecular targets of NPs is crucial in identifying potential candidates for NP-based drugs. However, the traditional process of determining compound−target interaction requires extensive in vitro or in vivo experiments. To address this limitation, utilizing ML tools to predict the compound−target interaction can significantly reduce the required effort.

Several ML tools have been developed to predict protein targets of bioactive compounds. Keum et al. used data from the DrugBank database to develop six classification-prediction models for compound−target interactions in humans. Using these models, the study predicted the interactions of compounds from NPs and identified several disease-related proteins, including G-protein-coupled receptors (GPCR), ion channels, enzymes, receptors, and transporters, as potential targets of natural herbal compounds.[85] Similarly, Cockroft et al. developed STarFish, a computational target fishing model that utilized *k*NN, RF, and MLP algorithms to identify protein targets of bioactive compounds by cross-referencing 20 NP databases with the ChEMBL bioactivity database. During cross-validation, the models achieved strong performance with AUROC scores ranging from 0.94 to 0.99 and Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) scores ranging from 0.89 to 0.94, but their performance decreased when tested on the NP dataset. However, the implementation of a model stacking approach significantly improved the performance of predicting protein targets of NPs with increased AUROC and BEDROC scores.[86]

Öztürk et al. proposed a deep learning model that predicted drug−target interaction (DTI) binding affinities by using only sequence information of both targets and drugs, which outperformed existing methods such as KronRLS and SimBoost. Unlike most computational methods that focus on binary classification, the proposed model utilized advanced deep learning algorithms such as CNNs to model protein sequences and compound 1D representations for binding affinity prediction.[87] Karimi et al. used a semisupervised deep learning model that combines recurrent and convolutional neural networks (RNN−CNN) and integrates domain knowledge to predict target selectivity. The model outperformed conventional options in achieving relative error in IC_{50} within 5-fold for test cases and 20-fold for protein classes not included in training,[88] while their subsequent study curated a dataset with both affinities and contacts of compound−protein interactions and assessed the interpretability of various DeepAffinity versions. The model showed generalizability in affinity prediction and superior interpretability, with potential applications in contact-assisted docking, structure-free binding site prediction, and structure−activity relationship studies.[89] Lee et al. developed a deep learning model which is capable of predicting DTIs on a large scale using raw protein sequences, which can handle a variety of protein lengths and target protein classes.[90] In addition, Rifaioglu et al. proposed DEEPScreen, a large-scale DTI prediction system for early-stage drug discovery that employed a deep CNN to learn complex features from readily available 2D structural representations of compounds.[91] Another study by Huang et al. described MolTrans, a deep learning model to improve DTI prediction for in silico drug discovery by incorporating a knowledge-inspired substructural pattern mining algorithm and interaction modeling module, resulting in DTI prediction with increased accuracy and interpretability, as well as utilizing an augmented transformer encoder to better extract and capture semantic relations among substructures from massive unlabeled biomedical data.[92]

In addition, ML tools have been developed for the prediction of specific target proteins such as protein kinase B (PKBβ),[93] cytochrome P450 (CYP450),[94] human plasma proteins,[95] sirtuin 1 (SITR1),[96] and estrogen receptor a (ERa).[97] For instance, Davis and Vasanthi utilized the QSAR model to identify potential anticancer compounds from a seaweed metabolite database. Using a hybrid genetic algorithm and multiple linear regression analysis, they identified molecular descriptors that played a role in anticancer activity, with Baumann's alignment-independent topological descriptors playing a significant role in variation of activity. Subsequently, they performed a docking study of two crystal structures of PKBβ to identify novel ATP-competitive inhibitors of PKBβ, with Callophycin A exhibiting better ligand efficiency than other PKBβ inhibitors. In silico pharmacokinetic and toxicity studies also showed that Callophycin A had a high drug score compared to other inhibitors.[93] Li et al. developed a multi-task DNN model to predict the inhibitive effect of a compound against five major CYP450 isoforms, namely, 1A2, 2C9, 2C19, 2D6, and 3A4. They also built linear regression models to quantify how the other tasks contributed to the prediction difference of a given task between single-task and multi-task models. Furthermore, sensitivity analysis was applied to extract useful knowledge about CYP450 inhibition, which may shed light on the structural features of these isoforms and give hints about how to avoid side effects during drug development.[94] Sun et al. used six ML algorithms and 26 molecular descriptors to develop QSAR models that could predict plasma protein binding

