Critical Reviews in Biotechnology

ISSN: 0738-8551 (Print) 1549-7801 (Online) Journal homepage: www.tandfonline.com/journals/ibty20

Advances in AI-based strategies and tools to facilitate natural product and drug development

Buddha Bahadur Basnet, Zhen-Yi Zhou, Bin Wei & Hong Wang

To cite this article: Buddha Bahadur Basnet, Zhen-Yi Zhou, Bin Wei & Hong Wang (2025) Advances in AI-based strategies and tools to facilitate natural product and drug development, Critical Reviews in Biotechnology, 45:7, 1527-1558, DOI: 10.1080/07388551.2025.2478094

To link to this article: https://doi.org/10.1080/07388551.2025.2478094

CRITICAL REVIEWS IN BIOTECHNOLOGY
2025, VOL. 45, NO. 7, 1527–1558
https://doi.org/10.1080/07388551.2025.2478094

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

Advances in AI-based strategies and tools to facilitate natural product and drug development

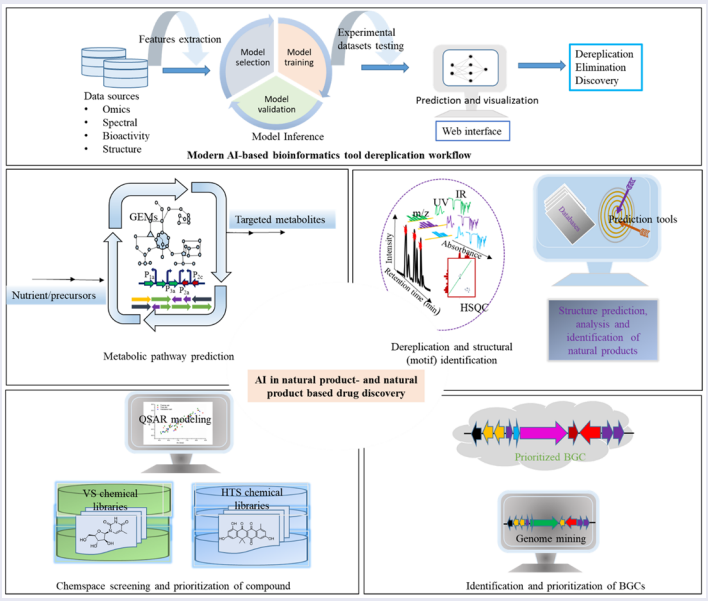
Buddha Bahadur Basnet[a,b], Zhen-Yi Zhou[a], Bin Wei[a], and Hong Wang[a,c]

[a]College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou, China; [b]Central Department of Biotechnology, Tribhuvan University, Kathmandu, Nepal; [c]Key Laboratory of Marine Fishery Resources Exploitment, Utilization of Zhejiang Province, Zhejiang University of Technology, Hangzhou, China

ABSTRACT

Natural products and their derivatives have been important for treating diseases in humans, animals, and plants. However, discovering new structures from natural sources is still challenging. In recent years, artificial intelligence (AI) has greatly aided the discovery and development of natural products and drugs. AI facilitates to: connect genetic data to chemical structures or vice-versa, repurpose known natural products, predict metabolic pathways, and design and optimize metabolites biosynthesis. More recently, the emergence and improvement in neural networks such as deep learning and ensemble automated web based bioinformatics platforms have sped up the discovery process. Meanwhile, AI also improves the identification and structure elucidation of unknown compounds from raw data like mass spectrometry and nuclear magnetic resonance. This article reviews these AI-driven methods and tools, highlighting their practical applications and guide for efficient natural product discovery and drug development.

GRAPHICAL ABSTRACT



ARTICLE HISTORY
Received 20 October 2024
Revised 11 February 2025
Accepted 16 February 2025

KEYWORDS
Natural products; natural product drug discovery; artificial intelligence; dereplication; (bio/chemo) informatics tools; metabolomics

Introduction

Natural products (NPs), produced by: bacteria, fungi, plants, and animals, encompass a diverse array of specialized metabolites, such as: peptides, polyketides, saccharides, terpenes, and alkaloids [1]. These compounds play crucial roles in inter-organismal interactions, acting as: signals, weapons, nutrient-scavenging agents, and stress protectants [2]. Historically, they

CONTACT Hong Wang hongw@zjut.edu.cn; Bin Wei binwei@zjut.edu.cn College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou, China

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07388551.2025.2478094.

have been successfully utilized as: antibiotics, chemotherapeutics, immunosuppressants, and crop protection agents, due to their natural origins [3]. In addition, NPs and their structural analogues have significantly contributed to pharmacotherapy, particularly for cancer and infectious diseases [4]. However, challenges in drug discovery, including: technical barriers to screening, isolation, characterization, and optimization, led to a decline in their pursuit by the pharmaceutical industry from the 1990s onwards [1,4,5]. Recent technological and scientific advancements, such as improved analytical tools, genome mining and engineering strategies, and microbial culturing advances, are overcoming these challenges and opening new opportunities, including the artificial intelligence (AI) incorporated strategies and tools.

Considerable benefits of NPs over synthetic molecules (e.g., diverse chemical scaffolds and biological friendliness) captivate both laboratory researcher and computer scientists [6]. Numerous researchers have created computational methods to assist the swift discovery and structural elucidation of NPs, as well as to aid molecular patterns for combinatorial design or target selectivity [7]. Informatics (chemo/bio), and related disciplines have significantly contributed to NP-based drug discovery, with their successes and limitations [8]. Computational platform, including AI approaches such as machine learning (ML) and deep learning (DL), have gradually integrated into natural product discovery and drug development ((NP) D/DD) tools [5]. Considering the growing need for innovative tools to decipher NPs by utilization the ever-expanding omics datasets, this review presents the latest AI-based tools and strategies in (NP) D/DD. It aims to help natural product scientists quickly prioritize AI-powered tools for the prompt identification of NPs and explore their potential for drug development. Unlike previous studies, this review provides a comprehensive overview of these tools and techniques, including advantage and limitation. By integrating modern AI-coupled dereplication tools and highlighting their applications, it uncovers previously unreported aspects of AI in (NP) D/DD, offering a fresh perspective and paving the way for future research.

Artificial intelligence in natural product discovery and drug development tools

Filtering and prioritizing NPs from vast chemical spaces and -omics compendium, current and next generation key challenges in natural and synthetic-based computational aided drug design (CADD) necessitate the development of high throughput approaches. In the past few years, AI-based automation strategies have offered significant possibilities in the (NP) D/DD, particularly through medium to high throughput screening, thereby expanding the chemical space of natural products libraries [5], to some extent. Since AI's inception in 1956 [9], it has gained remarkable attention and exciting successes, especially with the introduction of technologies like: CADD, quantitative structure-activity relationship (QSAR) [8], support vector machines (SVMs) [10], random forests (RFs) and long short-term memory (LSTM) [11] concepts in the 1990s. These innovations have rigorously expanded AI's applicability in (NP) D/DD [12].

DL and ML, both subfields of AI, advantage the "omics" and (bio/chemo) informatics technologies to model complex nonlinear relationships, pattern recognition and feature extraction from low-level data representations via various molecular graphs [13] or artificial neural architectures. These advances rely theoretically on the diverse array of fundamental mathematical algorithms paradigms, including optimization techniques statistical models, graph theory. Typically, they are implemented using supervised machine algorithms, as depicted in Table 1 and Figure 1(b), which act as single to multi-layer hidden architectures [14]. More specifically, DL represents the most recent neural network architecture, harboring multiple layers, usually three or more layers, as a hidden layer [15]. The first reported application of DL in metabolomics profiling dates back to 1992 [16].

In last couple of years, an array of automatic prediction or networking softwares have been implemented in (NP) D/DD, spanning from metabolite isolation to drug development [17]. These systems incorporate algorithms individually, or in combination, forming artificial network linking computing elements [18], ensembles of multiple algorithms boosting the accuracy and efficacy of the predictions [19]. Intriguingly, AI systems rapidly triggered attraction for applicability in (NP) D/DD domains, such as: surging and exploring the chemspace, repurposing of existing knowledge of molecules and target prediction, structural elucidation and *de novo* synthesis, dereplication and biological profiling of chemical entities, association of BGCs from genome data to chemical structure or vice versa, and metabolic engineering and pathway characterization (Figures 1 and 2).

Association of the BGCs from genomic data to chemical structure or vice versa

Since the early 2000s, next generation sequencing and synthetic biology advanced our understanding of NPs biosynthetic logic and genetic circuits [20]. Meanwhile, the renaissance of modern AI enhanced tools (Figure

Table 1. Supervised machine-learning algorithms in natural products discovery tools.

Approaches	Algorithms	Tools	Reference
Classification approach	K-nearest neighbors	DTi2Vec	[99]
	Naive Bayesian	Discovery of VEGFR2 and BuChE inhibitors	[295,296]
	Support vector networks	QSAR models	[297]
Ensemble approach	Boosting machine	XGBoost classifier	[298]
	Random forest	DEcRyPT	[107]
		QSAR Modeling	[299]
Regression model	Regression (Logistic, multiple, linear, lasso)	STarFish (logistic regression)	[98]
Neural and deep neural algorithm	Multilayer perceptron (MLP)	Prediction and optimization of the andrographolide content in *Andrographis paniculata*	[300]
	Long short-term memory network (LSTM)	HypoRiPPAtlas	[189]
		NPs-like molecules database	[301]
		MetaboListem and TABoLiSTM	[302]
	Convolutional neural networks (CNN)	NMR-based SMART 2.0	[303]
		BiGCARP	[304]
	Graph neural network (GNN)	MSGNN-DTA	[305]
		DTI-HETA	[306]
	Recurrent neural network (RNN)	de novo ligand generation	[307]
		QBMG	[308]
	Generative adversarial network (GAN)	Mol-CycleGAN	[309]
		ReLeaSE	[310]
	Deep Boltzmann machine	DTI prediction	[311]
	Deep belief networks	LBVS	[312]
		de Novo Drug Discovery	[313]
	Variational auto encoders	New chemical reaction generation	[314]
	Dense DNN	CANOPUS	[201]
		NP-VAE	[315]
	Adversarial auto encoders (AAR)	PCM-AAE	[316]
	Feedforward Neural Networks (FNN)	NPClassifier	[203]
	Transformer encoder-decoder neural network	MSNovelist	[197]

1(a); Tables 1 and 2) compiled the rapid accumulation of genomic information and excavated previously undiscovered genes or gene clusters, encoding the new scaffolds or biologically active features. For example, a combined computational pipeline of antiSMASH, the Minimum Information about a Biosynthetic Gene cluster (MIBiG), Big-SCAPE and CoRASON provide identification, compare and correlate the biosynthetic information with secondary metabolites in databases, and link evolutionary and maps phylogenetically [21]. These feature- or correlation-based hooking approaches mine genome based on potential target, compounds family, or bioactive features [22]. Historically and hitherto, Hidden Markov Model-based selection approaches (e.g., SMURF [23], antiSMASH [24], CO-OCCUR [25]), top-notch curated tools, identify and annotate BGCs and chemical entities per unprecedented speed and scale; nevertheless, flaws like false positive rates and inability to predict the novel BGCs, metabolic pathways, or structures still exist on the counterparts. Over the last decades, the renaissance of ML algorithms that incorporated scoring functions and pattern-recognition approaches accelerated the precision of identifying new chemical entities, metabolic pathways, or BGCs by recognizing unique features from the BGCs (e.g., NPlinker) [26], novel enzyme encoding BGCs domains (e.g., EFI-CGFP) [27] and biochemical predictions for poorly annotated genes (e.g., MAGI) [28]. Feature-based tools like: BAGEL4 [29], SANDPUMA [30], GNP [31], iSNAP [32] and MetaMiner [33] are crucial for predicting specific metabolite classes, such as: bacteriocin, nonribosomal peptides, polyketides, and ribosomally synthesized and post-translationally modified peptides. Meanwhile ML-based tools RODEO [34], GECCO [35], RiPPER [36] and RiPPMiner-Genome [37] identify unique superclusters or multi-precursor peptide RiPP BGCs based on special enzyme features.

Despite the limitations, such as: targeting single NP classes, requiring prior knowledge of known sequences, modifications, or core enzymatic machinery, DL tools such as DeepRiPP [38], RFs-based DeepBGC [39], ANN-based SanntiS [40], SVM-based NeuRiPP [41], and SVM integrated tool decRiPPter [42] and pHMM-based RRE-Finder [43] have introduced new possibilities. These tools surpass the performance of traditional ML methods in terms of: features extraction, pattern recognition, scalability, and precision. For example, DeepRiPP and NeuRiPP can identify complex patterns and relationships in data that traditional ML methods with more accurate predictions and annotations of NPs. Furthermore, recent updates on ML tools such as (e.g., PRISM-4 [44], antiSMASH 7.0 [45], DDAP [46], AdenPredictor [47] and PKSpop [48]) have greatly improved the prediction accuracy of chemical structures from genome sequences. For instance, the original version of antiSMASH, released in 2011, could identify a few classes of NPs [24]. The latest version, antiSMASH 7.0, released in 2023, adds more features, including the prediction of chemical structure classes, visualization of enzymatic assembly lines, and regulation of gene clusters [45]. Few important missing pieces, for example, systematic, automatic and extra large-scale (meta) genome mining for recognition and classification of BGCs that generate unfamiliar molecules remain underappreciated. To mitigate such problems, tools that deal with the massive datasets size and implement correlation-based similarity-networking

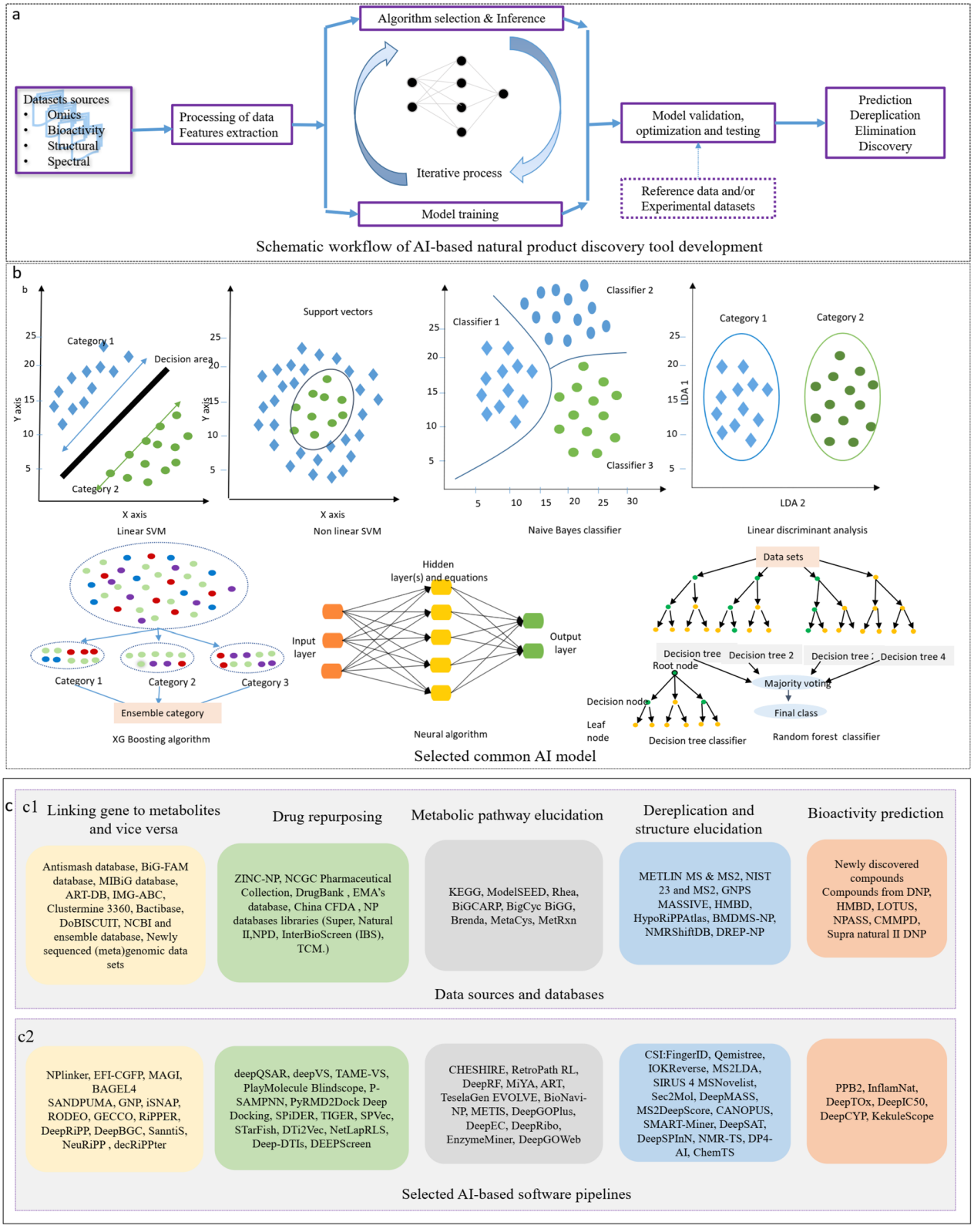


Figure 1. The game-changing potential of artificial intelligence, including deep learning tools, in different aspects of natural product discovery (NPD). (a) Schematic summary of AI-based tools workflow. (b) Common supervised learning algorithm frequently integrated in NPD tools. (c) Data sources and databases for AI model development in major five different areas of NPD; namely, linking gene to metabolites and vice versa, drug repurposing, metabolic pathway elucidation, dereplication and structure elucidation and bioactivity prediction. (c1) Selected data sources. (c2) Modern selected AI-based tools.

Table 2. Advanced AI-based natural product depository and annotation tools (2014–2023).

