Artificial intelligence for natural product drug discovery

nature reviews drug discovery

https://doi.org/10.1038/s41573-023-00774-7

Review article

Check for updates

Artificial intelligence for natural product drug discovery

Michael W. Mullowney[1,62], Katherine R. Duncan[2,62], Somayah S. Elsayed[3,62], Neha Garg[4,62], Justin J. J. van der Hooft[5,6,62], Nathaniel I. Martin[7,62], David Meijer[5,62], Barbara R. Terlouw[5,62], Friederike Biermann[5,8,9], Kai Blin[10], Janani Durairaj[11], Marina Gorostiola González[12,13], Eric J. N. Helfrich[8,9], Florian Huber[14], Stefan Leopold-Messer[15], Kohulan Rajan[16], Tristan de Rond[17], Jeffrey A. van Santen[18], Maria Sorokina[19,20], Marcy J. Balunas[21,22], Mehdi A. Beniddir[23], Doris A. van Bergeijk[3], Laura M. Carroll[24], Chase M. Clark[25], Djork-Arné Clevert[26], Chris A. Dejong[27], Chao Du[3], Scarlet Ferrinho[28], Francesca Grisoni[29,30], Albert Hofstetter[31], Willem Jespers[12], Olga V. Kalinina[32,33,34], Satria A. Kautsar[35], Hyunwoo Kim[36], Tiago F. Leao[37], Joleen Masschelein[38,39], Evan R. Rees[25], Raphael Reher[40,41], Daniel Reker[42,43], Philippe Schwaller[44], Marwin Segler[45], Michael A. Skinnider[27,46], Allison S. Walker[47,48], Egon L. Willighagen[49], Barbara Zdrazil[50], Nadine Ziemert[51], Rebecca J. M. Goss[28], Pierre Guyomard[52], Andrea Volkamer[34,53], William H. Gerwick[54], Hyun Uk Kim[55], Rolf Müller[32,56,57,58], Gilles P. van Wezel[3,59], Gerard J. P. van Westen[12] ✉, Anna K. H. Hirsch[32,56,57,58] ✉, Roger G. Linington[18] ✉, Serina L. Robinson[60] ✉ & Marnix H. Medema[5,61] ✉

Abstract

Developments in computational omics technologies have provided new means to access the hidden diversity of natural products, unearthing new potential for drug discovery. In parallel, artificial intelligence approaches such as machine learning have led to exciting developments in the computational drug design field, facilitating biological activity prediction and de novo drug design for molecular targets of interest. Here, we describe current and future synergies between these developments to effectively identify drug candidates from the plethora of molecules produced by nature. We also discuss how to address key challenges in realizing the potential of these synergies, such as the need for high-quality datasets to train deep learning algorithms and appropriate strategies for algorithm validation.

Sections

Introduction

Uses of AI in natural product research

Data sources and data standardization

Conclusions and outlook

A full list of affiliations appears at the end of the paper. ✉e-mail: gerard@lacdr.leidenuniv.nl; anna.hirsch@helmholtz-hips.de; rliningt@sfu.ca; serina.robinson@eawag.ch; marnix.medema@wur.nl

Introduction

Bacteria, fungi, plants and animals produce a wide range of specialized metabolites, also known as natural products. Across the tree of life, these comprise hundreds of thousands of different chemical structures — including peptides, polyketides, saccharides, terpenes and alkaloids — that facilitate an organism's ability to thrive in a particular environment. They have crucial roles in complex inter-organismal interactions, functioning as signals, weapons, nutrient-scavenging agents and stress protectants to mediate competition and collaboration. In the host–microbiome context, specialized metabolites mediate competition and collaboration between microbes and their host.

These natural products have historically been applied with remarkable success as antibiotics, chemotherapeutics, immunosuppressants and crop protection agents. Natural products remain a promising source for the discovery of such drugs based on characteristics such as their relatively high degree of three-dimensionality (as opposed to the often 'flat' synthetic structures), which may be important in modulating challenging drug targets, and their origins as natural metabolites, which makes them likely to be substrates for transporter systems that can enable drugs to reach their targets[1,2].

Although the popularity of natural product discovery programmes in the pharmaceutical industry diminished between roughly 1990 and 2010 owing to the rise of combinatorial chemistry and high-throughput screening[3], there has been a recent renaissance in natural products research in both academia and small biotech start-ups. This renaissance is catalysed by the availability of large-scale omics data, which allows deeper access to the hidden chemical treasure troves of the biosphere. The genes for most specialized metabolite biosynthetic pathways in bacteria and fungi (and some in plants and animals) appear as clusters in the genome of the producing organisms: more than 2,500 of these biosynthetic gene clusters (BGCs) and their products have now been characterized experimentally[4]. This physical clustering has the potential to facilitate the identification of millions of putative biosynthetic pathways for novel molecules through computational genomic analysis[5], which could provide starting points for drug discovery.