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

Table 3. Glossary of ML Terms

name	abbreviation	feature
support vector machine	SVM	Supervised ML algorithm used for classification, regression, and outlier detection analysis.
natural language processing	NLP	An ML technology that focuses on enabling computers to understand, interpret, and generate human language.
recurrent neural network	RNN	A type of artificial neural network designed to model sequential data by allowing the network to persist with information from previous time steps.
long short-term memory	LSTM	A type of RNN architecture designed to handle the vanishing gradient problem in standard RNNs.
bidirectional long short-term memory	BiLSTM	A variant of the LSTM network that captures the dependencies of a sequence in both forward and backward directions.
convolutional neural network	CNN	A type of neural network designed for image recognition and processing.
bidirectional encoder representations from transformers	BERT	A pretrained natural language processing model using an unsupervised learning approach.
naive Bayes	NB	A probabilistic classification algorithm based on Bayes' theorem, which is commonly used in text classification and spam filtering.
random forest	RF	A type of ensemble ML algorithm that combines multiple DTs to improve the accuracy and robustness of the model.
sequential minimization optimization	SMO	A popular algorithm for solving the optimization problem in SVMs to find the optimal values of the parameters that define the SVM hyperplane.
voted perceptron	VP	A type of Perceptron algorithm that uses multiple weight vectors instead of a single weight vector for binary classification.
iterative stochastic elimination	ISE	A type of wrapper method evaluating different subsets of features by iteratively removing one feature at a time based on their importance, until a desired level of accuracy is achieved.
Gaussian process	GP	A type of nonparametric model that is used to model complex, nonlinear relationships between variables, without making any assumptions about the underlying distribution of the data.
deep neural network	DNN	A type of artificial neural network that is composed of multiple layers of interconnected processing nodes.
directed-message passing deep neural network	DMP-DNN	A type of deep learning architecture that is used for processing and modeling graph-structured data.
classification tree/decision tree	CT/DT	An ML model that is constructed by recursively partitioning the input space into smaller regions and used for classification and regression tasks.
frequency-weighted fingerprint	FWF	A binary vector that encodes the presence or absence of certain chemical substructures in a molecule.
Tanimoto coefficient	TC	A similarity metric used to measure the similarity between two molecular fingerprints.
MinMax kernel	KMM	A type of kernel function and a similarity measure between two data points in a feature space.
rotation rorest	ROF	An ensemble learning method combining multiple DT classifiers into a single model.
linear discriminant analysis	LDA	A supervised learning method that seeks to find a linear combination of features that best separates the classes of a given dataset.
k-nearest neighbor	*k*NN	A nonparametric and simple algorithm that makes predictions based on the similarity between a new data point and its *k* nearest neighbors in the training dataset.
multilayer perceptron/multilayer neural network	MLP/MNN	A type of feedforward artificial neural network composed of multiple layers of interconnected processing nodes that is widely used for supervised learning tasks such as classification, regression, and prediction.
graph neural network	GNN	A type of neural network designed to operate on data structured as graphs and used for tasks such as node classification, link prediction, and graph classification.
graph convolutional network	GCN	A type of GNN that use a convolutional-like operation to aggregate information from neighboring nodes in a graph.
graph isomorphism network	GIN	A type of GNN that consists of multiple graph convolutional layers and aims to address the problem of graph isomorphism.
quantitative structure−activity relationship	QSAR	Using statistical and ML techniques to establish a relationship between a set of molecular descriptors (such as molecular weight, shape, and chemical properties) and the activity or property of interest (such as biological activity, solubility, or toxicity).
boost tree	BT	A type of ensemble learning method for combining multiple weak learners to form a strong learner and used for both regression and classification tasks.
multiple linear regression	MLR	A statistical modeling technique used to analyze the relationship between two or more independent variables and a dependent variable.
support vector regression	SVR	A variation of SVM and used for regression analysis.
recursive partitioning	RP	Involves recursively splitting the data into smaller subsets based on the values of the input variables to create a DT to make predictions and used for classification and regression tasks.
multitokenization transformer	MTT	A type of neural network architecture used in natural language processing tasks, such as language modeling and text classification.

(PPB) fractions of 967 pharmaceuticals. The models demonstrated excellent performance and could be useful for chemists in predicting PPB from molecular structure. Furthermore, the study identified important structural descriptors that contribute to the predictive power of the models, providing guidance for the modification of chemicals.[95] In another application, the QSAR model was used to generate an inhibitor structure pattern for SIRT1, a deacetylase enzyme associated with aging, diabetes, and cancer. The pattern was used for ligand-based virtual screening for over one million active compounds from Chinese herbs, leading to the identification of 12 compounds as SIRT1 inhibitors. Molecular docking software confirmed that three of these compounds had a high affinity for SIRT1.[96] In a separate study, Pang et al. developed two ML models, NB and recursive partitioning (RP), to identify ERα antagonists from an in-house NP library.

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

The models predicted 162 compounds as ER antagonists, which were then evaluated by molecular docking. Eight representative compounds were selected and tested for ERα competitor assay and luciferase reporter gene assay, showing varying levels of antagonistic activity against ERα.[97]