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
SNAP-MS	Groups chemical similarities in the Natural Products Atlas with mass spectrometry features from molecular networking, using reference spectra and an in-house microbial extract library to accurately predict compound families	Decision trees, SVMs, and neural networks	https://www.npatlas.org/discover/snapms/	Advantage • Rapid and user friendly in annotation • Process large datasets efficiently • Integration with databases Limitation • Depend on quality and comprehensiveness of reference databases • Expertise required to handle MS datasets	[221]
CANPA	Predict metabolite structures from identified compounds using an MS/MS prediction module and a collaborative library of (bio)chemical transformations for analogue discovery	Decision trees, SVM, and neural networks	https://metwork.pharmacie.parisdescartes.fr.	Advantage • Facilitate collaborative support by the sharing of libraries and annotations • Efficient and accurate in prediction Limitation • Complex setup and use • Requires MS and computational expertise	[317]
NP-analyst	Prioritize metabolites for isolation and create global network views of biologically active chemical space in large extract libraries	Different supervised learning and neural networks	www.npanalyst.org	Advantage • Useful in bioactive NPs discovery • Rapid, accurate, and flexible strategy Limitation • Depends on the quality and comprehensiveness of reference libraries • Potential for false positives/negatives result	[318]
R-FiBiCo Script	Target bioactive compounds in natural extracts	Spearman correlation, F-PCA, PLS, and PLS-DA		Advantage • Identification of bioactive compound from complex extract • Can process large datasets • Freely available Limitation • Require knowledge of statistical, programming language, chemical and biological data • High quality data and computational resources required	[319]
MS2Query	Integrates Spec2Vec, MS2Deepscore, and precursor masses to rank potential analogues and exact matches	Spec2Vec and MS2Deepscore learning	https://github.com/iomega/ms2query	Advantage • Freely available and rapid annotation of compounds • Process huge datasets • Integrated with databases Limitation • Dependent on the quality of databases • Require knowledge of MS • Risk of incorrect annotations	[195]
MAW	Automate complex annotation by integrating MS^2 data pre-processing, spectral and compound database matching, computational classification, and *in silico* annotation	Cosine similarity, Euclidean distance, and spectral correlation	https://github.com/zmahnoor14/MAW	Advantage • Automatic annotation • Provide precise annotation by leveraging comprehensive databases and advanced algorithms Limitation • Require complex initial setup and configuration • Results depends on the quality of input data and reference databases	[320]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
MolNetEnhancer	Facilitates chemical annotation, visualization, and discovery of substructure diversity within molecular families by integrating molecular networking, MS2LDA, *in silico* annotation tools (e.g., Network Annotation Propagation or DEREPLICATOR), and ClassyFire	Pattern recognition algorithms (PRA)	https://github.com/madeleineernst/pyMolNetEnhancer https://github.com/madeleineernst/RMolNetEnhancer	Advantage • Automated chemical classification as molecular families • Provide structural details for each fragmentation spectrum Limitation • Require knowledge of molecular networking and MS data analysis • Accuracy of annotation relate to the quality of input data and reference databases	[321]
SIRUS 4	Identify molecular structures using CSI:FingerID integrated with a molecular structure database	Random Forest (RF)	https://bio.informatik.uni-jena.de/sirius/	Advantage • Open accessible as graphical user interface (GUI) and command-line tools • Annotation is very accurate • Can handle wide range of metabolites Limitation • Require knowledge of MS data analysis and interpretation • Accuracy of annotation relate to the quality of input data and reference databases • Steep learning is need for new user or require comprehensive training or tutorials	[322]
MolDiscovery	Search small molecule MS/MS using efficient fragmentation algorithms and probabilistic models in an *in silico* database of over 8 million spectra	Probabilistic model	https://www.moldiscovery.com/	Advantage • Applicable to wide area of life science including NPD • Open accessible as graphical user interface (GUI) and command-line tools • Annotation is very accurate Limitation • Require knowledge of MS data analysis and interpretation • Accuracy of annotation relate to the quality of input data and reference databases	[190]
KGMN	Global annotation of unknown metabolites *via* knowledge-guided multi-layer metabolic networking	SVMs and Random Forests	http://metdna.zhulab.cn/	Advantage • Annotate both known and unknown metabolites • Annotation is very accurate as it integrates multiple layers of network, e.g., MS/MS, metabolic reaction, peak correlation to reach the conclusion Limitation$^{\Delta}$ • Might face difficulties in integrating with other bioinformatics tools and databases	[210]
MetDNA	Web server for metabolite identification and metabolic reaction network analysis from LC-MS/MS data	Recursive algorithm	http://metdna.zhulab.cn/	Advantage • No standard spectral library needed • Open access with web server interface Limitation$^{\Delta}$ • Might not support real-time or near-real-time data processing, which can be a limitation for some applications	[323]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
iSNAP	Analyzes individual spectra to reveal the significance of matches between MS/MS spectra and candidate NRP compounds	SVMs and RF	www-novo.cs.uwaterloo.ca:8180/isnap	Advantage • Dereplicate nonribosomal peptides in complex extract and identifies analogs with high accuracy • Automation, very sensitive and free accessible Limitation$^{\Delta}$ • Usually it is limited to NRPs class of Metabolites • Users may find limited flexibility in terms of customizing algorithms or parameters for specific needs	[32]
NAP	A tool in the GNPS web platform creates a network consensus of re-ranked structural candidates using molecular network topology and structural similarity to enhance *in silico* annotations	Graph-based algorithms	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp	Advantage • Automated annotation and open access *via* GNPS web-platform • Uses molecular network topology even when no match to a MS/MS spectrum in spectral libraries Limitation$^{\Delta}$ • The underlying network construction might be complex and not easily understandable for all users • High hardware requirements might limit its usability on standard desktop computers	[324]
BioCAn	Annotation tools for untargeted LC-MS data combine database searches and *in silico* fragmentation analyses, placing results in the biological context of a metabolic model			Advantage • Searches including *in silico* fragmentation analyses and database improved accuracy • Open access and process large datasets Limitation$^{\Delta}$ • Technical expertise required for initial setup • Challenges in interoperability with other commonly used bioinformatics tools and platforms • Lack of comprehensive learning resources and tutorials might be a hurdle for new users	[325]
xMS-Annotator	Uses a multi-criteria scoring algorithm to classify database matches by confidence levels, aiding in metabolite identification and prioritization for MS/MS confirmation	SVMs and RF	https://sourceforge.net/projects/xmsannotator/	Advantage • Utilizes a network-based approach to enhance the accuracy of metabolite annotation • Multi-database support and multi-criteria approach improve the confidence levels of prediction Limitation$^{\Delta}$ • Require knowledge of R programming and metabolomics data analysis • Limited or no graphical user interface (GUI) can be a challenge for users who prefer GUI-based tools	[326]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
METLIN and METLIN MS[2]	Database of ~3.0 million HRMS and 850,000 MS/MS molecular standards from diverse origins for identifying known and unknown metabolites and chemical entities	CNN and RNN	http://metlin.scripps.edu	Advantage • Open access extensive database with wide range of chemical entities • Data with positive and negative ionization increasing the robustness of dereplication Limitation[Δ] • The database might not be updated frequently enough to include the latest compounds and metabolites • Limited options for user-driven customization and addition of new entries	[327,328]
Nearest neighbor suspect spectral library	Contains 87,916 annotated MS/MS spectra derived from hundreds of millions of spectra from published untargeted metabolomics experiments	Nearest neighbor algorithms	https://github.com/bittremieux/gnps_suspect_library	Advantage • Free aviable via GNPS platform • Comprehensive and robust library with wide range of chemical entities • Annotations from known spectra to structurally related unknowns Limitation[Δ]	[329]
GNPS	Community-curated ecosystem for analyzing untargeted metabolomics data via small molecule mass spectrometry	PRA and graph-based algorithms	http://gnps.ucsd.edu	Advantage • User friendly, highthrought and versatile and open accesible • Integrated with many tools and database enhancing the overall workflow efficiency and data interoperability Limitation • Sharing data on a public platform might raise concerns about data privacy and proprietary information • Limited direct user support might make troubleshooting difficult	[215]
CliqueMS	Annotate in-source LC-MS[1] data in untargeted metabolomics based on coelution profile similarity	Graph-based algorithms	https://CRAN.R-project.org/package=cliqueMS	Advantage • Open access and rapidly process of large datasets • Uses a network-based algorithm to annotate isotopes, in-source adducts, and fragments from LC/MS data, reducing the complexity of data as similar compound to have similar signal Limitation[Δ] • It allows for custom adduct lists, some users might find the customization options limited compared to other tools	[330]
CANOPUS	Systematically annotate compound classes using fragmentation spectra	DNN	https://bio.informatik.uni-jena.de/software/canopus/	Advantage • Database independent and suitable for classification of unknown compounds • Offers both a graphical user interface (GUI) and command-line version • Provide the structural information even without the reference spectral Limitation[Δ] • Results are less accurate for very large or highly complex molecules • Limited availability of comprehensive learning resources or tutorials	[201]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
SCORE-metabolite-ID	Dereplicate known metabolites and dynamically analyze unknown compounds in complex mixtures by semi-automatically detecting correlated NMR and MS data, linking NMR signals to mass-to-charge ratios from ESI mass spectra	Three-dimensional correlation		Advantage • Enables fast and reliable dereplication • Analyze complex mixtures without the need for individual isolation of compounds Limitation$^{\Delta}$ • Require expertise in NMR, MS, and LC techniques, as well as proficiency in using MATLAB for analysis	[287]
VarQuest	A GNPS web tool, annotates and discovers novel peptide NPs	Variant search algorithm	http://cab.spbu.ru/software/varquest	Advantage • Capable of processing large datasets • Integrates with GNPS enhancing its utility • Can identify both peptidic and novel meatbolites Limitation$^{\Delta}$ • Require specific expertise in peptidic natural products • Designed for large datasets therefore require highthroughput settings	[331]
MSNovelist	Generate structures de novo from MS^2 spectra using fingerprint prediction with an encoder-decoder neural network	Encoder–decoder neural network	https://github.com/meowcat/MSNovelist	Advantage • Can identify the novel compound not present in existing spectral libraries • Dereplicate molecular structure prediction from MS2 spectra Limitation$^{\Delta}$	[197]
DEREPLICATOR +	Annotate metabolites from MS/MS data using *in silico* fragmentation graphs through single- and multistage fragmentation	*In silico* fragmentation algorithms	http://gnps.ucsd.edu/	Advantage • Annotate both peptidic and non-peptidic natural products • Can be integrated with other software and it is frequently upadted • Has hihg accuracy and resolution *via* use of multi satge fragmentation Limitation$^{\Delta}$ • Limit in discovering the completley new compound • High level of interpretation of results required • Diffcult to run in lab with less powerful computing resources	[188]
Ms2lda.org	Discover unknown compounds by extracting Mass2Motifs, then annotate and store these annotations from MS/MS data	Latent Dirichlet Allocation	http://ms2lda.org	Advantage • User friendly interface • Useful for scientist with less extensive knowledge in computational skills • Allows to create and share own motif sets Limitation$^{\Delta}$ • Useful in identifying the known compounds • Risk of false prediction	[332
MS-DIAL	A versatile program for untargeted metabolomics, enabling spectral deconvolution for GC-MS, LC-MS, and data-independent MS/MS	Data-dependent acquisition and data-independent acquisition	http://prime.psc.riken.jp/	Advantage • Supports a wide range of MS datasets • User friendly intuitive interface • Capable of peak detection, identification, and quantification • Integrated to several bioinformatic tools and databases Limitation • Risk of flase positive result, necessitating careful validation • Less efficient in identifying the novel compound	[333]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
MS2Analyzer	Searches for specific neutral loss, product ions, *m/z* differences, and precursor ions in MS/MS spectra, mainly for lipid and glycoside annotation	PRA such as SVMs and RF	https://fiehnlab.ucdavis.edu/projects/ms2analyzer	Advantage • Enables detailed substructure annotation • Provide user friendly interface • Allows users to create custom query files to search for specific mass spectral features Limitation$^{\Delta}$ • Not updated and maintenance timely • Less effective in discovering entirely new molecule	[334]
NPvis	A tool for matching the peptidic NPs from MS/MS spectra	Interactive visualization	http://cab.cc.spbu.ru/npvis/	Advantage • Allows users to interactively map annotated spectrum peaks to corresponding structure fragments • Accessible as a web service and a standalone application • Supports an extended alphabet of amino acids and PNP-specific linkage bonds Limitation$^{\Delta}$ • Not updated and maintenance timely • Only focus on peptidic natural product	[335]
MetWork	Predict *in silico* metabolization, aiding NPs annotation and discovery		https://metwork.pharmacie.parisdescartes.fr	Advantage • Integrated with the GNPS and use the collaborative library of reactions • Predict potential metabolization pathways and (un) known or rare compound • User-friendly web-based interface • Metwork prediction maybe valuable assest for optimization the production of bioactive compounds Limitation • Regular software updates and maintenance is necessary • Require specific data formats or preprocessing steps • Specially focus only on metabolization	[336]
DREP-NP	Repository for DREP-NP DE replication database files, enabling rapid DE replication of known NPs using mass spectrometry and fast NMR (^{1}H, HSQC, HMBC) data		https://github.com/clzani/DEREP-NP	Advantage • Robust approach to NP identification as it use both MS and NMR spectral • Integrates with comprehensive database • User-friendly intuitive interface Limitation • Limited instrument compatibility so researchers need to adapt their workflows or invest in compatible equipment • Timely update integrated database and DREP-NP tool is necessary	[337]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
MyCompoundID MS/MS Search	375,809 Predicted metabolites MS/MS spectra	Spectral matching algorithms	www.MyCompoundID.org	Advantage • Supports batch processing of multiple spectra • Features automated identification programs, such as DnsID for dansyl labeled metabolites • Can be integrated to other bioinformatic tools Limitation$^{\Delta}$ • Limited spectral coverage • Integration require additional customization • Limited data format compatability	[338]
MS-FINDER	Elucidates compound structures from unknown EI-MS (GC/MS) and MS/MS spectra using hydrogen rearrangement	*In-silico* fragmentation algorithms	http://prime.psc.riken.jp/	Advantage • Highly accurate in identifying metabolites from MS/MS spectra • Use large comprehensive database and user friendly interface • Support batch processing Limitation$^{\Delta}$ • Accuracy of MS-FINDER relies on the quality and comprehensiveness of external spectral libraries	[339]
ChemDistiller	Annotates metabolites with tandem MS data from a vast database of pre-calculated "fingerprints" and fragmentation patterns	PRA	https://bitbucket.org/iAnalytica/chemdistillerpython	Advantage • Process large datasets efficiently • Use compound with pre-calculated fingerprints and fragmentation patterns • Customizable making it suitable for both small and large-scale studies Limitation$^{\Delta}$ • Specific data formats or preprocessing of spectral	[340]
MAGMa	Annotate untargeted metabolomics MS data by generating theoretical MS/MS fragments from target structures	Multi-task Gaussian processes	https://nlesc.github.io/MAGMa/	Advantage • Supports hierarchical and internally consistent fragmentation trees • User-friendly intuitive interface • Utilize large spectral libraries Limitation$^{\Delta}$ • Struggle with accurately predicting MS/MS spectra for entirely new or highly unconventional metabolites • Requiring additional customization or technical expertise for integration	[341]
MIDAS	Identify unknown compounds in untargeted metabolomics by comparing theoretical MS/MS fragmentations to experimental tandem mass spectra	Autoencoders deep learning algorithm	http://midas.omicsbio.org	Advantage • Enumerates possible fragments from metabolites • Accurately match experimental MS/MS spectra with theoretical spectra • Automatic identification process using large database Limitation$^{\Delta}$ • Some results might require manual curation • Diffcult to process highly complex spectra	[342]

(*Continued*)

Table 2. Continued.

Tool	Description	AI-algorithm or techniques	Website/source code link	Advantage & limitation	Reference
MetFrag2.5	Annotate high-precision tandem mass spectra of metabolites using a library of 100,000+ fragmentation rules based on standard reactions	Bayesian models and statistical learning methods	http://cruttkies.github.io/MetFrag/	Advantage • Offer user-friendly intuitive interface • Processing large datasets efficiently • Utilizes extensive databases like PubChem and ChemSpider, enhancing the likelihood of correct identification • Timely update and refinements for improving its performance and accuracy Limitation$^{\Delta}$	[343]
HAMMER	Annotate metabolites in complex samples by generating *in silico* MS^n libraries and batch matching of *in silico* mass spectral data	*In silico* fragmentation algorithms	http://www.biosciences-labs.bham.ac.uk/viant/hammer/	Advantage • Facilitate the *in silico* MS^n libraries generation for enhance identification • Perform batch matching of *in-silico* mass spectral data Limitation$^{\Delta}$ • Complex setting up and configuring • High computational capacity for libraries generation	[344]
CFM-ID 4.0	Predict, annotate and identify for small molecules from ESI-QTOF-MS/MS	Competitive fragmentation modeling	https://cfmid.wishartlab.com	Advantage • Accurately predict ESI-MS/MS spectra for a given compound structure • Cover wide range of chemical classes • Timely update and maintenance with improved performance • User friendly and can integrate with other tools Limitation$^{\Delta}$ • Requires substantial computational power • Potential risk of false prediction	[345]
HypoRiPPAtlas	Hypothetical tandem mass spectra datasets from NPs in 22,671 complete microbial genomes	SVMs and RF	https://github.com/mohimanilab/seq2ripp	Advantage • Facilitate the identification of novel RiPPs from MS and genomic data • Process of large datasets quickly and efficiently • Offer user-friendly intuitive interface Limitation$^{\Delta}$ • Require both genomic and MS data • Limited experimental library coverage	[189]

$^{\Delta}$Require high quality and completeness of input data; substantial computational resources; Potential for false positive/negative prediction; require good understanding of tool and expertise to interpret the results.

workflow have been developed. For example, BiG-SCAPE and CORASON [21] together with BiG-SLICE [49], enabled reconstruction of BGCs phylogenies from different sources and groups into gene cluster families, respectively. Additionally, the BiG-FAM database [50], guided by ML offers an improved, user-friendly interface that facilitate the display and comparison of gene clusters directly from query sequences. Moreover, MlBiG effectively connect biosynthetic enzymes to the chemical transformations they catalyze [51]. The RiPPQuest [31], NRPQuest [52], Pep2Path [53], NRPSPredictor2 [54], NPOmix [55], and molecular networking (MN) approach identify the potential gene clusters or gene cluster families from analytical datasets (e.g., MS/MS fragmentation) [56]. Recent examples, such as: linaridins, tambromycin, tyrobetaines, pristinin A3, deepflavo and deepginsen, thioviramide and thiovarsolin, and corynaridin augment the potential of

AI-based prediction softwares in discovering new BGCs classes or NCEs, including unusual NPs families like imiditides [1,5] (Figure 2).