In the field studying natural products, artificial intelligence (AI) approaches are now being developed to predict (parts of) chemical structures of BGC products based on DNA sequence alone, fuelled by data on known biosynthetic pathways and their chemical products, which is increasingly standardized and stored in public databases. Although this helps in identifying molecules with new rather than known chemical structures (dereplication) and in linking molecules to their biosynthetic genes[6], there is an urgent need for more effective ways to filter and prioritize the enormous predicted natural product biosynthetic diversity to identify drug leads.

In the field of computational drug design, AI strategies are being developed that may help to address this challenge by providing better understanding of structure–activity relationships and by predicting macromolecular targets for natural products based on their chemical structures. Here, two main approaches are traditionally used: on the one hand, statistical modelling focuses on finding correlations between chemical structure and biological activity, termed quantitative structure–activity relationship (QSAR) modelling; on the other hand, structure-based research attempts to fit 3D chemical structures to protein targets (docking) and subsequently study their behaviour on the nano- to millisecond timescale (molecular dynamics).

For both fields, AI methods have opened up new possibilities in the design, synthesis and biological profiling of existing and new small molecules. Central to these methods are public databases that provide biological activity data for large numbers of (protein) targets and chemical structures. On the basis of chemical similarity, advanced machine learning techniques can use these data to obtain models that are able to predict the potential activity of untested chemical structures within these extensive chemical collections. Moreover, these methods can also be used to systematically analyse large datasets routinely produced from extended molecular dynamics studies and identify hidden patterns in the protein dynamics[7,8]. This has led to exciting successes that have advanced the understanding of the complex interplay between small molecules and protein macromolecules. Examples include new computer-suggested chemical structures (de novo design)[9], drug repurposing through the prediction of unexpected activities and guiding medicinal chemistry approaches to modify and optimize drug molecules for their biological effects (both on and off target)[10].

There is thus great potential for cross-fertilization between the fields of omics-based natural product discovery and computational drug design (Fig. 1). The use of AI could lead to a rapid acceleration of scientific progress in these fields and to a convergence of their methods and directions. For example, scientists have started to apply machine learning — a subfield of AI that generates insights by using algorithms to recognize patterns from data — to the discovery and structural characterization of natural products and to predict relationships between structure and pharmaceutical properties.

However, researchers in these fields have interacted very little so far. In this Review, we present an integrated perspective of a group of scientists from both areas based on an interactive workshop that discussed new ways to connect these research areas and jointly leverage the power of AI to use the vast chemical diversity of the biosphere for the development of new drugs. We first describe applications of AI in natural product research, including genome and metabolome mining, structural characterization of natural products and prediction of the targets and biological activities of natural products. We then discuss a key challenge in realizing the potential of AI in the field — the creation and maintenance of large, high-quality datasets with which to train algorithms — and how this could be addressed. We also consider the pitfalls in training algorithms, such as overfitting, and approaches to avoid them (Box 1).

Uses of AI in natural product research

Natural product genome and metabolome mining

Several AI technologies have been developed to accelerate the discovery of natural products by predicting biosynthetic genes and metabolite structures from sequence or spectral data, respectively. Identifying natural product BGCs still largely relies on rule-based methods such as those used in antiSMASH[11] and PRISM[12]. Although these approaches are successful at detecting known BGC classes, they are less proficient at identifying novel types of BGC or unclustered pathways[13,14]. In these more complex cases, machine learning algorithms have been shown to offer significant advantages over rule-based methods. For example, the hidden Markov model-based method ClusterFinder[15], the deep learning approaches DeepBGC[16], GECCO[17] and SanntiS[18], and several genome mining algorithms for ribosomally synthesized and post-translationally modified peptides (RiPPs)[19–22] each use deep learning or support vector machines to identify BGCs not captured using canonical rule-based annotation approaches. These methods were trained on sequence-based features such as gene families, protein domains and amino acid sequence properties. Although they still have a higher false positive rate than rule-based approaches and also suffer from false negatives for known types of BGC, they have already demonstrated utility in identifying novel classes of natural product

biosynthetic pathways[13]. For example, the decRiPPter algorithm, aimed to predict novel RiPP families, identified pristinin, which belongs to a novel class of lanthipeptides[19] (Fig. 2). In addition, DeepRiPP, thanks to its deep learning-based RiPP precursor detection module, enabled the discovery of the RiPPs deepflavo and deepginsen, whose precursor peptides were encoded distantly from any of their associated biosynthetic enzymes[21].

Whereas genome mining algorithms can hint at biosynthetic potential, metabolomics allows direct detection of biosynthesized components, even if their precise structures are unknown. However, inferring molecular structures and substructures from mass spectrometry (MS) data is far from straightforward. Therefore, AI has been leveraged to target common challenges in MS-based metabolome mining[23], including library matching and searching using mass spectral similarity metrics[24,25], molecular formula annotation[26,27], molecular class annotation[28,29] and retention time prediction[30]. The efficacy of these algorithms is still limited by the relatively small sets of tandem MS (MS/MS) spectra annotated with the fragment ion chemical structures of their corresponding metabolites. However, these algorithms can be enhanced by imputing missing data; for example, by predicting molecular fingerprints or simulated spectra from metabolite structures directly[28]. Similarly, NMR metabolome mining tasks are undergoing transformation[31], as deep learning provides new avenues towards improving NMR spectrum reconstruction, denoising[32], peak picking, *J*-coupling prediction[33] and spectral deconvolution[34].