FUTURE PERSPECTIVES

ML has shown valuable potential in NP research, especially in genome mining and scaffold prediction, and predicting properties of NPs, such as drug-likeness, toxicity, and biological activity.[98] However, there are several technical limitations that need to be addressed in order to fully exploit the potential of ML for NPs.[99,100] One of the main limitations is the lack of integrated and standardized NP databases, which can serve as the training data for ML models. The available databases with structure and bioactivity information for NPs (e.g., ChEMBL, PubChem, and ZINC NPs) and databases for BGCs (e.g., antiSMASH, MIBiG, and BiG-FAM) have been extensively reviewed.[9,11] The existing databases are often incomplete, contain errors, and lack standardized annotations, making it difficult to train accurate ML models. The solution to this limitation is to construct high-quality and large-scale NP databases that are standardized and comprehensive, such as the recently launched NPAtlas database.[101] Another limitation is the featurization of NP structures, which involves transforming chemical structures into numerical descriptors that can be used as inputs for ML models.[12] Traditional featurization methods may not capture the unique structural features of NPs, requiring the development of new featurization methods that incorporate the structural diversity and complexity of NPs. An example of such a method is the DeepChem library,[102] which uses deep learning to generate molecular representations that capture 3D structural information. A third limitation is the lack of ML algorithms that can handle small and biased datasets, which are common in NP research.[103] Traditional ML algorithms may not perform well on small datasets or when the classes are imbalanced. To overcome the challenges posed by small and imbalanced datasets in NP discovery, various techniques to enhance the performance of ML models have been proposed, such as data augmentation, transfer learning, contrastive learning, and ensemble methods. By applying these methods, ML models can better handle limited and unevenly distributed data, leading to improved prediction performance on NP discovery.[104,105] Leveraging transfer learning and multitask learning strategies can significantly boost the efficiency and efficacy of ML models for NP discovery. By pretraining models on vast datasets from related domains and subsequently fine-tuning them on smaller NP datasets, the models can adapt and generalize to the specific context of NPs. This approach not only leads to more accurate predictions but also reduces the data requirements for training, making it particularly valuable in scenarios with limited available data. The prospect of detecting NPs with true novelty and accuracy remains a challenge due to the limited and unbalanced training data consisting of canonical BGCs. A possible solution to this limitation is the integration of ML with rule-based models that use predefined rules or logic to make decisions. In the context of imbalanced datasets, combining ML with rule-based models can help improve the performance and generalization of the predictions. This approach could improve the detection of BGCs that deviate significantly from existing biosynthetic schemes. Finally, the integration of ML with other computational approaches, such as molecular docking, molecular dynamics simulations, and quantum chemical calculations, offers a promising direction in NP research. Hybrid models that combine ML with these complementary techniques can provide a more comprehensive understanding of the interactions and activities of NPs.[106] This synergy allows researchers to gain deeper insights into the molecular mechanisms underlying NP actions. Additionally, the use of NLP can improve the efficiency of data extraction from the vast amount of literature on NPs. However, the use of NLPs in NP research is still in its early stage, and there are several challenges to overcome, such as the complexity and variability of natural language and the lack of standardized annotations.[47]

CONCLUSIONS

ML has emerged as a powerful tool for NP discovery, assisting in genome mining and enabling the prediction of bioactivity. This review summarizes the various ML tools utilized in genome mining and bioactivity prediction, along with the associated limitations and potential solutions in the NP research field. Although there are many technical challenges associated with the use of ML tools for NPs, the ongoing development and application of these tools hold immense promise in the discovery of new NPs and understanding of their biological effects.

Table 3 contains a glossary of ML terms.

AUTHOR INFORMATION

Corresponding Author

Huimin Zhao − *Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States; Department of Chemical and Biomolecular Engineering and Departments of Chemistry, Biochemistry, and Bioengineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-9069-6739; Email: zhao5@illinois.edu

Authors

Yujie Yuan − *Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

Chengyou Shi − *Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States; Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00234

Author Contributions

H. Z and Y. Y contributed to conception and critical revision of the article; Y. Y and C. S wrote the initial manuscript; H. Z, Y. Y, and C. S revised and approved the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by an AI Research Institutes program supported by the U.S. National Science Foundation under grant no. 2019897 (H.Z.) and a grant from the National Institutes of Health (AI144967 to H.Z.).

REFERENCES

(1) Newman, D. J.; Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770−803.

(2) Ayikpoe, R. S.; Shi, C.; Battiste, A. J.; Eslami, S. M.; Ramesh, S.; Simon, M. A.; Bothwell, I. R.; Lee, H.; Rice, A. J.; Ren, H.; et al. A scalable platform to discover antimicrobials of ribosomal origin. *Nat. Commun.* **2022**, *13*, 6135.

(3) Yuan, Y.; Cheng, S.; Bian, G.; Yan, P.; Ma, Z.; Dai, W.; Chen, R.; Fu, S.; Huang, H.; Chi, H.; et al. Efficient exploration of terpenoid biosynthetic gene clusters in filamentous fungi. *Nat. Catal.* **2022**, *5*, 277−287.

(4) Zhang, M. M.; Wong, F. T.; Wang, Y.; Luo, S.; Lim, Y. H.; Heng, E.; Yeo, W. L.; Cobb, R. E.; Enghiad, B.; Ang, E. L.; et al. CRISPR−Cas9 strategy for activation of silent Streptomyces biosynthetic gene clusters. *Nat. Chem. Biol.* **2017**, *13*, 607−609.