Repurposing of existing knowledge of natural products

Drug repurposing or drug repositioning, identifies anecdote uses of previously approved drugs or known metabolites [57] (e.g., Cyclosporin from *Tolypocladium inflatum* Gams was discovered as an antifungal agent [58] and later approved as an immune suppressant drug [59]), has propelled much attention as it overcomes the time, safety and cost constraints of the traditional NPDD [60]. In the earliest demonstration various CADD approaches [61] smeared individually or in combination systematically benchmark substantial-scale information to obtain meaningful interpretations of *in silico*, NPs, or synthetic drug repurposing. QSAR, computer-aided VS and proteochemometric modelling [62,63] being the widely accepted, among others. In recent times, incorporating DL [64] and ML architectures (both supervised and unsupervised) [19] into CADD tools has substantially contributed to NPs-based drug discovery [65] (e.g., fusidic acid, isolated from the fungus *Fusidium coccineum* as an antibacterial agent [66], later rediscovered as an antiviral agent [67]). QSAR models, tracing back to 1962 [68], predict and prioritize biological activities based on the features derived from their molecular structures and statistical modeling emphasized to find the correlations between extracted features [69]. However, classical QSAR numerically characterize molecular structures at different layers of structure representation without model development, a tectonic drawback [70]. To fill this gap, coupling QSAR with the ML/DL algorithm significantly improve the unstructured data size processing and complex data sets (e.g., both linear and non-linear), by generating models and unleashing accurate predictions [71]. Over the past few years, AI-based QSAR models such as deep QSAR [70] have successfully implemented and improved its NPDD performance [72]. For example, Julian Ivanov et al. (2022) applied QSAR machine-learning predictive models to identify novel drug candidates for the viral targets three chymotrypsin-like protease and RNA-dependent RNA polymerase [73]. Likewise, VS, initiated in the early 1970s, prioritized targeted molecules by scanning commercial available databases (e.g., ZINC [74], Drugbank [75]) or manually curated libraries (e.g., libraries created by compiling the compounds structure from literature manually or by use of AI tools, for instance, BioNLP and tensors [76], Word2vec [77], DECIMER [78]). In particular, approaches such as: structure-based drug discovery (SBDD) [79], ligand-based drug discovery (LBDD) [18], and hybrid methods [80] emerge as major noteworthy VS techniques for drug discovery and development. Mechanistically, structure-based VS softwares relay on scoring functions (SFs) calculation as the kernel logics [81], and recently, incorporating AI algorithm into VS tools (e.g., DeepVS [80], PlayMolecule BindScope [82], and TAME-VS [83]) improved the SFs precision for binding affinity prediction [84]. Notably, in the last couple of years, AI-driven tools integrated with docking software (e.g., PyRMD2Dock [85], Deep Docking [86]) have enabled high throughput screening of ultra-large chemical libraries [87]. In particular, together with synthesized and arbitrary compounds, several NPs have repurposed and repositioned using ML/DL architectures in VS framework [88]. For example, Gaudênci and Pereira (2020) repositioned five marine natural products (MNP) as SARS-CoV-2 M^{pro} inhibitors using the CADD workflow with an initial 11,162 MNPs and 5276 organic compound-derived QSAR architecture [89]. Likewise, Grisoni et al. (2019) identified two NP-mimetic against Alzheimer's disease using two ML module, namely, SPiDER (SOM-based Prediction of Drug Equivalence Relationships) and TIGER (Target Inference GEneratoR) [90]. Another example, identification of NPs as EGFR double mutants inhibitors by screening approximately 0.15 million molecules from various natural products libraries [91]. To overcome the imperfection of VS, for instance, the need for target proteins 3D structures in SBDD and poor prediction with limited binding ligands in LBDD, computational drug-target interactions (DTI) prediction methods can efficiently identify putative new drugs, or novel targets for existing drugs [92], overcoming high cost and time of experimental methods [93]. DTIs predicting platforms include feature-based, ML and DL-based models. Feature-based approaches such as SwissTargetPrediction [94], Pharm-IF [95], PharmMapper [96] rely on known DTI chemical descriptors for molecules and the descriptors for the targets to generate feature vectors, while ML algorithm-integrated tools (e.g., SPVec [97], STarFish [98], DTi2Vec [99], and NetLapRLS [100]) uses similarity-based associations. More leveraging high throughput AI-based DTI tools (e.g., DeepConv-DTI [101], DNN-DTIs [102], and LDS-CNN [103]) uses individual or in combination with DL neural architectures (Table 1; Figure 1(b)), often with heterogeneous data sources scalable platform (e.g., HIDTI [104], DTINet [92]). Recent efforts for broader scale productivity focus on AI-based approaches in DTI prediction using multiple prediction algorithms [98], network-based prediction tool (e.g., AOPEDF [105]), or multi-modal representation framework involving more than two heterogeneous networks [106]. For instance, Rodrigues et al. (2018) augmented a RF regression-based DTI prediction road map named DEcRyPT (Drug-Target Relationship

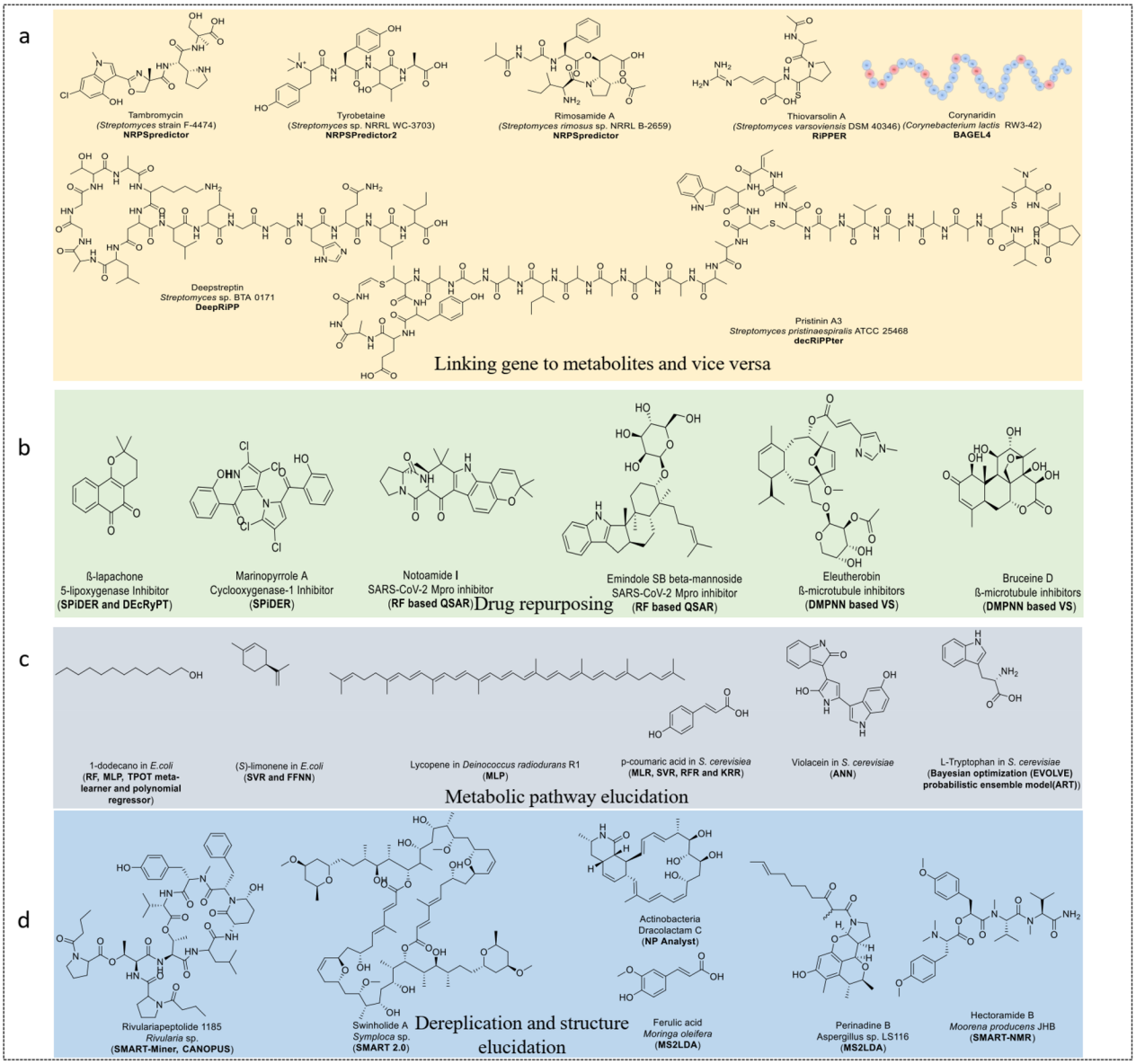


Figure 2. Illustrative examples of natural products in five major different areas of natural product discovery and drug development. Color code indicates the application areas of AI in natural product mining tools and datasets sources (Figure 1(c)). (a) Light yellow-association of the BGCs from genomic data to chemical structure or vice versa. (b) Light green- repurposing of existing knowledge of NPs. (c) Light grayish cyan- metabolic pathway-prediction, optimization and design. (d) Light blue- structure characterization and dereplication. Each compound structure is provided with its name, source information about its source (except in the case of repurposing drugs) and the tool applied for prediction.

Predictor) to identify *β*-lapachone as an allosteric modulator of 5-lipoxygenase [107]. Another example, TIGER, combining multiple SOMs predict NPs target [90,108].

Metabolic pathway-prediction, optimization and design

Until now, only approximately 10% of biochemical reactions involved in NPs biosynthesis are characterized, and organized in public metabolic pathway databases (e.g., PathBank 2.0 [109], KEGG 80.2 [110], EcoCyc [111]), that comprehensively annotated gene-reaction-metabolite connectivity [112]. In particular, reference-based procedures (also known as rule-based approaches) (e.g., PathWhiz [113], RAVEN [114], MRE [115]) generate the metabolic pathways by aligning and mapping the information to the reference template from the public databases [115], or manually curated databases, hitherto ensuring unprecedented performance as exemplified by methods like mIXGPR [116]. Nonetheless, these strategies struggle with the rapidly escalating metabolic and genomic datasets, leaving many pathways

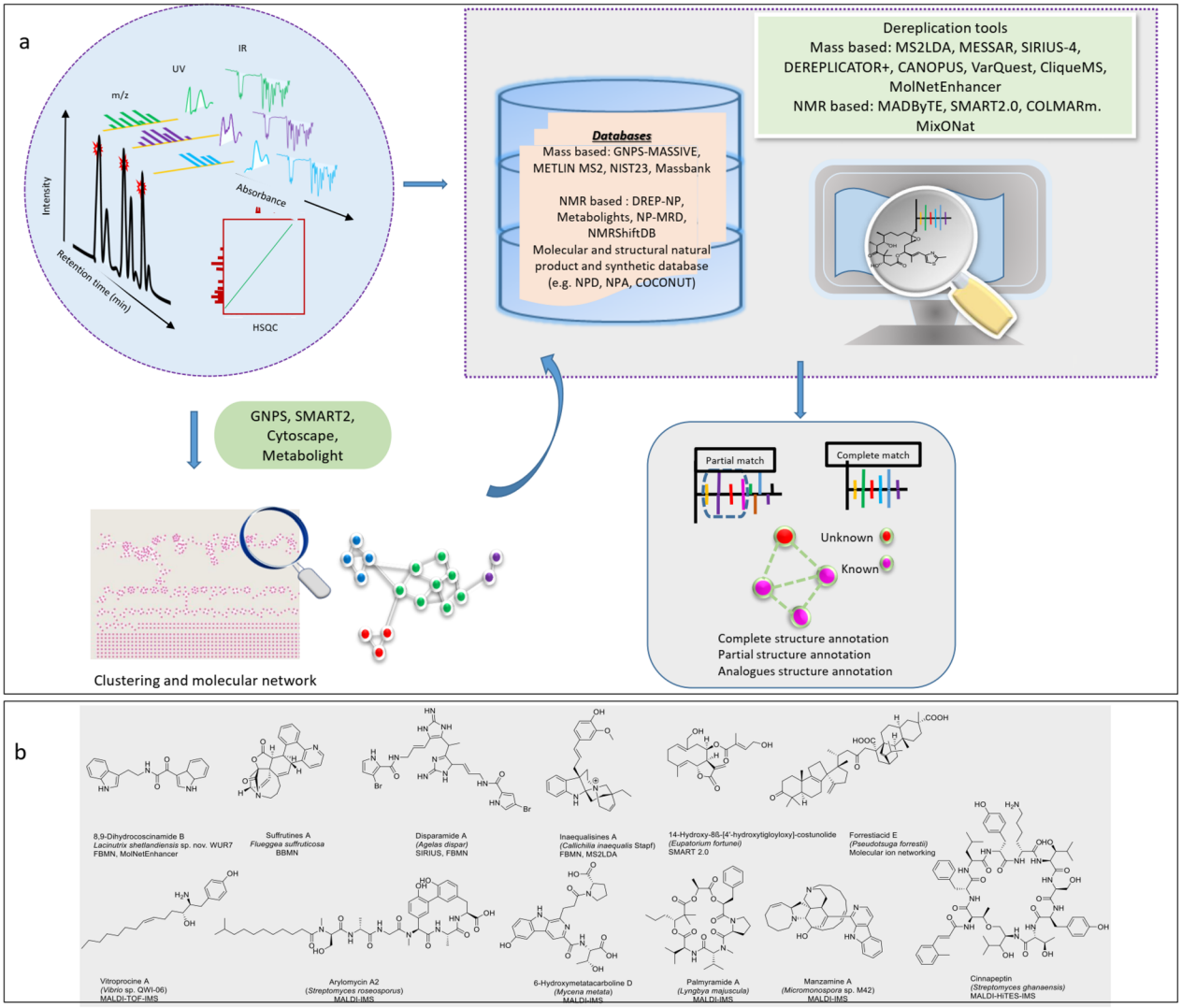


Figure 3. Interplay of molecular networking (MN) and AI based dereplication prediction tools in natural product discovery (NPD). (a) Schematic illustration of MN and dereplication tools for NPD. (b) Selected example of compounds discovered using different MN techniques and imaging mass spectrosocpy hyphenated techniques. The chemical structures are provided along with their names, sources, and the dereplication tools utilized during the structure discovery process.

uncharacterized, or inadequately understood [112]. To enhance the perceptiveness of complex NPs metabolic domains, and to predict reactions without predefined rules, AI algorithms have been integrated into three key areas of metabolic and biosynthetic networking: prediction [117], optimization [118] and redesign [119]. These models tailored time-series multi-omics data [120], genome-scale metabolic networks [121] and genome-scale metabolic models (e.g., CHESHIRE) [122] for scalable and accurate pathway prediction [123]. For example, Baranwal et al. (2020) adopted the hybrid framework of RF and GCN networks [124], Jia et al. (2020) proposed a similarity-based model for metabolic pathways prediction [125] while Koch et al. (2020) implemented the Monte Carlo Tree Search reinforcement learning procedure for synthetic pathways prediction within biological systems (e.g., RetroPath RL) [126]. Likewise Shah et al. (2022) devised a supervised learning model DeepRF to predict all metabolic pathways with high-performance accuracy (>97%), recall (>95%), and precision (>99%) [127]. The AI-driven biosynthetic pathway optimization relies on maximizing the product titers, rates, and yields (TRY). Novel techniques like promoter and ribosome binding site optimization for targeted gene expression have improved NPs production in various organisms [118]. For instance, Zhou et al. (2020) applied ML MiYA model to increase violacein production in *Saccharomyces cerevisiae* by approximately 2.5 folds [128]. While, Zhang et al. (2020) used ML tools such as ART [129] TeselaGen EVOLVE algorithm, together with mechanistic models, to train over 0.124 million experimental time series data in

yeast to improve the tryptophan titer and productivity by up to 74 and 43%, respectively. Likewise, Zheng et al. (2022) used the ML tool BioNavi-NP [112] for predicting building blocks and biosynthetic pathway intermediate with over 90% accuracy. Additional examples include: application of RF, polynomial, multilayer perceptron and TPOT meta-learner to optimize a 3-step pathway for dodecanol production [130], limonene production in *Escherichia coli* using support vector regression [118] and lycopene synthesis in *E. coli* using Gaussian processes [131], and METIS, for improving productivity and magnitude of crotonyl-CoA/ethyl malonyl-CoA/hydroxybutyryl-CoA (CETCH) cycle [132]. The reconstruction or design of metabolic pathways involve selecting the right substrate and enzyme to produce a desired product. Traditional metabolic design bioinformatics tools (e.g., PGDB PathoLogic program [133], Meta-Cyc [134]) used HMM annotation based on sequence homology searches against a database of previously characterized proteins for building the metabolic pathways [135]. Of late, AI algorithms forecast the most effective routes for synthesizing target molecules by providing a nexus among: genes, substrate, enzymes, and metabolites. In particular, tools like ESP [136] aid in substrate selection while EnzymeMiner [137], DEDAL [138], DeepRibo [139] identify enzymes using sequence and structural data and enzyme function by: DeepGOPlus [140], DeepGOWeb [141], DeepEC [142] and enzyme-substrate Michaelis constant K_M [143] forecast the biosynthetic routes of desired molecules. Further, reviews by Mendoza et al. (2019), Lawson et al. (2021), and Sveshnikova et al. (2022) on metabolic reconstruction tools [144], prediction and reconstruction paradigm [145], and computational tools for novel pathways design [146], respectively, provide deeper insights into these innovations, including those listed in Table S1.

Bioactivity prediction

Prediction of biological activity of known or recently discovered NPs is crucial in NPDD. Tools like: CODD-Pred [147], LiP-Quant [148] predict target, SwissADME [149], DBPP-Predictor [150], and CDRUG [151] assess drug likeness and pkCSM [152] evaluate toxicity. *In silico* biological activity prediction links molecular structure to biological activities using reverse docking (PASS target) [153]. Established bioactivity prediction emphasized mainly on similarity-based searching (e.g., MuSSeL [154], MolTarPred [155], and Turbo prediction [156]), or bioactivity spectra analysis (e.g., PASS online) [157] based on structural properties from available biologically annotated chemical compendium. However, these rules-based approaches have limitation, necessitating the development of molecular structure databases for better extrapolation. Recently, new methods "bottom-up activity predicting" have been developed to predict the activity using ML regression and SVM classifiers [158]. ML-powered bioinformatics tools such as PPB2 [159], InflamNat [160] predict drug target and anti-inflammatory activity including ANN (DeepTox [161], DeepIC_{50} [162], and DeepCYP [163]) use deep learning for toxicity, IC_{50}, and metabolism predictions, respectively. Additionally the DL-based PGMG tool generates bioactive molecules to meet the specific NPs pharmacophore (e.g., lavendustin A) [164].

Artificial intelligence in chemical structure characterization and dereplication tools

Improvements in mass spectroscopy (MS) [165], nuclear magnetic resonance (NMR) [166], X-ray crystallography [167] and microcrystal electron diffraction (MicroED) [168,169], have enabled rapid and efficient NPs detection, isolation, identification, and elucidation in complex biological extracts, even from nanoscale concentration [170]. These platforms and methods contrast sharply with older techniques that struggled with low concentrations and were prone to human bias and errors [171]. Undoubtedly, analytical tools and technologies have been foundational for the structure elucidation since the nineteenth century, often in high-throughput manners using hyphenated approaches [172], coined in 1980s [173], amalgamates multiple separation techniques and detection protocols (Table S2), often enhanced with ML-based predictive tools [174] and NPs databases [175] and, significantly, exploitation since the mid 1990s [176]. Additionally, newer specialized NPs dereplication hyphenated setups coupled with MN [177], bioactivity profiling [178], *in silico* screening of ultra-large databases such as FastEI [179] and high throughput programs [180]. Most of the time-honored to modern compound identification and structure elucidation strategies often rely on bioinformatics approaches that align spectra or chemical shift features of unknown compounds to curated spectral databases of reference compounds [181] (e.g., MassIVE [182], NMRShiftDB2 [183]) (Figure 3). However, lately, with the imminence of AI, including DNN and ANN, these approaches enable rapid data processing, analysis and predicting chemical structure with (out) prior knowledge [184]. Nevertheless, the use of AI-addendum rule-based approaches for the *de novo* identification of unknown compounds from MS data traced back to early 1960s [185].

Mass based structure identification

MS, being more sensitive and specific in comparison to NMR [181], has witnessed the dramatic upgrade in spectral databases and dereplication solutions of compounds over the past two decades [186]. In particular, modern prediction and characterization ML approaches rely on identifying metabolites by comparing query MS/MS spectrum against reference compounds or databases using the similarity-, distance- or probability functions [187]. In addition, *in silico* spectral or fragmental library tools (e.g., DEREPLICATOR + [188], HypoRiPPAtlas [189], molDiscovery [190] predict structure based on fragmentation rules [187]. In particular, ML-based prediction approaches relay on Markov modules (e.g., CFM-ID 3.0) [191], SVM (e.g., CSI: FingerID [192], Qemistree [193]), Kernel regression models (IOKRreverse and IOKRfusion) [194], RF architecture (e.g., MS2Query) [195] and Latent Dirichlet Allocation (e.g., MS2LDA) [196]. Moreover, DL frameworks (e.g., MSNovelist [197], MetFID [198], DeepNovo [199], and 3D-MSnet [200]) have improved automatic prediction of complete or fragments of small molecule structures from MS/MS spectra. In recent years, DL-based methodologies like CANOPUS [201] and NPClassifier [202] have been used to cluster mass spectra for systematic classification of unknown NP classes (e.g., rivulariapeptolides [203]), or identify features of specific metabolites classes (e.g., DeepIso [204], PointIso [205]). This contrast classical hierarchical class annotation tools such as ChEBI [206], ChemOnt and ClassyFire which uses the SMiles ARbitrary Target Specification (SMARTS) [207] for large-scale compounds set annotation based on substructure presence or absence.

Mass based molecular networking and dereplication

MS/MS based MN visualization platform displays the chemspace by aligning experimental tandem mass spectrometry (MS/MS) spectra of massive datasets size with each other and compare spectral similarities, assuming analogous have similar fragmentation, rather than comparison with reference MS/MS data [208] and preventing re-characterization via comparing the MS-based databases [209]. Comprehensive analysis of both targeted and untargeted metabolomics constructs multiple networks by analyzing the experimental and knowledge-based MS/MS data sets [210]. One remarkable tool, Global Natural Products Social Molecular Networking (GNPS), established in 2013, is an MS/MS fragmentation similarity-based social networking and dereplication tool. It supports MZmine 3 [211] or OpenMS 3 [212] files, databases like MASST [213] and tools such as ReDU [214] for dereplication and unbiased chemical relationship mapping molecules within, or across different samples [215]. Logically, there is a strong likelihood of identifying structurally similar metabolites through spectral similarity of MS data within the same or closely related clusters, thereby accelerating the annotation of homologous compounds [215]. More recently, classical MS/MS-based MN, pioneered in 2012 by *Dorrestein* et al. [216], further updated with additional features like the: feature-based molecular network [217], ion-identifying molecular network [218], building blocks-based molecular network [219], substructure-based molecular network (MS2LDA) [196], and bioactive molecular network [220], ameliorate the sensitiveness, high throughput, and robustness. Recently improved MN tools include SNAP-MS [221] for NP family annotation, Qemistree [193] for linking DNA sequences to metabolomics data, MetFID [198] for compounds fingerprinting, and Spec2Vec [222] for identifying patterns using natural language processing. Astoundingly, modern MS/MS spectra-based MN, and dereplication have opened an innovative veritable avalanche for the: identification of known analogues [220], discovery of unique classes [221], characterization of biosynthetic pathways [20], chemical ecology relationship elucidation [223], including guidance for the development of metabolomics protocols, methods and tools- MESSAR [224], CSI-FingerID [192], Qemistree [193] and microbeMASST [225]. Furthermore, the synergy created by combining MN with the (meta) genomic mining [226], mass spectrometry imaging [227,228], and stable isotope labeling techniques (e.g., MetExtract II [229], X^{13}CMS [230], geoRge [231], and MS spectra to genomic information [232] expanded access to decipher NPs chemical space.

Dereplication, first coined in the late 1970s, gained significant attention from natural product scientists after the 1990s [233], usually employ similarity-based modules to prevent the re-discovery of previously known compounds. These strategies rely on hyphenated techniques, bioactivity fingerprints, and database matching to identify metabolites from complex samples [234]. Until twenty first century, tools such as solid-phase extraction (e.g., SPE) [235], ion exchange chromatography (e.g., CPC, CEC) [236], and simple hyphenated techniques (Table S2) were fundamental, together with a few computerized databases (e.g., NAPRALERT [237], Berdy Database [238]) and limited biological screening systems (e.g., COMPARE program [239]). A significant flourish of analytical instruments, and techniques evolved between the 1960s and the 1990s. In early twenty first century, high-resolution Fourier transform mass spectrometry (FTMS) spectra, tandem mass spectrometry (MS/MS) spectra, and LC/

MS data and databases (e.g., METLIN [240], HMBD [241], NAPROC-13 [242]), including complex hyphenated techniques (Table S2) became dereplication tools [243]. Matching NMR- and MS spectra fragmentation from different annotated databases (e.g., XCMS2 [244], Massbank [245]) became the next stage of dereplication strategies for complex samples. The aforementioned dereplication strategies still need more confidence to unveil the unique structures and capture related analogues. Consequently, furtherance tools and databases (e.g., iMet [246], CAMERA [247]) that harbors extended parameters for baiting, for instance, retention time, adjunct ion, and multiple charges came into effect. In the last 10 years, continuous preference in the scope and datasets of metabolomics have led to the development of high throughput annotation, identification, and repository tools (Table 2), often incorporated with the artificial intelligence algorithm. The modern annotation and repertoire tools [248] offer unique addition comprehensive analysis for precursor ion ($MS^1/MS^2/MS^n$), adduct, isotope, and complex search algorithms (e.g., ChelomEx [249]), including data generating platforms for antibiotics screening (e.g., Bio MAP [250]), and cytological profiling [251].

Mass based imaging spectroscopy

Imaging mass spectroscopy (IMS), a mass based robust strategy, begun in the 1960s [252], has enabled: direct discovery of NPs [253], elucidation of biosynthetic pathway [254], cellular localization [255] quantification and observation of NPs-mediated microbial interactions [256]. In recent time, hyphenation in IMS such as high-throughput elicitor screening (HiTES) coupled with reporter systems and IMS [257] has garnered considerable scientific concern for mining the bacterial strains metabolomics profile [258]. For example, together with genome mining, it identifies the biosynthetic route for siphonazole in *Herpetosiphon* sp. B060 strain [259], discovered unprecedented NPs skeletons, including stendomycin (I–VI), lantipeptide AmfS, lasso peptide SSV-2083 from *Streptomyces hygroscopicus* ATCC 53653, *Streptomyces griseus* IFO 13350 and *Streptomyces sviceus* ATCC 20983, respectively [253]. In addition, Zhang & Seyedsayamdost (2020) discovered cinnapeptin, a cyclic cryptic depsipeptide from *Streptomyces ghanaensis* using MALDI-MS-guided HiTES [260].

NMR based structure identification

NMR based structure- elucidation, validation, and annotation are more challenging than MS-based structure characterization [171]. Notably, the forward approach for automatic prediction of chemical shifts, splitting patterns directly from chemical structures has achieved great success (Table 3) often facilitated by structure elucidation systems (e.g., NMRDB (https://www.nmrdb.org/), ChemDraw [261], Mestrenova [262], DetaNet [263] and ACD/Labs [264]). However, in the juxtaposed scenario, the rapid, accurate and automated prediction of the structure from NMR data remains challenging in metabolomics and NPD. Computer-assisted structure elucidation (CASE) is one of the noteworthy and earliest programs for suggesting best-fit structures from NMR and MS spectroscopic data, database information, and computational algorithms of NPs [265,266]. Recently, the CASE workflow has been synergistic with NMR predictive tools (e.g., Mestrenova [262], ACD/Structure Elucidator [267] and ACD/Labs (https://www.acdlabs.com/)) or atomic-force microscopy (AFM) and density functional theory (DFT) [268] to overcome the stereochemistry bottleneck in complex structure elucidation. Still, most CASE framework rely on databases containing chemical structures and spectra. In this scenario, Pesek et al. (2020) designed a database-independent rule-based computational algorithm to elucidate the structure of an unknown compound from spectroscopic ^1H and ^{13}C NMR, IR, and mass spectra, and, furthermore, bioinformatics tools, such as: SpecSolv [269], HOUDINI [270], COCON [271], StrucEluc [272], and LSD [273], play indispensable roles in the structure elucidation of complex organic compounds. Nevertheless, the enormous chemical space of compounds continues to limit fast, accurate and automatic structure prediction from NMR and other spectroscopic data. In recent decades, AI algorithms broadly applied for NMR spectra reconstruction from poor experimental data [274], including predictive tools for structure validation [275], automatic assignment [267] and elucidation [166,276] mounted exponentially, although the earliest application of ML/DL architectures in NMR dates back to 1970s [277], for prediction and solving the NMR-based structure elucidation and annotation of NPs [278]. During the previous years, Huang et al. (2021) proposed an ML framework for automated elucidation of unknown compounds using ^1H and ^{13}C NMR spectra [166]. Another, the SVM-M protocol, a versatile tool for identifying the structural and stereo chemical assignment of complex organic compounds per sublime confidence based on the ^{13}C NMR chemical shifts [279]. The SMART [278], SMART-Miner [280] and DeepSAT [281], CNN based methods use ^1H-^{13}C heteronuclear single quantum coherence (HSQC) spectra for efficient identification and rapid annotation of NPs molecular structures [282] and compound classes [278]. DeepSPInN predicts the molecular structures when given IR and ^{13}C NMR spectra without referring to any preexisting spectral databases

Table 3. NMR-based dereplication prediction tools.

Tool	Description	AI-algorithmn or techniques	Website/source code link	Adavantage & limitation	Reference
SMART 2.0	Dereplicate compounds using non-uniform sampling ^{1}H-^{13}C HSQC	Deep CNN		Advantage • Freely available, user-friendly, and offering high accuracy in prediction • Link to external databases such as GNPS, MiBIG, NPAtlas Limitation • Need high quality NMR data and technical expertise • Libarary limtation • Computational resources	[278]
MADByTE	Dereplicate using TOCSY and HSQC spectra to identify and match spin system features, creating a chemical similarity network for complex mixtures	Pattern recognition algorithm (PRA)	https://www.madbyte.org/resources	Advantage • Efficient in scafflods characterization • Open source and automation • Helpful for large sample sets Limitation • Need high quality NMR data and technical expertise • Complexity in setup and operation	[346]
MixONat	Allows dereplication of NPs from complex mixtures using ^{13}C NMR	PRA	https://sourceforge.net/projects/mixonat/	Advantage • Freely avialable and user friendly • Can perform annotation only from ^{13}C NMR spectra • Also useful for stereoisomers analysis Limitation • Less sensitive then MS based dereplication • Need high quality NMR data and technical expertise • Complexity in setup and operation	[291]
COLMARm	Dereplicate via use of 2D NMR (^{13}C-^{1}H HSQC, HSQC-TOCSY and ^{1}H-^{1}H TOCSY spectra	PRA	https://spin.ccic.osu.edu/index.php/colmarm/index2	Advantage • User friendly web-based interface • Provide the detailed structural information • Automatic processing of spectral data • Support wide range of input file format Limitation • Need high quality NMR data and technical expertise • Complexity in setup and operation • Need to learn the tools features and various functionalities to use it	[347]
^{1}H and ^{13}C and Predictor	Predict ^{13}C or ^{1}H NMR spectra for molecules, incorporated in software like ChemDraw, MestReNova, and NMRShiftDB	Ensemble approach (e.g., ML and Hose code)	https://mestrelab.com/software/mnova-software/ https://nmrshiftdb.nmr.uni-koeln.de/	Advantage • High accuracy and user friendly in prediction • Support wide range of input file format Limitation • Subscription based software • Need to learn the tools features and various functionalities to use it	[261]

(*Continued*)

Table 3. Continued.

Tool	Description	AI-algorithmn or techniques	Website/source code link	Adavantage & limitation	Reference
IMPRESSION	Predicts NMR parameters	Kernel Ridge Regression algorithm		Advantage • Offers high accuracy and speed in NMR predictions, while being user-friendly and accessible • Advantage in solving molecular conformation and stereoisomerism Limitation • Need computer resources and technical expertise • Requires high quality chemical data • May struggle with rare functional group	[348]
DeepSAT	Dereplicate known compound using ^{1}H-^{13}C HSQC NMR spectrum	End-to-end learning framework		Advantage • Offers high accuracy and speed in structure annotation and scaffold prediction, while being user-friendly and accessible • Deals with large datasets, and predicts compound classes Limitation • May struggle with rare functional group • Resources intensive and requires technical expertise	[281]

or molecular fragment knowledge [276]. Other recent methods, such as NMR-TS [283], ChemTS [284], DP4-AI [285] uses Monte Carlo tree search guided by an RNN to solve the molecular structure. Similarly, Kuhn et al. (2021) implemented CNN to detect substructures by using 2D HSQC and heteronuclear multiple bond correlation spectra [286]. Additionally, Zanardi and Sarotti augmented ANN mediated multidimensional pattern recognition from experimental and calculated 2D C–H COrrelation SpectroscopY to overcome the regio- or stereo chemical errors in NPs identification [275]. The SCORE-metabolite-ID [287] explores the unknown metabolite from complex mixtures using 3D correlation of ^{1}H-NMR, MS and LC data and ELINA [288] connects small molecule structures with their biological functions before isolation via bioactivity correlation of ^{1}H NMR signals of crude extracts.

Challenges in development and implementation of AI-powered bioinformatics tools

It is worth noting that although there has been immense success and meticulous improvement in AI-powered bioinformatics tools for NPs prioritization and prediction, several central issues remain in the design and prediction accuracy of these tools. Key challenges include the: poor-quality and incompleteness of training datasets, necessity of technical expertise, high-performance computing infrastructure, and paid subscription [289]. The quality of these datasets is often influenced by limited size, instrument parameters, and variability in annotation methods. Additionally, biases and false assignments, such as those involving structures in metabolomics or genomic datasets linked to chemical structures, can affect model accuracy. In particular, AI tools heavily rely on well-curated and accessible omics databases for model selection and training. However, accessibility of databases, data curation techniques or errors in automated data curation and restrictive copyright rules also hampered the completeness of datasets. Maintaining and updating these data sources and discovered tools are crucial for AI-driven discoveries (e.g., enhancing the prediction accuracy, fixing the bug, incorporating latest datasets), but this requires continuous validation and financial support. The next challenge is validation of model; the quality of standard datasets, the accuracy of established methods, and the availability of reliable experimental datasets critically hinder the validation of

innovated AI tools. Without high-quality data and robust validation methods, AI models predictions can lead to impractical and intangible results, often resulting in a false positive/negative.