(5) Culp, E. J.; Yim, G.; Waglechner, N.; Wang, W.; Pawlowski, A. C.; Wright, G. D. Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. *Nat. Biotechnol.* **2019**, *37*, 1149−1154.

(6) Xu, F.; Wu, Y.; Zhang, C.; Davis, K. M.; Moon, K.; Bushin, L. B.; Seyedsayamdost, M. R. A genetics-free method for high-throughput discovery of cryptic microbial metabolites. *Nat. Chem. Biol.* **2019**, *15*, 161−168.

(7) Bok, J. W.; Keller, N. P. LaeA, a regulator of secondary metabolism in Aspergillus spp. *Eukaryot. Cell* **2004**, *3*, 527−535.

(8) Mao, X. M.; Xu, W.; Li, D.; Yin, W. B.; Chooi, Y. H.; Li, Y. Q.; Tang, Y.; Hu, Y. Epigenetic genome mining of an endophytic fungus leads to the pleiotropic biosynthesis of natural products. *Angew. Chem.* **2015**, *127*, 7702−7706.

(9) Hemmerling, F.; Piel, J. Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nat. Rev. Drug Discovery* **2022**, *21*, 359−378.

(10) Ren, H.; Shi, C.; Zhao, H. Computational tools for discovering and engineering natural product biosynthetic pathways. *iScience* **2020**, *23*, 100795.

(11) Zhang, R.; Li, X.; Zhang, X.; Qin, H.; Xiao, W. Machine learning approaches for elucidating the biological effects of natural products. *Nat. Prod. Rep.* **2021**, *38*, 346−361.

(12) Jeon, J.; Kang, S.; Kim, H. U. Predicting biochemical and physiological effects of natural products from molecular structures using machine learning. *Nat. Prod. Rep.* **2021**, *38*, 1954−1966.

(13) Prihoda, D.; Maritz, J. M.; Klempir, O.; Dzamba, D.; Woelk, C. H.; Hazuda, D. J.; Bitton, D. A.; Hannigan, G. D. The application potential of machine learning and genomics for understanding natural product diversity, chemistry, and therapeutic translatability. *Nat. Prod. Rep.* **2021**, *38*, 1100−1108.

(14) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(15) James, C. A. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc., 2004. http://www.daylight.com/dayhtml/doc/theory/theory.toc.html.

(16) Mayhoub, M.; Carter, D. Towards hybrid lighting systems: A review. *Light. Res. Technol.* **2010**, *42*, 51−71.

(17) *OEChem TK*; Openeye Scientific Software Inc.: Santa Fe, NM, USA, 2012.

(18) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminf.* **2015**, *7*, 23−34.

(19) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. **2018**, ChemRxiv:7097960.v1.

(20) Tiidenberg, K.; Gómez Cruz, E. Selfies, image and the re-making of the body. *Body Soc.* **2015**, *21*, 77−102.

(21) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(22) Polton, D. Installation and operational experiences with MACCS (Molecular Access System). *Online Rev.* **1982**, *6*, 235−242.

(23) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Network.* **2009**, *20*, 61−80.

(24) Yamashita, R.; Nishio, M.; Do, R. K. G.; Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into Imag.* **2018**, *9*, 611−629.

(25) Wen, W.; Wu, C.; Wang, Y.; Chen, Y.; Li, H. Learning structured sparsity in deep neural networks. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2016; Vol. *29*.

(26) Balakrishnama, S.; Ganapathiraju, A., Linear Discriminant Analysis-a Brief Tutorial. *Institute for Signal and information Processing*; Mississippi, 1998; Vol. *18*, pp 1−8.

(27) Zhang, H. The optimality of naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*; AAAI Press, 2004; Vol. *2*, p 3.

(28) Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565−1567.

(29) Hyafil, L.; Rivest, R. L. Constructing optimal binary decision trees is NP-complete. *Inf. Process. Lett.* **1976**, *5*, 15−17.

(30) Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.* **2016**, *114*, 24−31.

(31) Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; van Wezel, G. P.; Medema, M. H.; Weber, T. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **2021**, *49*, W29−W35.

(32) Skinnider, M. A.; Johnston, C. W.; Gunabalasingam, M.; Merwin, N. J.; Kieliszek, A. M.; MacLellan, R. J.; Li, H.; Ranieri, M. R. M.; Webster, A. L. H.; Cao, M. P. T.; Pfeifle, A.; Spencer, N.; To, Q. H.; Wallace, D. P.; Dejong, C. A.; Magarvey, N. A. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **2020**, *11*, 6058.

(33) Walsh, C. T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* **2016**, *33*, 127−135.

(34) Rottig, M.; Medema, M. H.; Blin, K.; Weber, T.; Rausch, C.; Kohlbacher, O. NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **2011**, *39*, W362−W367 Web Server issue).