Moreover, the heterogeneity of data derived from sources like NMR, MS, and omics, presents a significant challenge (e.g., integration of ultra large size experimental data). Redundant information can introduce bias, and the complexity of varying sources, processing methods, and analytical approaches makes standardization difficult. This complexity can lead to predictive models missing important NPs classes. Most of the current tools are limited to specific applicability domains, which skews discoveries. Given the diverse data types, it is unlikely that a single informatics tool can encompass all necessary information, though there is a strong emphasis on developing multiomic integration algorithms based comprehensive tools (e.g., multi-modal XCMS Online) [290] to facilitate the NPD. Finally, paid subscriptions (e.g., MixONat [291], Deepsat [281]) can indeed pose challenges when using dereplication software. Researchers with limited research funds might find it difficult to afford the recurring costs associated with these subscriptions, limiting their access to advanced tools and databases.

Conclusions

AI advancements have revolutionized (NP) D/DD. Analytical techniques like MS and NMR, combined with AI, have significantly improved the detection and characterization of NPs. Modern AI-based tools, including Tables 2 and 3 dereplicate (un)known NPs, link them to gene clusters, and suggest biosynthesis routes and predict DTIs.

Despite significant progress in AI for (NP) D/DD, there remains a need for new algorithms to extract meaningful features from heterogeneous data sources. In a nutshell, the future of AI-driven NPR is promising, with substantial advancements in key areas. For example, federated learning that will enable collaborative model training across institutions without sharing raw data, enhancing privacy and security. This approach facilitates diverse dataset pooling, improving model accuracy and robustness [292,293]. Likewise, the improvement of a multi-modal data integration tool that will combine genomic, proteomic, and metabolomics data for a comprehensive understanding of NPs biosynthesis pathway, accelerating NPs discovery [294] is necessary. Lastly, continued efforts to enhance accuracy, reliability, and accessibility, along with global data sharing within the scientific community, will drive advancements in AI-based innovation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Key Research and Development Programs [nos. 2022YFC2804104 and 2022YFC2804700, China], the Fundamental Research Funds for the Provincial Universities of Zhejiang [no. RF-A2022013, China], the programs of the National Natural Science Foundation of China [no. 42276137, China], and Zhejiang International Sci-Tech Cooperation Base for the Exploitation and Utilization of Nature Product.

ORCID

Buddha Bahadur Basnet http://orcid.org/0000-0003-1153-4860
Zhen-Yi Zhou http://orcid.org/0000-0003-1342-6054
Bin Wei http://orcid.org/0000-0002-2862-2907
Hong Wang http://orcid.org/0000-0003-0058-060X

References

[1] Atanasov AG, Zotchev SB, Dirsch VM, The International Natural Product Sciences Taskforce, et al. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021;20:200–216. doi: 10.1038/s41573-020-00114-z.
[2] Meinwald J. Natural products as molecular messengers. J Nat Prod. 2011;74:305–309. doi: 10.1021/np100754j.
[3] Newman DJ. Natural products and drug discovery. Natl Sci Rev. 2022;9:nwac206. doi: 10.1093/nsr/nwac206.
[4] Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nat Rev Drug Discov. 2015;14:111–129. doi: 10.1038/nrd4510.
[5] Mullowney MW, Duncan KR, Elsayed SS, et al. Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. 2023;22:895–916. doi: 10.1038/s41573-023-00774-7.
[6] Rodrigues T, Reker D, Schneider P, et al. Counting on natural products for drug design. Nat Chem. 2016;8:531–541. doi: 10.1038/nchem.2479.
[7] Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nat Chem Biol. 2015;11:639–648. doi: 10.1038/nchembio.1884.
[8] Rifaioglu AS, Atas H, Martin MJ, et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform. 2019;20:1878–1912. doi: 10.1093/bib/bby061.
[9] Schmidhuber J. Annotated history of modern AI and deep learning. 2022. arXiv:2212.11279. doi: 10.48550/ARXIV.2212.11279.
[10] Bernhard EB, Isabelle MG, Vladimir NV. A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual ACM Workshop on Computational

Learning Theory; 1992 Jul 27–29. Pittsburgh, PA; New York, NY: ACM; 1992.
[11] Saitta L. European Coordinating Committee for Artificial Intelligence, Associazione Italiana per l'Intelligenza Artificiale, editors. Machine learning: proceedings of the thirteenth international conference, Bari, Italy, July 3–6. San Francisco, California: Morgan Kaufmann; 1996.
[12] Melo MCR, Maasch JRMA, De La Fuente-Nunez C. Accelerating antibiotic discovery through artificial intelligence. Commun Biol. 2021;4:1050. doi: 10.1038/s42003-021-02586-0.
[13] Kim J, Park S, Min D, et al. Comprehensive survey of recent drug discovery using deep learning. Int J Mol Sci. 2021;22:9983. doi: 10.3390/ijms22189983.
[14] Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expert Opin Drug Discov. 2016;11:785–795. doi: 10.1080/17460441.2016.1201262.
[15] Jayatunga MKP, Xie W, Ruder L, et al. AI in small-molecule drug discovery: a coming wave? Nat Rev Drug Discov. 2022;21:175–176. doi: 10.1038/d41573-022-00025-1.
[16] Goodacre R, Kell DB, Bianchi G. Neural networks and olive oil. Nature. 1992;359:594–594. doi: 10.1038/359594a0.
[17] Nag S, Baidya ATK, Mandal A, et al. Deep learning tools for advancing drug discovery and development. 3 Biotech. 2022;12:110. doi: 10.1007/s13205-022-03165-8.
[18] Selvaraj C, Chandra I, Singh SK. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. Mol Divers. 2022;26:1893–1913. doi: 10.1007/s11030-021-10326-z.
[19] Wu Z, Zhu M, Kang Y, et al. Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. Brief Bioinform. 2021;22:bbaa321.
[20] Medema MH, De Rond T, Moore BS. Mining genomes to illuminate the specialized chemistry of life. Nat Rev Genet. 2021;22:553–571. doi: 10.1038/s41576-021-00363-7.
[21] Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol. 2020;16:60–68. doi: 10.1038/s41589-019-0400-9.
[22] Bauman KD, Butler KS, Moore BS, et al. Genome mining methods to discover bioactive natural products. Nat Prod Rep. 2021;38:2100–2129. doi: 10.1039/d1np00032b.
[23] Khaldi N, Seifuddin FT, Turner G, et al. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47:736–741. doi: 10.1016/j.fgb.2010.06.003.
[24] Medema MH, Blin K, Cimermancic P, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39:W339–W346. doi: 10.1093/nar/gkr466.
[25] Gluck-Thaler E, Haridas S, Binder M, et al. The architecture of metabolism maximizes biosynthetic diversity in the largest class of fungi. Mol Biol Evol. 2020;37:2838–2856. doi: 10.1093/molbev/msaa122.
[26] Hjörleifsson Eldjárn G, Ramsay A, Van Der Hooft JJJ, et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. PLOS Comput Biol. 2021;17:e1008920. doi: 10.1371/journal.pcbi.1008920.
[27] Zallot R, Oberg N, Gerlt JA. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry. 2019;58:4169–4182. doi: 10.1021/acs.biochem.9b00735.
[28] Erbilgin O, Rübel O, Louie KB, et al. MAGI: a method for metabolite annotation and gene integration. ACS Chem Biol. 2019;14:704–714. doi: 10.1021/acschembio.8b01107.
[29] van Heel AJ, de Jong A, Song C, et al. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res. 2018;46: W278–W281. doi: 10.1093/nar/gky383.
[30] Chevrette MG, Aicheler F, Kohlbacher O, et al. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across *Actinobacteria*. Bioinformatics. 2017;33:3202–3210. doi: 10.1093/bioinformatics/btx400.
[31] Johnston CW, Skinnider MA, Wyatt MA, et al. An automated genomes-to-natural products platform (GNP) for the discovery of modular natural products. Nat Commun. 2015;6:8421. doi: 10.1038/ncomms9421.
[32] Ibrahim A, Yang L, Johnston C, et al. Dereplicating nonribosomal peptides using an informatic search algorithm for natural products (iSNAP) discovery. Proc Natl Acad Sci USA. 2012;109:19196–19201. doi: 10.1073/pnas.1206376109.
[33] Cao L, Gurevich A, Alexander KL, et al. MetaMiner: a scalable peptidogenomics approach for discovery of ribosomal peptide natural products with blind modifications from microbial communities. Cell Syst. 2019;9:600–608.e4. doi: 10.1016/j.cels.2019.09.004.
[34] Tietz JI, Schwalen CJ, Patel PS, et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat Chem Biol. 2017;13:470–478. doi: 10.1038/nchembio.2319.
[35] Carroll LM, Larralde M, Fleck JS, et al. Accurate *de novo* identification of biosynthetic gene clusters with GECCO. Bioinformatics. 2021. doi: 10.1101/2021.05.03.442509.
[36] Van Riper SK, Higgins L, Carlis JV, et al. RIPPER: a framework for MS1 only metabolomics and proteomics label-free relative quantification. Bioinformatics. 2016;32:2035–2037. doi: 10.1093/bioinformatics/btw091.
[37] Agrawal P, Amir S, Barua D, et al. RiPPMiner-genome: a web resource for automated prediction of crosslinked chemical structures of RiPPs by genome mining. J Mol Biol. 2021;433(11):166887. doi: 10.1016/j.jmb.2021.166887.
[38] Merwin NJ, Mousa WK, Dejong CA, et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. Proc Natl Acad Sci USA. 2020;117:371–380. doi: 10.1073/pnas.1901493116.
[39] Hannigan GD, Prihoda D, Palicka A, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019;47:e110–e110–e110. doi: 10.1093/nar/gkz654.
[40] Sanchez S, Rogers JD, Rogers AB, et al. Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. Bioinformatics. 2023. doi: 10.1101/2023.05.23.540769.
[41] De Los Santos ELC. NeuRiPP: neural network identification of RiPP precursor peptides. Sci Rep. 2019;9:13406. doi: 10.1038/s41598-019-49764-z.

[42] Kloosterman AM, Cimermancic P, Elsayed SS, et al. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. PLOS Biol. 2020;18:e3001026. doi: 10.1371/journal.pbio.3001026.
[43] Kloosterman AM, Shelton KE, Van Wezel GP, et al. RRE-Finder: a genome-mining tool for class-independent RiPP discovery. mSystems. 2020;5:e00267-20. doi: 10.1128/mSystems.00267-20.
[44] Skinnider MA, Johnston CW, Gunabalasingam M, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 2020;11:6058. doi: 10.1038/s41467-020-19986-1.
[45] Blin K, Shaw S, Augustijn HE, et al. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. Nucleic Acids Res. 2023;51:W46–W50. doi: 10.1093/nar/gkad344.
[46] Li T, Tripathi A, Yu F, et al. DDAP: docking domain affinity and biosynthetic pathway prediction tool for type I polyketide synthases. Bioinformatics. 2020;36:942–944. doi: 10.1093/bioinformatics/btz677.
[47] Mongia M, Baral R, Adduri A, et al. AdenPredictor: accurate prediction of the adenylation domain specificity of nonribosomal peptide biosynthetic gene clusters in microbial genomes. Bioinformatics. 2023;39:i40–i46. doi: 10.1093/bioinformatics/btad235.
[48] Wang Y, Correa Marrero M, Medema MH, et al. Coevolution-based prediction of protein–protein interactions in polyketide biosynthetic assembly lines. Bioinformatics. 2020;36:4846–4853. doi: 10.1093/bioinformatics/btaa595.
[49] Kautsar SA, van der Hooft JJJ, de Ridder D, et al. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience. 2021;10(1):giaa154. doi: 10.1093/gigascience/giaa154.
[50] Kautsar SA, Blin K, Shaw S, et al. BiG-FAM: the biosynthetic gene cluster families database. Nucleic Acids Res. 2021;49:D490–D497. doi: 10.1093/nar/gkaa812.
[51] Terlouw BR, Blin K, Navarro-Muñoz JC, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 2023;51:D603–D610. doi: 10.1093/nar/gkac1049.
[52] Mohimani H, Liu W-T, Kersten RD, et al. NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. J Nat Prod. 2014;77:1902–1909. doi: 10.1021/np500370c.
[53] Medema MH, Paalvast Y, Nguyen DD, et al. Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLOS Comput Biol. 2014;10:e1003822. Gardner PP, editor doi: 10.1371/journal.pcbi.1003822.
[54] Röttig M, Medema MH, Blin K, et al. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 2011;39:W362–W367. doi: 10.1093/nar/gkr323.
[55] Leão TF, Wang M, da Silva R, et al. NPOmix: a machine learning classifier to connect mass spectrometry fragmentation data to biosynthetic gene clusters. PNAS Nexus. 2022;1:pgac257. doi: 10.1093/pnasnexus/pgac257.
[56] Nguyen DD, Wu C-H, Moree WJ, et al. MS/MS networking guided analysis of molecule and gene cluster families. Proc Natl Acad Sci. 2013;110(28):E2611–E2620. doi: 10.1073/pnas.1303471110.
[57] Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18:41–58. doi: 10.1038/nrd.2018.168.
[58] Merluzzi VJ, Adams J. The search for anti-inflammatory drugs: case histories from concept to clinic. Boston: Birkhäuser; 1995.
[59] Tedesco D, Haragsim L. Cyclosporine: a review. J Transplant. 2012;2012:230386–230387. doi: 10.1155/2012/230386.
[60] Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9:203–214. doi: 10.1038/nrd3078.
[61] Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616:673–685. doi: 10.1038/s41586-023-05905-z.
[62] Kim PT, Winter R, Clevert D-A. Unsupervised representation learning for proteochemometric modeling. Int J Mol Sci. 2021;22:12882. doi: 10.3390/ijms222312882.
[63] Bongers BJ, IJzerman AP, Van Westen GJP. Proteochemometrics – recent developments in bioactivity and selectivity modeling. Drug Discov Today Technol. 2019;32–33:89–98. doi: 10.1016/j.ddtec.2020.08.003.
[64] Kimber TB, Chen Y, Volkamer A. Deep learning in virtual screening: recent applications and developments. Int J Mol Sci. 2021;22:4435. doi: 10.3390/ijms22094435.
[65] Koromina M, Pandi M-T, Patrinos GP. Rethinking drug repositioning and development with artificial intelligence, machine learning, and omics. OMICS. 2019;23:539–548. doi: 10.1089/omi.2019.0151.
[66] Godtfredsen W, Roholt K, Tybring L. FUCIDIN. The Lancet. 1962;279:928–931. doi: 10.1016/S0140-6736(62)91968-2.
[67] Hetmann M, Langner C, Durmaz V, et al. Identification and validation of fusidic acid and flufenamic acid as inhibitors of SARS-CoV-2 replication using DrugSolver CavitomiX. Sci Rep. 2023;13:11783. doi: 10.1038/s41598-023-39071-z.
[68] Hansch C, Maloney PP, Fujita T, et al. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature. 1962;194:178–180. doi: 10.1038/194178b0.
[69] Xu Y. Deep neural networks for QSAR. In: Heifetz A, editor. Artififfical intelligence in drug design. New York, NY: Springer US; 2022. p. 233–260.
[70] Tropsha A, Isayev O, Varnek A, et al. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov. 2024;23:141–155. doi: 10.1038/s41573-023-00832-0.
[71] Mao J, Akhtar J, Zhang X, et al. Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models. iScience. 2021;24:103052. doi: 10.1016/j.isci.2021.103052.
[72] Hu S, Chen P, Gu P, et al. A deep learning-based chemical system for QSAR prediction. IEEE J Biomed Health Inform. 2020;24:3020–3028. doi: 10.1109/JBHI.2020.2977009.
[73] Ivanov J, Polshakov D, Kato-Weinstein J, et al. Quantitative structure–activity relationship machine learning models and their applications for identifying

viral 3CLpro- and RdRp-targeting compounds as potential therapeutics for COVID-19 and related viral infections. ACS Omega. 2020;5:27344–27358. doi: 10.1021/acsomega.0c03682.

[74] Irwin JJ, Shoichet BK. ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005;45:177–182. doi: 10.1021/ci049714+.

[75] Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46:D1074–D1082. doi: 10.1093/nar/gkx1037.

[76] Gachloo M, Wang Y, Xia J. A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. Genomics Inform. 2019;17:e18. doi: 10.5808/GI.2019.17.2.e18.

[77] Buchan DWA, Jones DT. Learning a functional grammar of protein domains using natural language word embedding techniques. Proteins Struct Funct Bioinforma. 2020;88:616–624. doi: 10.1002/prot.25842.

[78] Rajan K, Zielesny A, Steinbeck C. DECIMER 1.0: deep learning for chemical image recognition using transformers. J Cheminform. 2021;13:61. doi: 10.1186/s13321-021-00538-8.

[79] Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. Int J Mol Sci. 2019;20:2783. doi: 10.3390/ijms20112783.

[80] Pereira JC, Caffarena ER, Dos Santos CN. Boosting docking-based virtual screening with deep learning. J Chem Inf Model. 2016;56:2495–2506. doi: 10.1021/acs.jcim.6b00355.

[81] Huang S-Y, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Phys Chem Chem Phys. 2010;12:12899–12908. doi: 10.1039/c0cp00151a.

[82] Skalic M, Martínez-Rosell G, Jiménez J, et al. PlayMolecule BindScope: large scale CNN-based virtual screening on the web. Bioinformatics. 2019;35:1237–1238. doi: 10.1093/bioinformatics/bty758.

[83] Bian Y, Kwon JJ, Liu C, et al. Target-driven machine learning-enabled virtual screening (TAME-VS) platform for early-stage hit identification. Front Mol Biosci. 2023;10:1163536. doi: 10.3389/fmolb.2023.1163536.

[84] Ye W-L, Shen C, Xiong G-L, et al. Improving docking-based virtual screening ability by integrating multiple energy auxiliary terms from molecular docking scoring. J Chem Inf Model. 2020;60:4216–4230. doi: 10.1021/acs.jcim.9b00977.

[85] Roggia M, Natale B, Amendola G, et al. Streamlining large chemical library docking with artificial intelligence: the PyRMD2Dock approach. J Chem Inf Model. 2024;64:2143–2149. doi: 10.1021/acs.jcim.3c00647.

[86] Gentile F, Agrawal V, Hsing M, et al. Deep docking: a deep learning platform for augmentation of structure based drug discovery. ACS Cent Sci. 2020;6:939–949. doi: 10.1021/acscentsci.0c00229.

[87] Gentile F, Yaacoub JC, Gleave J, et al. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. Nat Protoc. 2022;17:672–697. doi: 10.1038/s41596-021-00659-2.

[88] Moumbock AFA, Li J, Mishra P, et al. Current computational methods for predicting protein interactions of natural products. Comput Struct Biotechnol J. 2019;17:1367–1376. doi: 10.1016/j.csbj.2019.08.008.

[89] Gaudêncio SP, Pereira F. A computer-aided drug design approach to predict marine drug-like leads for SARS-CoV-2 main protease inhibition. Mar Drugs. 2020;18:633. doi: 10.3390/md18120633.

[90] Grisoni F, Merk D, Friedrich L, et al. Design of natural-product-inspired multitarget ligands by machine learning. ChemMedChem. 2019;14:1129–1134. doi: 10.1002/cmdc.201900097.

[91] Agarwal SM, Nandekar P, Saini R. Computational identification of natural product inhibitors against EGFR double mutant (T790M/L858R) by integrating ADMET, machine learning, molecular docking and a dynamics approach. RSC Adv. 2022;12:16779–16789. doi: 10.1039/d2ra00373b.

[92] Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun. 2017;8:573. doi: 10.1038/s41467-017-00680-8.

[93] Chen X, Yan CC, Zhang X, et al. Drug–target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016;17:696–712. doi: 10.1093/bib/bbv066.

[94] Gfeller D, Grosdidier A, Wirth M, et al. SwissTargetPrediction: a web server for target prediction of bioactive small molecules. Nucleic Acids Res. 2014;42: W32–W38. doi: 10.1093/nar/gku293.

[95] Sato T, Honma T, Yokoyama S. Combining machine learning and pharmacophore-based interaction fingerprint for *in silico* screening. J Chem Inf Model. 2010;50:170–185. doi: 10.1021/ci900382e.

[96] Liu X, Ouyang S, Yu B, et al. PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res. 2010;38: W609–W614. doi: 10.1093/nar/gkq300.

[97] Zhang Y-F, Wang X, Kaushik AC, et al. SPVec: a Word2vec-inspired feature representation method for drug-target interaction prediction. Front Chem. 2019;7:895. doi: 10.3389/fchem.2019.00895.

[98] Cockroft NT, Cheng X, Fuchs JR. STarFish: a stacked ensemble target fishing approach and its application to natural products. J Chem Inf Model. 2019;59:4906–4920. doi: 10.1021/acs.jcim.9b00489.

[99] Thafar MA, Olayan RS, Albaradei S, et al. DTi2Vec: drug–target interaction prediction using network embedding and ensemble learning. J Cheminform. 2021;13:71. doi: 10.1186/s13321-021-00552-w.

[100] Xia Z, Wu L-Y, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010;4(Suppl 2):S6. doi: 10.1186/1752-0509-4-S2-S6.

[101] 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:e1007129. doi: 10.1371/journal.pcbi.1007129.

[102] Chen C, Shi H, Jiang Z, et al. DNN-DTIs: improved drug-target interactions prediction using XGBoost feature selection and deep neural network. Comput Biol Med. 2021;136:104676. doi: 10.1016/j.compbiomed.2021.104676.

[103] Wang Y, Zhang Z, Piao C, et al. LDS-CNN: a deep learning framework for drug-target interactions prediction based on large-scale drug screening. Health Inf Sci Syst. 2023;11:42. doi: 10.1007/s13755-023-00243-w.
[104] Soh J, Park S, Lee H. HIDTI: integration of heterogeneous information to predict drug-target interactions. Sci Rep. 2022;12:3793. doi: 10.1038/s41598-022-07608-3.
[105] Zeng X, Zhu S, Hou Y, et al. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. Bioinformatics. 2020;36:2805–2812. doi: 10.1093/bioinformatics/btaa010.
[106] Wei J, Lu L, Shen T. Predicting drug–protein interactions by preserving the graph information of multi source data. BMC Bioinf. 2024;25:10. doi: 10.1186/s12859-023-05620-6.
[107] Rodrigues T, Werner M, Roth J, et al. Machine intelligence decrypts β-lapachone as an allosteric 5-lipoxygenase inhibitor. Chem Sci. 2018;9:6899–6903. doi: 10.1039/c8sc02634c.
[108] Schneider P, Schneider G. De-orphaning the marine natural product (±)-marinopyrrole A by computational target prediction and biochemical validation. Chem Commun. 2017;53:2272–2274. doi: 10.1039/c6cc09693j.
[109] Wishart DS, Kruger R, Sivakumaran A, et al. PathBank 2.0—the pathway database for model organism metabolomics. Nucleic Acids Res. 2024;52: D654–D662. doi: 10.1093/nar/gkad1041.
[110] Jin Z, Sato Y, Kawashima M, et al. KEGG tools for classification and analysis of viral proteins. Protein Sci. 2023;32:e4820. doi: 10.1002/pro.4820.
[111] Karp PD, Paley S, Caspi R, et al. The EcoCyc database (2023). EcoSal Plus. 2023;11:eesp00022023. doi: 10.1128/ecosalplus.esp-0002-2023.
[112] Zheng S, Zeng T, Li C, et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. Nat Commun. 2022;13:3342. doi: 10.1038/s41467-022-30970-9.
[113] Ramirez-Gaona M, Marcu A, Pon A, et al. A web tool for generating high quality machine-readable biological pathways. J Vis Exp. 2017;120:54869.
[114] Wang H, Marcišauskas S, Sánchez BJ, et al. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. PLOS Comput Biol. 2018;14:e1006541. doi: 10.1371/journal.pcbi.1006541.
[115] Kuwahara H, Alazmi M, Cui X, et al. MRE: a web tool to suggest foreign enzymes for the biosynthesis pathway design with competing endogenous reactions in mind. Nucleic Acids Res. 2016;44:W217–W225. doi: 10.1093/nar/gkw342.
[116] Joe H, Kim H-G. Multi-label classification with XGBoost for metabolic pathway prediction. BMC Bioinf. 2024;25:52. doi: 10.1186/s12859-024-05666-0.
[117] Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. BMC Bioinf. 2010;11:15. doi: 10.1186/1471-2105-11-15.
[118] Jervis AJ, Carbonell P, Vinaixa M, et al. Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli*. ACS Synth Biol. 2019;8:127–136. doi: 10.1021/acssynbio.8b00398.
[119] Wang L, Dash S, Ng CY, et al. A review of computational tools for design and reconstruction of metabolic pathways. Synth Syst Biotechnol. 2017;2:243–252. doi: 10.1016/j.synbio.2017.11.002.
[120] Costello Z, Martin HG. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. NPJ Syst Biol Appl. 2018;4:19. doi: 10.1038/s41540-018-0054-3.
[121] Faust K, Croes D, Van Helden J. Prediction of metabolic pathways from genome-scale metabolic networks. Biosystems. 2011;105:109–121. doi: 10.1016/j.biosystems.2011.05.004.
[122] Chen C, Liao C, Liu Y-Y. Teasing out missing reactions in genome-scale metabolic networks through hypergraph learning. Nat Commun. 2023;14:2375. doi: 10.1038/s41467-023-38110-7.
[123] Yu T, Boob AG, Volk MJ, et al. Machine learning-enabled retrobiosynthesis of molecules. Nat Catal. 2023;6:137–151. doi: 10.1038/s41929-022-00909-w.
[124] Baranwal M, Magner A, Elvati P, et al. A deep learning architecture for metabolic pathway prediction. Bioinformatics. 2020;36:2547–2553. doi: 10.1093/bioinformatics/btz954.
[125] Jia Y, Zhao R, Chen L. Similarity-based machine learning model for predicting the metabolic pathways of compounds. IEEE Access. 2020;8:130687–130696. doi: 10.1109/ACCESS.2020.3009439.
[126] Koch M, Duigou T, Faulon J-L. Reinforcement learning for bioretrosynthesis. ACS Synth Biol. 2020;9:157–168. doi: 10.1021/acssynbio.9b00447.
[127] Shah HA, Liu J, Yang Z, et al. DeepRF: a deep learning method for predicting metabolic pathways in organisms based on annotated genomes. Comput Biol Med. 2022;147:105756. doi: 10.1016/j.compbiomed.2022.105756.
[128] Zhou Y, Li G, Dong J, et al. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in Saccharomyces cerevisiae. Metab Eng. 2018;47:294–302. doi: 10.1016/j.ymben.2018.03.020.
[129] Radivojević T, Costello Z, Workman K, et al. A machine learning automated recommendation tool for synthetic biology. Nat Commun. 2020;11:4879. doi: 10.1038/s41467-020-18008-4.
[130] Opgenorth P, Costello Z, Okada T, et al. Lessons from two design–build–test–learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. ACS Synth Biol. 2019;8:1337–1351. doi: 10.1021/acssynbio.9b00020.
[131] HamediRad M, Chao R, Weisberg S, et al. Towards a fully automated algorithm driven platform for biosystems design. Nat Commun. 2019;10:5150. doi: 10.1038/s41467-019-13189-z.
[132] Pandi A, Diehl C, Yazdizadeh Kharrazi A, et al. A versatile active learning workflow for optimization of genetic and metabolic networks. Nat Commun. 2022;13:3876. doi: 10.1038/s41467-022-31245-z.
[133] Karp PD, Paley SM, Krummenacker M, et al. Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 2010;11:40–79. doi: 10.1093/bib/bbp043.

[134] Caspi R, Foerster H, Fulcher CA, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 2006;34: D511–D516. doi: 10.1093/nar/gkj128.
[135] Karp PD, Latendresse M, Caspi R. The pathway tools pathway prediction algorithm. Stand Genomic Sci. 2011;5:424–429. doi: 10.4056/sigs.1794338.
[136] Kroll A, Ranjan S, Engqvist MKM, et al. A general model to predict small molecule substrates of enzymes based on machine and deep learning. Nat Commun. 2023;14:2787. doi: 10.1038/s41467-023-38347-2.
[137] Hon J, Borko S, Stourac J, et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res. 2020;48:W104–W109. doi: 10.1093/nar/gkaa372.
[138] Llinares-López F, Berthet Q, Blondel M, et al. Deep embedding and alignment of protein sequences. Nat Methods. 2023;20:104–111. doi: 10.1038/s41592-022-01700-2.
[139] Clauwaert J, Menschaert G, Waegeman W. DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. Nucleic Acids Res. 2019;47:e36–e36–e36. doi: 10.1093/nar/gkz061.
[140] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2021;37:1187–1187. doi: 10.1093/bioinformatics/btaa763.
[141] Kulmanov M, Zhapa-Camacho F, Hoehndorf R. DeepGOWeb: fast and accurate protein function prediction on the (semantic) web. Nucleic Acids Res. 2021;49:W140–W146. doi: 10.1093/nar/gkab373.
[142] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci USA. 2019;116:13996–14001. doi: 10.1073/pnas.1821905116.
[143] Kroll A, Engqvist MKM, Heckmann D, et al. Deep learning allows genome-scale prediction of Michaelis constants from structural features. PLOS Biol. 2021;19:e3001402. doi: 10.1371/journal.pbio.3001402.]
[144] Mendoza SN, Olivier BG, Molenaar D, et al. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biol. 2019;20:158. doi: 10.1186/s13059-019-1769-1.
[145] Shah HA, Liu J, Yang Z, et al. Review of machine learning methods for the prediction and reconstruction of metabolic pathways. Front Mol Biosci. 2021;8:634141. doi: 10.3389/fmolb.2021.634141.
[146] Sveshnikova A, MohammadiPeyhani H, Hatzimanikatis V. Computational tools and resources for designing new pathways to small molecules. Curr Opin Biotechnol. 2022;76:102722. doi: 10.1016/j.copbio.2022.102722.
[147] Yin X, Wang X, Li Y, et al. CODD-Pred: a web server for efficient target identification and bioactivity prediction of small molecules. J Chem Inf Model. 2023;63:6169–6176. doi: 10.1021/acs.jcim.3c00685.
[148] Piazza I, Beaton N, Bruderer R, et al. A machine learning-based chemoproteomic approach to identify drug targets and binding sites in complex proteomes. Nat Commun. 2020;11:4200. doi: 10.1038/s41467-020-18071-x.
[149] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017;7:42717. doi: 10.1038/srep42717.
[150] Gu Y, Wang Y, Zhu K, et al. DBPP-Predictor: a novel strategy for prediction of chemical drug-likeness based on property profiles. J Cheminform. 2024;16:4. doi: 10.1186/s13321-024-00800-9.
[151] Li G-H, Huang J-F. CDRUG: a web server for predicting anticancer activity of chemical compounds. Bioinformatics. 2012;28:3334–3335. doi: 10.1093/bioinformatics/bts625.
[152] Pires DEV, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015;58:4066–4072. doi: 10.1021/acs.jmedchem.5b00104.
[153] Pogodin PV, Lagunin AA, Filimonov DA, et al. PASS targets: ligand-based multi-target computational system based on a public data and naïve Bayes approach. SAR QSAR Environ Res. 2015;26:783–793. doi: 10.1080/1062936X.2015.1078407.
[154] Alberga D, Trisciuzzi D, Montaruli M, et al. A new approach for drug target and bioactivity prediction: the Multifingerprint Similarity Search Algorithm (MuSSeL). J Chem Inf Model. 2019;59:586–596. doi: 10.1021/acs.jcim.8b00698.
[155] Peón A, Li H, Ghislat G, et al. MolTarPred: a web tool for comprehensive target prediction with reliability estimation. Chem Biol Drug Des. 2019;94:1390–1401. doi: 10.1111/cbdd.13516.
[156] Abdo A, Pupin M. Turbo prediction: a new approach for bioactivity prediction. J Comput Aided Mol Des. 2022;36:77–85. doi: 10.1007/s10822-021-00440-3.
[157] Lagunin A, Stepanchikova A, Filimonov D, et al. PASS: prediction of activity spectra for biologically active substances. Bioinformatics. 2000;16:747–748. doi: 10.1093/bioinformatics/16.8.747.
[158] Walker AS, Clardy J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J Chem Inf Model. 2021;61:2560–2571. doi: 10.1021/acs.jcim.0c01304.
[159] Awale M, Reymond J-L. Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. J Chem Inf Model. 2019;59:10–17. doi: 10.1021/acs.jcim.8b00524.
[160] Zhang R, Ren S, Dai Q, et al. InflamNat: web-based database and predictor of anti-inflammatory natural products. J Cheminform. 2022;14:30. doi: 10.1186/s13321-022-00608-5.
[161] Mayr A, Klambauer G, Unterthiner T, et al. DeepTox: toxicity prediction using deep learning. Front Environ Sci. 2016;3:3. doi: 10.3389/fenvs.2015.00080.
[162] Joo M, Park A, Kim K, et al. A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients. Int J Mol Sci. 2019;20:6276. doi: 10.3390/ijms20246276.
[163] Li X, Xu Y, Lai L, et al. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. Mol Pharm. 2018;15:4336–4345. doi: 10.1021/acs.molpharmaceut.8b00110.
[164] Zhu H, Zhou R, Cao D, et al. A pharmacophore-guided deep learning approach for bioactive molecular generation. Nat Commun. 2023;14:6234. doi: 10.1038/s41467-023-41454-9.
[165] Ma X. Recent advances in mass spectrometry-based structural elucidation techniques. Molecules. 2022;27:6466. doi: 10.3390/molecules27196466.