(35) Montalban-Lopez, M.; Scott, T. A.; Ramesh, S.; Rahman, I. R.; van Heel, A. J.; Viel, J. H.; Bandarian, V.; Dittmann, E.; Genilloud, O.; Goto, Y.; Grande Burgos, M. J.; Hill, C.; Kim, S.; Koehnke, J.; Latham, J. A.; Link, A. J.; Martinez, B.; Nair, S. K.; Nicolet, Y.; Rebuffat, S.; Sahl, H. G.; Sareen, D.; Schmidt, E. W.; Schmitt, L.; Severinov, K.; Sussmuth, R. D.; Truman, A. W.; Wang, H.; Weng, J. K.; van Wezel, G. P.; Zhang, Q.; Zhong, J.; Piel, J.; Mitchell, D. A.; Kuipers, O. P.; van der Donk, W. A. New developments in RiPP discovery, enzymology and engineering. *Nat. Prod. Rep.* **2021**, *38*, 130−239.

(36) Agrawal, P.; Khater, S.; Gupta, M.; Sain, N.; Mohanty, D. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. *Nucleic Acids Res.* **2017**, *45*, W80−W88.

(37) Agrawal, P.; Amir, S.; Deepak; Barua, D.; Mohanty, D. RiPPMiner-Genome: A Web Resource for Automated Prediction of Crosslinked Chemical Structures of RiPPs by Genome Mining. *J. Mol. Biol.* **2021**, *433*, 166887.

(38) Tietz, J. I.; Schwalen, C. J.; Patel, P. S.; Maxson, T.; Blair, P. M.; Tai, H. C.; Zakai, U. I.; Mitchell, D. A. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **2017**, *13*, 470−478.

(39) Schwalen, C. J.; Hudson, G. A.; Kille, B.; Mitchell, D. A. Bioinformatic Expansion and Discovery of Thiopeptide Antibiotics. *J. Am. Chem. Soc.* **2018**, *140*, 9494−9501.

(40) Walker, M. C.; Eslami, S. M.; Hetrick, K. J.; Ackenhusen, S. E.; Mitchell, D. A.; van der Donk, W. A. Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genom.* **2020**, *21*, 387.