[166] Huang Z, Chen MS, Woroch CP, et al. A framework for automated structure elucidation from routine NMR spectra. Chem Sci. 2021;12:15329–15338. doi: 10.1039/d1sc04105c.
[167] Takaba K, Maki-Yonekura S, Inoue I, et al. Structural resolution of a small organic molecule by serial X-ray free-electron laser and electron crystallography. Nat Chem. 2023;15:491–497. doi: 10.1038/s41557-023-01162-9.
[168] Kim LJ, Ohashi M, Zhang Z, et al. Prospecting for natural products by genome mining and microcrystal electron diffraction. Nat Chem Biol. 2021;17:872–877. doi: 10.1038/s41589-021-00834-2.
[169] Jones CG, Martynowycz MW, Hattne J, et al. The CryoEM METHOD MicroED as a powerful tool for small molecule structure determination. ACS Cent Sci. 2018;4:1587–1592. doi: 10.1021/acscentsci.8b00760.
[170] Molinski TF. NMR of natural products at the 'nanomole-scale'. Nat Prod Rep. 2010;27:321–329. doi: 10.1039/b920545b.
[171] Valli M, Russo HM, Pilon AC, et al. Computational methods for NMR and MS for structure elucidation I: software for basic NMR. Phys Sci Rev. 2019;4:20180108. doi: 10.1515/psr-2018-0108.
[172] Sarker SD, Nahar L. Hyphenated techniques and their applications in natural products analysis. In: Sarker SD, Nahar L, editors. Natural products isolation. Totowa, NJ: Humana Press; 2012. p. 301–340.
[173] Hirschfeld T. The hyphenated methods. Anal Chem. 1980;52:297A–312A. doi: 10.1021/ac50052a002.
[174] Madsen CT, Refsgaard JC, Teufel FG, et al. Combining mass spectrometry and machine learning to discover bioactive peptides. Nat Commun. 2022;13:6235. doi: 10.1038/s41467-022-34031-z.
[175] Li K, Chung-Davidson Y-W, Bussy U, et al. Recent advances and applications of experimental technologies in marine natural product research. Mar Drugs. 2015;13:2694–2713. doi: 10.3390/md13052694.
[176] Gebretsadik T, Linert W, Thomas M, et al. LC–NMR for natural product analysis: a journey from an academic curiosity to a robust analytical tool. Sci. 2021;3:6. doi: 10.3390/sci3010006.
[177] Allard P-M, Péresse T, Bisson J, et al. Integration of molecular networking and *in-silico* MS/MS fragmentation for natural products dereplication. Anal Chem. 2016;88:3317–3323. doi: 10.1021/acs.analchem.5b04804.
[178] Lang G, Mayhudin NA, Mitova MI, et al. Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. J Nat Prod. 2008;71:1595–1599. doi: 10.1021/np8002222.
[179] Yang Q, Ji H, Xu Z, et al. Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale *in-silico* library. Nat Commun. 2023;14:3722. doi: 10.1038/s41467-023-39279-7.
[180] Bobzin SC, Yang S, Kasten TP. LC-NMR: a new tool to expedite the dereplication and identification of natural products. J Ind Microbiol Biotechnol. 2000;25:342–345. doi: 10.1038/sj.jim.7000057.
[181] Wishart DS. Computational strategies for metabolite identification in metabolomics. Bioanalysis. 2009;1:1579–1596. doi: 10.4155/bio.09.138.
[182] Aron AT, Gentry EC, McPhail KL, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc. 2020;15:1954–1991. doi: 10.1038/s41596-020-0317-5.
[183] Kuhn S, Kolshorn H, Steinbeck C, et al. Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry. Magn Reson Chem. 2024;62:74–83. doi: 10.1002/mrc.5418.
[184] McAlpine JB, Chen S-N, Kutateladze A, et al. The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. Nat Prod Rep. 2019;36:35–107. doi: 10.1039/c7np00064b.
[185] Lindsay RK, editor. Applications of artificial intelligence for organic chemistry: the DENDRAL project. New York: McGraw-Hill Book Co; 1980.
[186] Tian Z, Liu F, Li D, et al. Strategies for structure elucidation of small molecules based on LC–MS/MS data from complex biological samples. Comput Struct Biotechnol J. 2022;20:5085–5097. doi: 10.1016/j.csbj.2022.09.004.
[187] Nguyen DH, Nguyen CH, Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. Brief Bioinform. 2019;20:2028–2043. doi: 10.1093/bib/bby066.
[188] Mohimani H, Gurevich A, Shlemov A, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018;9:4035. doi: 10.1038/s41467-018-06082-8.
[189] Lee Y-Y, Guler M, Chigumba DN, et al. HypoRiPPAtlas as an atlas of hypothetical natural products for mass spectrometry database search. Nat Commun. 2023;14:4219. doi: 10.1038/s41467-023-39905-4.
[190] Cao L, Guler M, Tagirdzhanov A, et al. MolDiscovery: learning mass spectrometry fragmentation of small molecules. Nat Commun. 2021;12:3718. doi: 10.1038/s41467-021-23986-0.
[191] Djoumbou-Feunang Y, Pon A, Karu N, et al. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. Metabolites. 2019;9:72. doi: 10.3390/metabo9040072.
[192] Dührkop K, Shen H, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI: fingerID. Proc Natl Acad Sci USA. 2015;112:12580–12585. doi: 10.1073/pnas.1509788112.
[193] Tripathi A, Vázquez-Baeza Y, Gauglitz JM, et al. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. Nat Chem Biol. 2021;17:146–151. doi: 10.1038/s41589-020-00677-3.
[194] Brouard C, Bassé A, d'Alché-Buc F, et al. Improved small molecule identification through learning combinations of kernel regression models. Metabolites. 2019;9:160. doi: 10.3390/metabo9080160.
[195] De Jonge NF, Louwen JJR, Chekmeneva E, et al. MS2Query: reliable and scalable MS2 mass spectra-based analogue search. Nat Commun. 2023;14:1752. doi: 10.1038/s41467-023-37446-4.
[196] Van Der Hooft JJJ, Wandy J, Barrett MP, et al. Topic modeling for untargeted substructure exploration in metabolomics. Proc Natl Acad Sci USA. 2016;113:13738–13743. doi: 10.1073/pnas.1608041113.

[197] Stravs MA, Dührkop K, Böcker S, et al. MSNovelist: de novo structure generation from mass spectra. Nat Methods. 2022;19:865–870. doi: 10.1038/s41592-022-01486-3.
[198] Fan Z, Alley A, Ghaffari K, et al. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. Metabolomics. 2020;16:104. doi: 10.1007/s11306-020-01726-7.
[199] Litsa EE, Chenthamarakshan V, Das P, et al. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. Commun Chem. 2023;6:132. doi: 10.1038/s42004-023-00932-3.
[200] Wang R, Lu M, An S, et al. 3D-MSNet: a point cloud-based deep learning model for untargeted feature detection and quantification in profile LC-HRMS data. Bioinformatics. 2023;39:btad195. doi: 10.1093/bioinformatics/btad195.
[201] Dührkop K, Nothias L-F, Fleischauer M, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39:462–471. doi: 10.1038/s41587-020-0740-8.
[202] Kim HW, Wang M, Leber CA, et al. NPClassifier: a deep neural network-based structural classification tool for natural products. J Nat Prod. 2021;84:2795–2807. doi: 10.1021/acs.jnatprod.1c00399.
[203] Reher R, Aron AT, Fajtová P, et al. Native metabolomics identifies the rivulariapeptolide family of protease inhibitors. Nat Commun. 2022;13:4619. doi: 10.1038/s41467-022-32016-6.
[204] Zohora FT, Rahman MZ, Tran NH, et al. DeepIso: a deep learning model for peptide feature detection from LC-MS map. Sci Rep. 2019;9:17168. doi: 10.1038/s41598-019-52954-4.
[205] Zohora FT, Rahman MZ, Tran NH, et al. Deep neural network for detecting arbitrary precision peptide features through attention based segmentation. Sci Rep. 2021;11:18249. doi: 10.1038/s41598-021-97669-7.
[206] Hastings J, De Matos P, Dekker A, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41:D456–D463. doi: 10.1093/nar/gks1146.
[207] Djoumbou Feunang Y, Eisner R, Knox C, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform. 2016;8:61. doi: 10.1186/s13321-016-0174-y.
[208] Qin G-F, Zhang X, Zhu F, et al. MS/MS-based molecular networking: an efficient approach for natural products dereplication. Molecules. 2022;28:157. doi: 10.3390/molecules28010157.
[209] Lai Z, Tsugawa H, Wohlgemuth G, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods. 2018;15:53–56. doi: 10.1038/nmeth.4512.
[210] Zhou Z, Luo M, Zhang H, et al. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. Nat Commun. 2022;13:6656. doi: 10.1038/s41467-022-34537-6.
[211] Schmid R, Heuckeroth S, Korf A, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nat Biotechnol. 2023;41:447–449. doi: 10.1038/s41587-023-01690-2.
[212] Pfeuffer J, Bielow C, Wein S, et al. OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data. Nat Methods. 2024;21:365–367. doi: 10.1038/s41592-024-02197-7.
[213] Wang M, Jarmusch AK, Vargas F, et al. Mass spectrometry searches using MASST. Nat Biotechnol. 2020;38:23–26. doi: 10.1038/s41587-019-0375-9.
[214] Jarmusch AK, Wang M, Aceves CM, et al. ReDU: a framework to find and reanalyze public mass spectrometry data. Nat Methods. 2020;17:901–904. doi: 10.1038/s41592-020-0916-7.
[215] Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat Biotechnol. 2016;34:828–837. doi: 10.1038/nbt.3597.
[216] Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci. 2012;109:E1743–1752. doi: 10.1073/pnas.1203689109.
[217] Nothias L-F, Petras D, Schmid R, et al. Feature-based molecular networking in the GNPS analysis environment. Nat Methods. 2020;17:905–908. doi: 10.1038/s41592-020-0933-6.
[218] Schmid R, Petras D, Nothias L-F, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun. 2021;12:3832. doi: 10.1038/s41467-021-23953-9.
[219] He Q, Wu Z, Li L, et al. Discovery of neuritogenic *Securinega* alkaloids from *Flueggea suffruticosa* by a building blocks-based molecular network strategy. Angew Chem Int Ed Engl. 2021;60:19609–19613. doi: 10.1002/anie.202103878.
[220] Nothias L-F, Nothias-Esposito M, Da Silva R, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. J Nat Prod. 2018;81:758–767. doi: 10.1021/acs.jnatprod.7b00737.
[221] Morehouse NJ, Clark TN, McMann EJ, et al. Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. Nat Commun. 2023;14:308. doi: 10.1038/s41467-022-35734-z.
[222] Huber F, Ridder L, Verhoeven S, et al. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. PLOS Comput Biol. 2021;17:e1008724. doi: 10.1371/journal.pcbi.1008724.
[223] Anderton CR, Chu RK, Tolić N, et al. Utilizing a robotic sprayer for high lateral and mass resolution MALDI FT-ICR MSI of microbial cultures. J Am Soc Mass Spectrom. 2016;27:556–559. doi: 10.1007/s13361-015-1324-6.
[224] Liu Y, Mrzic A, Meysman P, et al. MESSAR: automated recommendation of metabolite substructures from tandem mass spectra. PLOS One. 2020;15:e0226770. doi: 10.1371/journal.pone.0226770.
[225] Zuffa S, Schmid R, Bauermeister A, et al. microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. Nat Microbiol. 2024;9:336–345. doi: 10.1038/s41564-023-01575-9.
[226] Kleigrewe K, Almaliti J, Tian IY, et al. Combining mass spectrometric metabolic profiling with genomic analysis: a powerful approach for discovering natural products from Cyanobacteria. J Nat Prod. 2015;78:1671–1682. doi: 10.1021/acs.jnatprod.5b00301.

[227] Veličković D, Zemaitis KJ, Bhattacharjee A, et al. Mass spectrometry imaging of natural carbonyl products directly from agar-based microbial interactions using 4-APEBA derivatization. mSystems. 2024;9:e00803-23. doi: 10.1128/msystems.00803-23.
[228] Vallet M, Vanbellingen QP, Fu T, et al. An integrative approach to decipher the chemical antagonism between the competing endophytes *Paraconiothyrium variabile* and *Bacillus subtilis*. J Nat Prod. 2017;80:2863–2873. doi: 10.1021/acs.jnatprod.6b01185.
[229] Bueschl C, Kluger B, Neumann NKN, et al. MetExtract II: a software suite for stable isotope-assisted untargeted metabolomics. Anal Chem. 2017;89:9518–9526. doi: 10.1021/acs.analchem.7b02518.
[230] Llufrio EM, Cho K, Patti GJ. Systems-level analysis of isotopic labeling in untargeted metabolomic data by X13CMS. Nat Protoc. 2019;14:1970–1990. doi: 10.1038/s41596-019-0167-1.
[231] Capellades J, Navarro M, Samino S, et al. geoRge: a computational tool to detect the presence of stable isotope labeling in LC/MS-based untargeted metabolomics. Anal Chem. 2016;88:621–628. doi: 10.1021/acs.analchem.5b03628.
[232] Krause J. Applications and restrictions of integrated genomic and metabolomic screening: an accelerator for drug discovery from Actinomycetes? Molecules. 2021;26:5450. doi: 10.3390/molecules26185450.
[233] Dias DA, Urban S, Roessner U. A historical overview of natural products in drug discovery. Metabolites. 2012;2:303–336. doi: 10.3390/metabo2020303.
[234] Wolfender J-L, Marti G, Thomas A, et al. Current approaches and challenges for the metabolite profiling of complex natural extracts. J Chromatogr A. 2015;1382:136–164. doi: 10.1016/j.chroma.2014.10.091.
[235] Liška I. Fifty years of solid-phase extraction in water analysis – historical development and overview. J Chromatogr A. 2000;885:3–16. doi: 10.1016/S0021-9673(99)01144-9.
[236] Cannell RJP, editor. Natural products isolation. Totowa, N.J: Humana Press; 1998.
[237] Loub WD, Farnsworth NR, Soejarto DD, et al. NAPRALERT: computer handling of natural product research data. J Chem Inf Comput Sci. 1985;25:99–103. doi: 10.1021/ci00046a009.
[238] Berdy J, Kertesz M. Bioactive natural products database: an aid for natural products identification. In: Collier HR, editor. Chem Inf. Berlin, Heidelberg: Springer Berlin Heidelberg; 1989. p. 237–251.
[239] Paull KD, Shoemaker RH, Hodes L, et al. Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. J Natl Cancer Inst. 1989;81:1088–1092. doi: 10.1093/jnci/81.14.1088.
[240] Smith CA, Maille GO, Want EJ, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit. 2005;27:747–751. doi: 10.1097/01.ftd.0000179845.53213.39.
[241] Wishart DS, Tzur D, Knox C, et al. HMDB: the human metabolome database. Nucleic Acids Res. 2007;35:D521–D526. doi: 10.1093/nar/gkl923.
[242] López-Pérez JL, Therón R, Del Olmo E, et al. NAPROC-13: a database for the dereplication of natural product mixtures in bioassay-guided protocols. Bioinformatics. 2007;23:3256–3257. doi: 10.1093/bioinformatics/btm516.
[243] Cui Q, Lewis IA, Hegeman AD, et al. Metabolite identification via the Madison metabolomics consortium database. Nat Biotechnol. 2008;26:162–164. doi: 10.1038/nbt0208-162.
[244] Benton HP, Wong DM, Trauger SA, et al. XCMS ^{2}: processing tandem mass spectrometry data for metabolite identification and structural characterization. Anal Chem. 2008;80:6382–6389. doi: 10.1021/ac800795f.
[245] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45:703–714. doi: 10.1002/jms.1777.
[246] Aguilar-Mogas A, Sales-Pardo M, Navarro M, et al. iMet: a network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. Anal Chem. 2017;89:3474–3482. doi: 10.1021/acs.analchem.6b04512.
[247] Kuhl C, Tautenhahn R, Böttcher C, et al. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84:283–289. doi: 10.1021/ac202450g.
[248] Gaudêncio SP, Bayram E, Lukić Bilela L, et al. Advanced methods for natural products discovery: bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. Mar Drugs. 2023;21:308. doi: 10.3390/md21050308.
[249] Baars O, Morel FMM, Perlman DH. ChelomEx: isotope-assisted discovery of metal chelates in complex media using high-resolution LC-MS. Anal Chem. 2014;86:11298–11305. doi: 10.1021/ac503000e.
[250] Wong WR, Oliver AG, Linington RG. Development of antibiotic activity profile screening for the classification and discovery of natural product antibiotics. Chem Biol. 2012;19:1483–1495. doi: 10.1016/j.chembiol.2012.09.014.
[251] Ochoa JL, Bray WM, Lokey RS, et al. Phenotype-guided natural products discovery using cytological profiling. J Nat Prod. 2015;78:2242–2248. doi: 10.1021/acs.jnatprod.5b00455.
[252] Amstalden Van Hove ER, Smith DF, Heeren RMA. A concise review of mass spectrometry imaging. J Chromatogr A. 2010;1217:3946–3954. doi: 10.1016/j.chroma.2010.01.033.
[253] Kersten RD, Yang Y-L, Xu Y, et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011;7:794–802. doi: 10.1038/nchembio.684.
[254] Lin Y, Jiang X, Zhu S, et al. Multi-omics combined with MALDI mass spectroscopy imaging reveals the mechanisms of biosynthesis of characteristic compounds in *Tetrastigma hemsleyanum* Diels et Gilg. Front Plant Sci. 2023;14:1294804. doi: 10.3389/fpls.2023.1294804.
[255] Nie L, Huang L, Jia X, et al. Enhanced identification and localization of metabolites in Scutellariae Radix using ion mobility enabled MALDI-Q-TOF/MS imaging. J Pharm Anal. 2024;14:284–286. doi: 10.1016/j.jpha.2023.09.018.
[256] Yang Y, Xu Y, Kersten RD, et al. Connecting chemotypes and phenotypes of cultured marine microbial assemblages by imaging mass spectrometry. Angew Chem Int Ed Engl. 2011;50:5839–5842. doi: 10.1002/anie.201101225.
[257] Xu F, Wu Y, Zhang C, et al. A genetics-free method for high-throughput discovery of cryptic microbial metab-