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

(41) Georgiou, M. A.; Dommaraju, S. R.; Guo, X.; Mast, D. H.; Mitchell, D. A. Bioinformatic and Reactivity-Based Discovery of Linaridins. *ACS Chem. Biol.* **2020**, *15*, 2976−2985.
(42) Hudson, G. A.; Burkhart, B. J.; DiCaprio, A. J.; Schwalen, C. J.; Kille, B.; Pogorelov, T. V.; Mitchell, D. A. Bioinformatic Mapping of Radical S-Adenosylmethionine-Dependent Ribosomally Synthesized and Post-Translationally Modified Peptides Identifies New C alpha, C beta, and C gamma-Linked Thioether-Containing Peptides. *J. Am. Chem. Soc.* **2019**, *141*, 8228−8238.
(43) Ramesh, S.; Guo, X.; DiCaprio, A. J.; De Lio, A. M.; Harris, L. A.; Kille, B. L.; Pogorelov, T. V.; Mitchell, D. A. Bioinformatics-Guided Expansion and Discovery of Graspetides. *ACS Chem. Biol.* **2021**, *16*, 2787−2797.
(44) Kloosterman, A. M.; Cimermancic, P.; Elsayed, S. S.; Du, C.; Hadjithomas, M.; Donia, M. S.; Fischbach, M. A.; van Wezel, G. P.; Medema, M. H. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS Biol.* **2020**, *18*, No. e3001026.
(45) de Los Santos, E. L. C. NeuRiPP: Neural network identification of RiPP precursor peptides. *Sci. Rep.* **2019**, *9*, 13406.
(46) Merwin, N. J.; Mousa, W. K.; Dejong, C. A.; Skinnider, M. A.; Cannon, M. J.; Li, H. X.; Dial, K.; Gunabalasingam, M.; Johnston, C.; Magarvey, N. A. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 371−380.
(47) Nadkarni, P. M.; Ohno-Machado, L.; Chapman, W. W. Natural language processing: an introduction. *J. Am. Med. Inf. Assoc.* **2011**, *18*, 544−551.
(48) Sharma, R.; Shrivastava, S.; Kumar Singh, S.; Kumar, A.; Saxena, S.; Kumar Singh, R. AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom. *Briefings Bioinf.* **2021**, *22*, bbab242.
(49) Ma, Y.; Guo, Z.; Xia, B.; Zhang, Y.; Liu, X.; Yu, Y.; Tang, N.; Tong, X.; Wang, M.; Ye, X.; et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **2022**, *40*, 921−931.
(50) Yan, J.; Cai, J.; Zhang, B.; Wang, Y.; Wong, D. F.; Siu, S. W. Recent Progress in the Discovery and Design of Antimicrobial Peptides Using Traditional Machine Learning and Deep Learning. *Antibiotics* **2022**, *11*, 1451.
(51) Hannigan, G. D.; Prihoda, D.; Palicka, A.; Soukup, J.; Klempir, O.; Rampula, L.; Durcak, J.; Wurst, M.; Kotowski, J.; Chang, D.; Wang, R. R.; Piizzi, G.; Temesi, G.; Hazuda, D. J.; Woelk, C. H.; Bitton, D. A. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **2019**, *47*, No. e110.
(52) Yang, Z.; Liao, B.; Hsieh, C.; Han, C.; Fang, L.; Zhang, S., Deep-BGCpred: A unified deep learning genome-mining framework for biosynthetic gene cluster prediction. **2021**, bioRxiv:2021.11.15.468547.
(53) Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Wieland Brown, L.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Linington, R. G.; Fischbach, M. A. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. *Cell* **2014**, *158*, 412−421.
(54) Rios-Martinez, C.; Bhattacharya, N.; Amini, A. P.; Crawford, L.; Yang, K. K. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. **2022**, bioRxiv:2022.07.22.500861.
(55) Kalchbrenner, N.; Espeholt, L.; Simonyan, K.; van den Oord, A.; Graves, A.; Kavukcuoglu, K., Neural Machine Translation in Linear Time. **2016**, arXiv:1610.10099.
(56) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.
(57) Medema, M. H.; Kottmann, R.; Yilmaz, P.; Cummings, M.; Biggins, J. B.; Blin, K.; de Bruijn, I.; Chooi, Y. H.; Claesen, J.; Coates, R. C.; Cruz-Morales, P.; Duddela, S.; Dusterhus, S.; Edwards, D. J.; Fewer, D. P.; Garg, N.; Geiger, C.; Gomez-Escribano, J. P.; Greule, A.; Hadjithomas, M.; Haines, A. S.; Helfrich, E. J. N.; Hillwig, M. L.; Ishida, K.; Jones, A. C.; Jones, C. S.; Jungmann, K.; Kegler, C.; Kim, H. U.; Kotter, P.; Krug, D.; Masschelein, J.; Melnik, A. V.; Mantovani, S. M.; Monroe, E. A.; Moore, M.; Moss, N.; Nutzmann, H. W.; Pan, G. H.; Pati, A.; Petras, D.; Reen, F. J.; Rosconi, F.; Rui, Z.; Tian, Z. H.; Tobias, N. J.; Tsunematsu, Y.; Wiemann, P.; Wyckoff, E.; Yan, X. H.; Yim, G.; Yu, F. G.; Xie, Y. C.; Aigle, B.; Apel, A. K.; Balibar, C. J.; Balskus, E. P.; Barona-Gomez, F.; Bechthold, A.; Bode, H. B.; Borriss, R.; Brady, S. F.; Brakhage, A. A.; Caffrey, P.; Cheng, Y. Q.; Clardy, J.; Cox, R. J.; De Mot, R.; Donadio, S.; Donia, M. S.; van der Donk, W. A.; Dorrestein, P. C.; Doyle, S.; Driessen, A. J. M.; Ehling-Schulz, M.; Entian, K. D.; Fischbach, M. A.; Gerwick, L.; Gerwick, W. H.; Gross, H.; Gust, B.; Hertweck, C.; Hofte, M.; Jensen, S. E.; Ju, J. H.; Katz, L.; Kaysser, L.; Klassen, J. L.; Keller, N. P.; Kormanec, J.; Kuipers, O. P.; Kuzuyama, T.; Kyrpides, N. C.; Kwon, H. J.; Lautru, S.; Lavigne, R.; Lee, C. Y.; Linquan, B.; Liu, X. Y.; Liu, W.; Luzhetskyy, A.; Mahmud, T.; Mast, Y.; Mendez, C.; Metsa-Ketela, M.; Micklefield, J.; Mitchell, D. A.; Moore, B. S.; Moreira, L. M.; Muller, R.; Neilan, B. A.; Nett, M.; Nielsen, J.; O'Gara, F.; Oikawa, H.; Osbourn, A.; Osburne, M. S.; Ostash, B.; Payne, S. M.; Pernodet, J. L.; Petricek, M.; Piel, J.; Ploux, O.; Raaijmakers, J. M.; Salas, J. A.; Schmitt, E. K.; Scott, B.; Seipke, R. F.; Shen, B.; Sherman, D. H.; Sivonen, K.; Smanski, M. J.; Sosio, M.; Stegmann, E.; Sussmuth, R. D.; Tahlan, K.; Thomas, C. M.; Tang, Y.; Truman, A. W.; Viaud, M.; Walton, J. D.; Walsh, C. T.; Weber, T.; van Wezel, G. P.; Wilkinson, B.; Willey, J. M.; Wohlleben, W.; Wright, G. D.; Ziemert, N.; Zhang, C. S.; Zotchev, S. B.; Breitling, R.; Takano, E.; Glockner, F. O. Minimum Information about a Biosynthetic Gene cluster. *Nat. Chem. Biol.* **2015**, *11*, 625−631.
(58) Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Munoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Duran, H. G. S.; Andreu, V. P.; Selem-Mojica, N.; Alanjary, M.; Robinson, S. L.; Lund, G.; Epstein, S. C.; Sisto, A. C.; Charkoudian, L.; Collemare, J.; Linington, R. G.; Weber, T.; Medema, M. H. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **2020**, *48*, D454−D458.
(59) Skinnider, M. A.; Johnston, C. W.; Edgar, R. E.; Dejong, C. A.; Merwin, N. J.; Rees, P. N.; Magarvey, N. A. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, E6343−E6351.
(60) Li, J.; Qu, X. D.; He, X. Y.; Duan, L.; Wu, G. J.; Bi, D. X.; Deng, Z. X.; Liu, W.; Ou, H. Y. ThioFinder: A Web-Based Tool for the Identification of Thiopeptide Gene Clusters in DNA Sequences. *PLoS One* **2012**, *7*, No. e45878.
(61) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; García-Jacas, C. R.; Chavez, E.; Beltran, J. A.; Guillen-Ramirez, H. A.; Brizuela, C. A. Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Sci. Rep.* **2020**, *10*, 18074.
(62) Veltri, D.; Kamath, U.; Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **2018**, *34*, 2740−2747.
(63) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087−D1093.
(64) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. $CAMP_{R3}$: a database on sequences, structures and signatures of antimicrobial peptides: Table 1. *Nucleic Acids Res.* **2016**, *44*, D1094−D1097.
(65) Zhao, X.; Wu, H.; Lu, H.; Li, G.; Huang, Q. LAMP: a database linking antimicrobial peptides. *PLoS One* **2013**, *8*, No. e66557.
(66) Walker, A. S.; Clardy, J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.* **2021**, *61*, 2560−2571.
(67) Dias, T.; Gaudêncio, S. P.; Pereira, F. A computer-driven approach to discover natural product leads for methicillin-resistant Staphylococcus aureus infection therapy. *Mar. Drugs* **2018**, *17*, 16.