olites. Nat Chem Biol. 2019;15:161–168. doi: 10.1038/s41589-018-0193-2.
[258] Hou J, Zhang Z, Wu W, et al. Mass spectrometry imaging: new eyes on natural products for drug research and development. Acta Pharmacol Sin. 2022;43:3096–3111. doi: 10.1038/s41401-022-00990-8.
[259] Tong Y, Whitford CM, Robertsen HL, et al. Highly efficient DSB-free base editing for streptomycetes with CRISPR-BEST. Proc Natl Acad Sci USA. 2019;116:20366–20375. doi: 10.1073/pnas.1913493116.
[260] Zhang C, Seyedsayamdost MR. Discovery of a cryptic depsipeptide from *Streptomyces ghanaensis* via MALDI-MS-guided high-throughput elicitor screening. Angew Chem Int Ed Engl. 2020;59:23005–23009. doi: 10.1002/anie.202009611.
[261] Cousins KR. Computer review of ChemDraw ultra 12.0. J Am Chem Soc. 2011;133:8388–8388. doi: 10.1021/ja204075s.
[262] Willcott MR. MestRe Nova. J. Am. Chem. Soc. 2009;131:13180–13180. doi: 10.1021/ja906709t.
[263] Zou Z, Zhang Y, Liang L, et al. A deep learning model for predicting selected organic molecular spectra. Nat Comput Sci. 2023;3:957–964. doi: 10.1038/s43588-023-00550-y.
[264] Elyashberg M, Williams A. ACD/structure elucidator: 20 years in the history of development. Molecules. 2021;26:6623. doi: 10.3390/molecules26216623.
[265] Bitchagno GTM, Fobofou Tanemossu SA. Computational methods for NMR and MS for structure elucidation III: more advanced approaches. Phys Sci Rev. 2019;4:20180109. doi: 10.1515/psr-2018-0109.
[266] Burns DC, Mazzola EP, Reynolds WF. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. Nat Prod Rep. 2019;36:919–933. doi: 10.1039/c9np00007k.
[267] Williams RB, O'Neil-Johnson M, Williams AJ, et al. Dereplication of natural products using minimal NMR data inputs. Org Biomol Chem. 2015;13:9957–9962. doi: 10.1039/c5ob01713k.
[268] Buevich AV, Elyashberg ME. Enhancing computer-assisted structure elucidation with DFT analysis of *J* -couplings. Magn Reson Chem. 2020;58:594–606. doi: 10.1002/mrc.4996.
[269] Will M, Fachinger W, Richert JR. Fully automated structure ElucidationA spectroscopist's dream comes true. J Chem Inf Comput Sci. 1996;36:221–227. doi: 10.1021/ci950092p.
[270] Korytko A, Schulz K-P, Madison MS, et al. HOUDINI: a new approach to computer-based structure generation. J Chem Inf Comput Sci. 2003;43:1434–1446. doi: 10.1021/ci034057r.
[271] Meiler J, Sanli E, Junker J, et al. Validation of structural proposals by substructure analysis and ^{13}C NMR chemical shift prediction. J Chem Inf Comput Sci. 2002;42:241–248. doi: 10.1021/ci010294n.
[272] Elyashberg ME, Blinov KA, Williams AJ, et al. *Structure Elucidator:* a versatile expert system for molecular structure elucidation from 1D and 2D NMR data and molecular fragments. J Chem Inf Comput Sci. 2004;44:771–792. doi: 10.1021/ci0341060.
[273] Plainchont B, De Paulo Emerenciano V, Nuzillard J. Recent advances in the structure elucidation of small organic molecules by the LSD software. Magn Reson Chem. 2013;51:447–453. doi: 10.1002/mrc.3965.
[274] Qu X, Huang Y, Lu H, et al. Accelerated nuclear magnetic resonance spectroscopy with deep learning. Angew Chem. 2020;132:10383–10386. doi: 10.1002/ange.201908162.
[275] Zanardi MM, Sarotti AM. GIAO C–H COSY simulations merged with artificial neural networks pattern recognition analysis. Pushing the structural validation a step forward. J Org Chem. 2015;80:9371–9378. doi: 10.1021/acs.joc.5b01663.
[276] Devata S, Sridharan B, Mehta S, et al. DeepSPInN – multimodal deep learning for molecular structure prediction from infrared and NMR spectra. Chemistry; 2023 [cited 2024 Jan 16]. Available from: https://chemrxiv.org/engage/chemrxiv/article-details/655de74b29a13c4d47ba0140.
[277] Cobas C. NMR signal processing, prediction, and structure verification with machine learning techniques. Magn Reson Chem. 2020;58:512–519. doi: 10.1002/mrc.4989.
[278] Zhang C, Idelbayev Y, Roberts N, et al. Small molecule accurate recognition technology (SMART) to enhance natural products research. Sci Rep. 2017;7:14243. doi: 10.1038/s41598-017-13923-x.
[279] Wu A, Ye Q, Zhuang X, et al. Elucidating structures of complex organic compounds using a machine learning model based on the ^{13}C NMR chemical shifts. Precis Chem. 2023;1:57–68. doi: 10.1021/prechem.3c00005.
[280] Kim HW, Zhang C, Cottrell GW, et al. SMART-Miner: a convolutional neural network-based metabolite identification from ^{1}H-^{13}C HSQC spectra. Magn Reson Chem. 2022;60:1070–1075. doi: 10.1002/mrc.5240.
[281] Kim HW, Zhang C, Reher R, et al. DeepSAT: learning molecular structures from nuclear magnetic resonance data. J Cheminform. 2023;15:71. doi: 10.1186/s13321-023-00738-4.
[282] Lee J, Park J, Kim J, et al. Targeted isolation of cytotoxic sesquiterpene lactones from *Eupatorium fortunei* by the NMR annotation tool, SMART 2.0. ACS Omega. 2020;5:23989–23995. doi: 10.1021/acsomega.0c03270.
[283] Zhang J, Terayama K, Sumita M, et al. NMR-TS: de novo molecule identification from NMR spectra. Sci Technol Adv Mater. 2020;21:552–561. doi: 10.1080/14686996.2020.1793382.
[284] Yang X, Zhang J, Yoshizoe K, et al. ChemTS: an efficient python library for *de novo* molecular generation. Sci Technol Adv Mater. 2017;18:972–976. doi: 10.1080/14686996.2017.1401424.
[285] Howarth A, Ermanis K, Goodman JM. DP4-AI automated NMR data analysis: straight from spectrometer to structure. Chem Sci. 2020;11:4351–4359. doi: 10.1039/d0sc00442a.
[286] Kuhn S, Tumer E, Colreavy-Donnelly S, et al. A pilot study for fragment identification using 2D NMR and deep learning. Magn Reson Chem. 2022;60:1052–1060. doi: 10.1002/mrc.5212.
[287] Watermann S, Bode M-C, Hackl T. Identification of metabolites from complex mixtures by 3D correlation of 1H NMR, MS and LC data using the SCORE-metabolite-ID approach. Sci Rep. 2023;13:15834. doi: 10.1038/s41598-023-43056-3.

[288] Grienke U, Foster PA, Zwirchmayr J, et al. 1H NMR-MS-based heterocovariance as a drug discovery tool for fishing bioactive compounds out of a complex mixture of structural analogues. Sci Rep. 2019;9:11113. doi: 10.1038/s41598-019-47434-8.

[289] Liang W, Tadesse GA, Ho D, et al. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell. 2022;4:669–677. doi: 10.1038/s42256-022-00516-1.

[290] Huan T, Palermo A, Ivanisevic J, et al. Autonomous multimodal metabolomics data integration for comprehensive pathway analysis and systems biology. Anal Chem. 2018;90:8396–8403. doi: 10.1021/acs.analchem.8b00875.

[291] Bruguière A, Derbré S, Dietsch J, et al. MixONat, a software for the dereplication of mixtures based on ^{13}C NMR spectroscopy. Anal Chem. 2020;92:8793–8801. doi: 10.1021/acs.analchem.0c00193.

[292] Yang H, Li J, Hao M, et al. An efficient personalized federated learning approach in heterogeneous environments: a reinforcement learning perspective. Sci Rep. 2024;14:28877. doi: 10.1038/s41598-024-80048-3.

[293] Yurdem B, Kuzlu M, Gullu MK, et al. Federated learning: overview, strategies, applications, tools and future directions. Heliyon. 2024;10:e38137. doi: 10.1016/j.heliyon.2024.e38137.

[294] Sanches PHG, De Melo NC, Porcari AM, et al. Integrating molecular perspectives: strategies for comprehensive multiomics integrative data analysis and machine learning applications in transcriptomics, proteomics, and metabolomics. Biology (Basel). 2024;13:848. doi: 10.3390/biology13110848.

[295] Kang D, Pang X, Lian W, et al. Discovery of VEGFR2 inhibitors by integrating naïve Bayesian classification, molecular docking and drug screening approaches. RSC Adv. 2018;8:5286–5297. doi: 10.1039/c7ra12259d.

[296] Fang J, Yang R, Gao L, et al. Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. J Chem Inf Model. 2013;53:3009–3020. doi: 10.1021/ci400331p.

[297] Shi Y. Support vector regression-based QSAR models for prediction of antioxidant activity of phenolic compounds. Sci Rep. 2021;11:8806. doi: 10.1038/s41598-021-88341-1.

[298] Martínez-Treviño SH, Uc-Cetina V, Fernández-Herrera MA, et al. Prediction of natural product classes using machine learning and ^{13}C NMR spectroscopic data. J Chem Inf Model. 2020;60:3376–3386. doi: 10.1021/acs.jcim.0c00293.

[299] Ferreira LT, Borba JVB, Moreira-Filho JT, et al. QSAR-based virtual screening of natural products database for identification of potent antimalarial hits. Biomolecules. 2021;11:459. doi: 10.3390/biom11030459.

[300] Champati BB, Padhiari BM, Ray A, et al. Application of a multilayer perceptron artificial neural network for the prediction and optimization of the andrographolide content in *Andrographis paniculata*. Molecules. 2022;27:2765. doi: 10.3390/molecules27092765.

[301] Tay DWP, Yeo NZX, Adaikkappan K, et al. 67 Million natural product-like compound database generated via molecular language processing. Sci Data. 2023;10:296. doi: 10.1038/s41597-023-02207-x.

[302] Yeung CS, Beck T, Posma JM. MetaboListem and TABoLiSTM: two deep learning algorithms for metabolite named entity recognition. Metabolites. 2022;12:276. doi: 10.3390/metabo12040276.

[303] Reher R, Kim HW, Zhang C, et al. A convolutional neural network-based approach for the rapid annotation of molecularly diverse natural products. J Am Chem Soc. 2020;142:4114–4120. doi: 10.1021/jacs.9b13786.

[304] Rios-Martinez C, Bhattacharya N, Amini AP, et al. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. PLOS Comput Biol. 2023;19:e1011162. doi: 10.1371/journal.pcbi.1011162.

[305] Wang S, Song X, Zhang Y, et al. MSGNN-DTA: multi-scale topological feature fusion based on graph neural networks for drug–target binding affinity prediction. Int J Mol Sci. 2023;24:8326. doi: 10.3390/ijms24098326.

[306] Shao K, Zhang Y, Wen Y, et al. DTI-HETA: prediction of drug–target interactions based on GCN and GAT on heterogeneous graph. Brief Bioinform. 2022;23:bbac109. doi: 10.1093/bib/bbac109.

[307] Merk D, Grisoni F, Friedrich L, et al. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. Commun Chem. 2018;1:68. doi: 10.1038/s42004-018-0068-1.

[308] Zheng S, Yan X, Gu Q, et al. QBMG: quasi-biogenic molecule generator with deep recurrent neural network. J Cheminform. 2019;11:5. doi: 10.1186/s13321-019-0328-9.

[309] Maziarka Ł, Pocha A, Kaczmarczyk J, et al. Mol-CycleGAN: a generative model for molecular optimization. J Cheminform. 2020;12:2. doi: 10.1186/s13321-019-0404-1.

[310] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci Adv. 2018;4:eaap7885. doi: 10.1126/sciadv.aap7885.

[311] Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics. 2013;29:i126–i134. doi: 10.1093/bioinformatics/btt234.

[312] Aries F, Ito W, Arida FS, et al. Deep belief networks for ligand-based virtual screening of drug design. Proceedings of 2016 6th International Workshop on Computer Science Engineering [Internet]. WCSE; 2016 [cited 2024 Mar 9]. Available from: http://www.wcse.org/WCSE_2016/115.pdf.

[313] Polykovskiy D, Zhebrak A, Vetrov D, et al. Entangled conditional adversarial autoencoder for de novo drug discovery. Mol Pharm. 2018;15:4398–4405. doi: 10.1021/acs.molpharmaceut.8b00839.

[314] Tempke R, Musho T. Autonomous design of new chemical reactions using a variational autoencoder. Commun Chem. 2022;5:40. doi: 10.1038/s42004-022-00647-x.

[315] Ochiai T, Inukai T, Akiyama M, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. Commun Chem. 2023;6:249. doi: 10.1038/s42004-023-01054-6.

[316] Bai X, Yin Y. Exploration and augmentation of pharmacological space via adversarial auto-encoder model for facilitating kinase-centric drug development. J Cheminform. 2021;13:95. doi: 10.1186/s13321-021-00574-4.

[317] Fox Ramos AE, Pavesi C, Litaudon M, et al. CANPA: computer-assisted natural products anticipation. Anal Chem. 2019;91:11247–11252. doi: 10.1021/acs.analchem.9b02216.

[318] Lee S, Van Santen JA, Farzaneh N, et al. NP analyst: an open online platform for compound activity mapping.

ACS Cent Sci. 2022;8:223–234. doi: 10.1021/acscentsci.1c01108.
[319] Ory L, Nazih E-H, Daoud S, et al. Targeting bioactive compounds in natural extracts – development of a comprehensive workflow combining chemical and biological data. Anal Chim Acta. 2019;1070:29–42. doi: 10.1016/j.aca.2019.04.038.
[320] Zulfiqar M, Gadelha L, Steinbeck C, et al. MAW: the reproducible metabolome annotation workflow for untargeted tandem mass spectrometry. J Cheminform. 2023;15:32. doi: 10.1186/s13321-023-00695-y.
[321] Ernst M, Kang KB, Caraballo-Rodríguez AM, et al. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. Metabolites. 2019;9:144. doi: 10.3390/metabo9070144.
[322] Dührkop K, Fleischauer M, Ludwig M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019;16:299–302. doi: 10.1038/s41592-019-0344-8.
[323] Shen X, Wang R, Xiong X, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun. 2019;10:1516. doi: 10.1038/s41467-019-09550-x.
[324] Da Silva RR, Wang M, Nothias L-F, et al. Propagating annotations of molecular networks using *in silico* fragmentation. PLOS Comput Biol. 2018;14:e1006089. doi: 10.1371/journal.pcbi.1006089.
[325] Alden N, Krishnan S, Porokhin V, et al. Biologically consistent annotation of metabolomics data. Anal Chem. 2017;89:13097–13104. doi: 10.1021/acs.analchem.7b02162.
[326] Uppal K, Walker DI, Jones DP. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. Anal Chem. 2017;89:1063–1067. doi: 10.1021/acs.analchem.6b01214.
[327] Guijas C, Montenegro-Burke JR, Domingo-Almenara X, et al. METLIN: a technology platform for identifying knowns and unknowns. Anal Chem. 2018;90:3156–3164. doi: 10.1021/acs.analchem.7b04424.
[328] Xue J, Guijas C, Benton HP, et al. METLIN MS2 molecular standards database: a broad chemical and biological resource. Nat Methods. 2020;17:953–954. doi: 10.1038/s41592-020-0942-5.
[329] Bittremieux W, Avalon NE, Thomas SP, et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. Nat Commun. 2023;14:8488. doi: 10.1038/s41467-023-44035-y.
[330] Senan O, Aguilar-Mogas A, Navarro M, et al. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. Bioinformatics. 2019;35:4089–4097. doi: 10.1093/bioinformatics/btz207.
[331] Gurevich A, Mikheenko A, Shlemov A, et al. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. Nat Microbiol. 2018;3:319–327. doi: 10.1038/s41564-017-0094-2.
[332] Wandy J, Zhu Y, Van Der Hooft JJJ, et al. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. Bioinformatics. 2018;34:317–318. doi: 10.1093/bioinformatics/btx582.
[333] Tsugawa H, Cajka T, Kind T, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015;12:523–526. doi: 10.1038/nmeth.3393.
[334] Ma Y, Kind T, Yang D, et al. MS2Analyzer: a software for small molecule substructure annotations from accurate tandem mass spectra. Anal Chem. 2014;86:10724–10731. doi: 10.1021/ac502818e.
[335] Kunyavskaya O, Mikheenko A, Gurevich A. NPvis: an interactive visualizer of peptidic natural product–MS/MS matches. Metabolites. 2022;12:706. doi: 10.3390/metabo12080706.
[336] Beauxis Y, Genta-Jouve G. MetWork: a web server for natural products anticipation. Bioinformatics. 2019;35:1795–1796. doi: 10.1093/bioinformatics/bty864.
[337] Zani CL, Carroll AR. Database for rapid dereplication of known natural products using data from MS and fast NMR experiments. J Nat Prod. 2017;80:1758–1766. doi: 10.1021/acs.jnatprod.6b01093.
[338] Huan T, Tang C, Li R, et al. MyCompoundID MS/MS search: metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. Anal Chem. 2015;87:10619–10626. doi: 10.1021/acs.analchem.5b03126.
[339] Tsugawa H, Kind T, Nakabayashi R, et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Anal Chem. 2016;88:7946–7958. doi: 10.1021/acs.analchem.6b00770.
[340] Laponogov I, Sadawi N, Galea D, et al. ChemDistiller: an engine for metabolite annotation in mass spectrometry. Bioinformatics. 2018;34:2096–2102. doi: 10.1093/bioinformatics/bty080.
[341] Ridder L, Van Der Hooft JJJ, Verhoeven S. Automatic compound annotation from mass spectrometry data using MAGMa. Mass Spectrom. 2014;3: S0033–S0033. doi: 10.5702/massspectrometry.S0033.
[342] Wang Y, Kora G, Bowen BP, et al. MIDAS: a database-searching algorithm for metabolite identification in metabolomics. Anal Chem. 2014;86:9496–9503. doi: 10.1021/ac5014783.
[343] Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. BMC Bioinf. 2019;20:376. doi: 10.1186/s12859-019-2954-7.
[344] Zhou J, Weber RJM, Allwood JW, et al. HAMMER: automated operation of mass frontier to construct *in silico* mass spectral fragmentation libraries. Bioinformatics. 2014;30:581–583. doi: 10.1093/bioinformatics/btt711.
[345] Wang F, Allen D, Tian S, et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. Nucleic Acids Res. 2022;50: W165–W174. doi: 10.1093/nar/gkac383.
[346] Egan JM, Van Santen JA, Liu DY, et al. Development of an NMR-based platform for the direct structural annotation of complex natural products mixtures. J Nat Prod. 2021;84:1044–1055. doi: 10.1021/acs.jnatprod.0c01076.
[347] Bingol K, Li D-W, Zhang B, et al. Comprehensive metabolite identification strategy using multiple two-dimensional NMR spectra of a complex mixture implemented in the COLMARm web server. Anal Chem. 2016;88:12411–12418. doi: 10.1021/acs.analchem.6b03724.
[348] Gerrard W, Bratholm LA, Packer MJ, et al. IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. Chem Sci. 2020;11:508–515. doi: 10.1039/c9sc03854j.