https://doi.org/10.1021/acssynbio.3c00234
ACS Synth. Biol. 2023, 12, 2650−2662

(68) Masalha, M.; Rayan, M.; Adawi, A.; Abdallah, Z.; Rayan, A. Capturing antibacterial natural products with in silico techniques. *Mol. Med. Rep.* **2018**, *18*, 763−770.
(69) Rayan, M.; Abdallah, Z.; Abu-Lafi, S.; Masalha, M.; Rayan, A. Indexing natural products for their antifungal activity by filters-based approach: Disclosure of discriminative properties. *Curr. Comput.-Aided Drug Des.* **2019**, *15*, 235−242.
(70) 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.; et al. A deep learning approach to antibiotic discovery. *Cell* **2020**, *180*, 688−702.e13.
(71) Liu, G.; Catacutan, D. B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J. C.; Chiappino-Pepe, A.; Syed, S. A.; Fragis, M.; Rachwalski, K.; et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* **2023**, DOI: 10.1038/s41589-023-01349-8.
(72) Egieyeh, S.; Syce, J.; Malan, S. F.; Christoffels, A. Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. *PLoS One* **2018**, *13*, No. e0204644.
(73) Li, G.-H.; Huang, J.-F. CDRUG: a web server for predicting anticancer activity of chemical compounds. *Bioinformatics* **2012**, *28*, 3334−3335.
(74) Dai, S.-X.; Li, W.-X.; Han, F.-F.; Guo, Y.-C.; Zheng, J.-J.; Liu, J.-Q.; Wang, Q.; Gao, Y.-D.; Li, G.-H.; Huang, J.-F. In silico identification of anti-cancer compounds and plants from traditional Chinese medicine database. *Sci. Rep.* **2016**, *6*, 25462−25511.
(75) Yue, Z.; Zhang, W.; Lu, Y.; Yang, Q.; Ding, Q.; Xia, J.; Chen, Y. Prediction of cancer cell sensitivity to natural products based on genomic and chemical properties. *PeerJ* **2015**, *3*, No. e1425.
(76) Pereira, F.; Latino, D. A.; Gaudêncio, S. P. QSAR-assisted virtual screening of lead-like molecules from marine and microbial natural sources for antitumor and antibiotic drug discovery. *Molecules* **2015**, *20*, 4848−4873.
(77) Pereira, F.; Latino, D. A.; Gaudêncio, S. P. A chemoinformatics approach to the discovery of lead-like molecules from marine and microbial sources en route to antitumor and antibiotic drugs. *Mar. Drugs* **2014**, *12*, 757−778.
(78) Cortés-Ciriano, I.; Bender, A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminf.* **2019**, *11*, 41.
(79) Rayan, A.; Raiyn, J.; Falah, M. Nature is the best source of anticancer drugs: Indexing natural products for their anticancer bioactivity. *PLoS One* **2017**, *12*, No. e0187925.
(80) Wang, Y.; Jin, Y.; Zhou, C.; Qu, H.; Cheng, Y. Discovering active compounds from mixture of natural products by data mining approach. *Med. Biol. Eng. Comput.* **2008**, *46*, 605−611.
(81) Galvez-Llompart, M.; Zanni, R.; García-Domenech, R. Modeling natural anti-inflammatory compounds by molecular topology. *Int. J. Mol. Sci.* **2011**, *12*, 9481−9503.
(82) Galvez-Llompart, M.; del Carmen Recio Iglesias, M.; Gálvez, J.; García-Domenech, R. Novel potential agents for ulcerative colitis by molecular topology: suppression of IL-6 production in Caco-2 and RAW 264.7 cell lines. *Mol. Divers.* **2013**, *17*, 573−593.
(83) Aswad, M.; Rayan, M.; Abu-Lafi, S.; Falah, M.; Raiyn, J.; Abdallah, Z.; Rayan, A. Nature is the best source of anti-inflammatory drugs: Indexing natural products for their anti-inflammatory bioactivity. *Inflammation Res.* **2018**, *67*, 67−75.
(84) Zhang, R.; Ren, S.; Dai, Q.; Shen, T.; Li, X.; Li, J.; Xiao, W. InflamNat: web-based database and predictor of anti-inflammatory natural products. *J. Cheminf.* **2022**, *14*, 30.
(85) Keum, J.; Yoo, S.; Lee, D.; Nam, H. Prediction of compound-target interactions of natural products using large-scale drug and protein information. *BMC Bioinf.* **2016**, *17*, 219−425.
(86) Cockroft, N. T.; Cheng, X.; Fuchs, J. R. STarFish: a stacked ensemble target fishing approach and its application to natural products. *J. Chem. Inf. Model.* **2019**, *59*, 4906−4920.
(87) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug−target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821−i829.
(88) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. DeepAffinity: interpretable deep learning of compound−protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **2019**, *35*, 3329−3338.
(89) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. Explainable deep relational networks for predicting compound−protein affinities and contacts. *J. Chem. Inf. Model.* **2020**, *61*, 46−66.
(90) Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* **2019**, *15*, No. e1007129.
(91) Rifaioglu, A. S.; Nalbat, E.; Atalay, V.; Martin, M. J.; Cetin-Atalay, R.; Doğan, T. DEEPScreen: high performance drug−target interaction prediction with convolutional neural networks using 2-D structural compound representations. *Chem. Sci.* **2020**, *11*, 2531−2557.
(92) Huang, K.; Xiao, C.; Glass, L. M.; Sun, J. MolTrans: molecular interaction transformer for drug−target interaction prediction. *Bioinformatics* **2021**, *37*, 830−836.
(93) Davis, G. D. J.; Vasanthi, A. H. R. QSAR based docking studies of marine algal anticancer compounds as inhibitors of protein kinase B (PKBβ). *Eur. J. Pharm. Sci.* **2015**, *76*, 110−118.
(94) Li, X.; Xu, Y.; Lai, L.; Pei, J. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Mol. Pharm.* **2018**, *15*, 4336−4345.
(95) Sun, L.; Yang, H.; Li, J.; Wang, T.; Li, W.; Liu, G.; Tang, Y. In silico prediction of compounds binding to human plasma proteins by QSAR models. *ChemMedChem* **2018**, *13*, 572−581.
(96) Sun, Y.; Zhou, H.; Zhu, H.; Leung, S.-w. Ligand-based virtual screening and inductive learning for identification of SIRT1 inhibitors in natural products. *Sci. Rep.* **2016**, *6*, 19312−19410.
(97) Pang, X.; Fu, W.; Wang, J.; Kang, D.; Xu, L.; Zhao, Y.; Liu, A.-L.; Du, G.-H. Identification of Estrogen Receptor α Antagonists from Natural Products via *In Vitro* and *In Silico* Approaches. *Oxid. Med. Cell. Longev.* **2018**, *2018*, 1−11.
(98) Saldívar-González, F.; Aldas-Bulos, V.; Medina-Franco, J.; Plisson, F. Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* **2022**, *13*, 1526−1546.
(99) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40−55.
(100) Sapoval, N.; Aghazadeh, A.; Nute, M. G.; Antunes, D. A.; Balaji, A.; Baraniuk, R.; Barberan, C.; Dannenfelser, R.; Dun, C.; Edrisi, M.; et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* **2022**, *13*, 1728.
(101) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; et al. The Natural Products Atlas 2.0: A database of microbially-derived natural products. *Nucleic Acids Res.* **2022**, *50*, D1317−D1323.
(102) Ramsundar, B.; Pande, V.; Eastman, P.; Feinberg, E.; Gomes, J.; Leswing, K.; Pappu, A.; Wu, M. *Democratizing Deep-Learning for Drug Discovery, Quantum Chemistry, Materials Science and Biology*; GitHub repository, 2016.
(103) Kaur, H.; Pannu, H. S.; Malhi, A. K. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* **2019**, *52*, 1−36.
(104) Yu, T.; Cui, H.; Li, J. C.; Luo, Y.; Jiang, G.; Zhao, H. Enzyme function prediction using contrastive learning. *Science* **2023**, *379*, 1358−1363.
(105) Yu, T.; Boob, A. G.; Volk, M. J.; Liu, X.; Cui, H.; Zhao, H. Machine learning-enabled retrobiosynthesis of molecules. *Nat. Catal.* **2023**, *6*, 137−151.
(106) Tsai, C.-F.; Chen, M.-L. Credit rating by hybrid machine learning techniques. *Appl. Soft Comput.* **2010**, *10*, 374−380.