Ultimately, AI algorithms that link genome-mined BGCs and gene cluster families to untargeted metabolome-mined spectra and predicted molecular classes should be developed. For example, a new deep learning algorithm was recently published that can predict biosynthetic routes from natural product chemical structures, which could provide a basis for matching with BGCs[35]. Such algorithms will help to de-orphan BGCs and molecular structures to address the large annotation gap between genomics and metabolomics. This may allow the combination of sequence and metabolome data to predict metabolite structures synergistically.

Structural characterization of natural products

Successful natural product drug discovery studies require the ability to unambiguously solve the structures of isolated compounds[36]. This task is challenging owing to the chemical complexity of metabolites existing in nature. Structure elucidation requires the collection, analysis and compilation of multiple data types, which may include NMR, infrared (IR), ultraviolet (UV), electronic circular dichroism (ECD) and X-ray spectroscopy, high-resolution MS (HRMS), MS/MS, and experimental and/or computational inspection of the encoded enzymes within the producing BGC[37,38]. Recently, the microcrystal electron diffraction (MicroED) technique, which has the potential to accelerate structure elucidation by allowing analysis of submicron-sized crystals of chemical compounds, was added to this arsenal[39,40].

In general, significant efforts have been made to improve the structural characterization of natural products through methodological, instrumentational and computational means, such as quantum chemistry-based theoretical calculations and AI-based structure predictions from MS and NMR data. Since as early as 1960, AI has been used to complement rule-based approaches in de novo identification of unknown compounds from MS data[41,42]. Subsequently, AI has been used to predict molecular formulae from MS spectra[43], match MS spectra to compounds in molecular databases using deep neural networks[41,43], elucidate structures de novo as SMILES strings from MS/MS spectra[44] and predict chemical properties and identify small molecules from MS[1] and collisional cross section (CCS) data[45].

Similarly, AI has been used to augment NMR-based structure elucidation and annotation. Computer-assisted structure elucidation (CASE) programs[46] reduce erroneous structural assignments by generating a probability-based ranking of all possible structures given an NMR dataset, which can guide structure determination. Examples include the convolutional neural network-based tool SMART 2.0, which guided the discovery and structure elucidation of a novel class of natural products including the new macrolide symplocolide A[47], SMART-Miner[48] and COLMAR[49], which identify and annotate primary metabolites from the NMR spectra of complex mixtures, and DP4-AI, which combines quantum chemistry-based theoretical calculations of NMR shifts with a Bayesian approach that assigns correctness probabilities to candidate structures, and with objective model selection for picking peaks and reducing noise[50,51]. One drawback of quantum chemistry-based theoretical calculations of NMR shifts lies in the need for extensive exploration of a metabolite's conformational space, which is computationally demanding for conformationally flexible molecules. Machine learning models such as ASE-ANI[52] have been developed to address this issue by filtering force field-generated conformations and thus reducing the computational cost.

Predicting targets and biological activity

One of the most important application areas for AI in natural product drug discovery is prediction of the macromolecular targets of the natural products, their associated biological activities and possible toxicities. Accurate predictions of these characteristics will provide direct clues as to which areas of chemical space (Box 2) are most promising for drug discovery. This will be key to the potential success of genome mining, which currently results in lists of candidate BGCs that are too large, with few strategies available to target efforts towards parts of natural product space (Box 2) with actual pharmaceutical potential. AI techniques, in combination with other technologies, can help to address this challenge (Fig. 3).

Natural product target elucidation. The progress of novel natural products towards being selected as drug candidates is often hampered by lack of knowledge about their targets, which impedes their preclinical testing and rational optimization. Given the complexity of metabolite isolation and handling, large-scale experimental determination of mechanisms of action for these molecules is not feasible owing to the costs and effort required. Computational models that rapidly predict the most likely targets from the molecular structure are therefore an area of active research[53]. Virtually all computational drug discovery approaches have been successfully applied to elucidate targets of natural products, including docking[54], clustering[55], bioactivity fingerprints[56], pharmacophores[57] and machine learning[58]. In some cases, this has also led to new insights regarding the mechanisms of action of natural products that were already in clinical trials[59]. Although applicability is currently limited, given this success and the increasing accuracy of advanced machine learning models, we expect further developments in this area that will lead to tailored and further improved models.

Classical cheminformatics- and pharmacophore-based predictions of bioactivity. Methods that rely on the use of classical cheminformatics and computer-assisted drug discovery tools to predict bioactivities for natural products are plentiful[53]. For example, the direct application

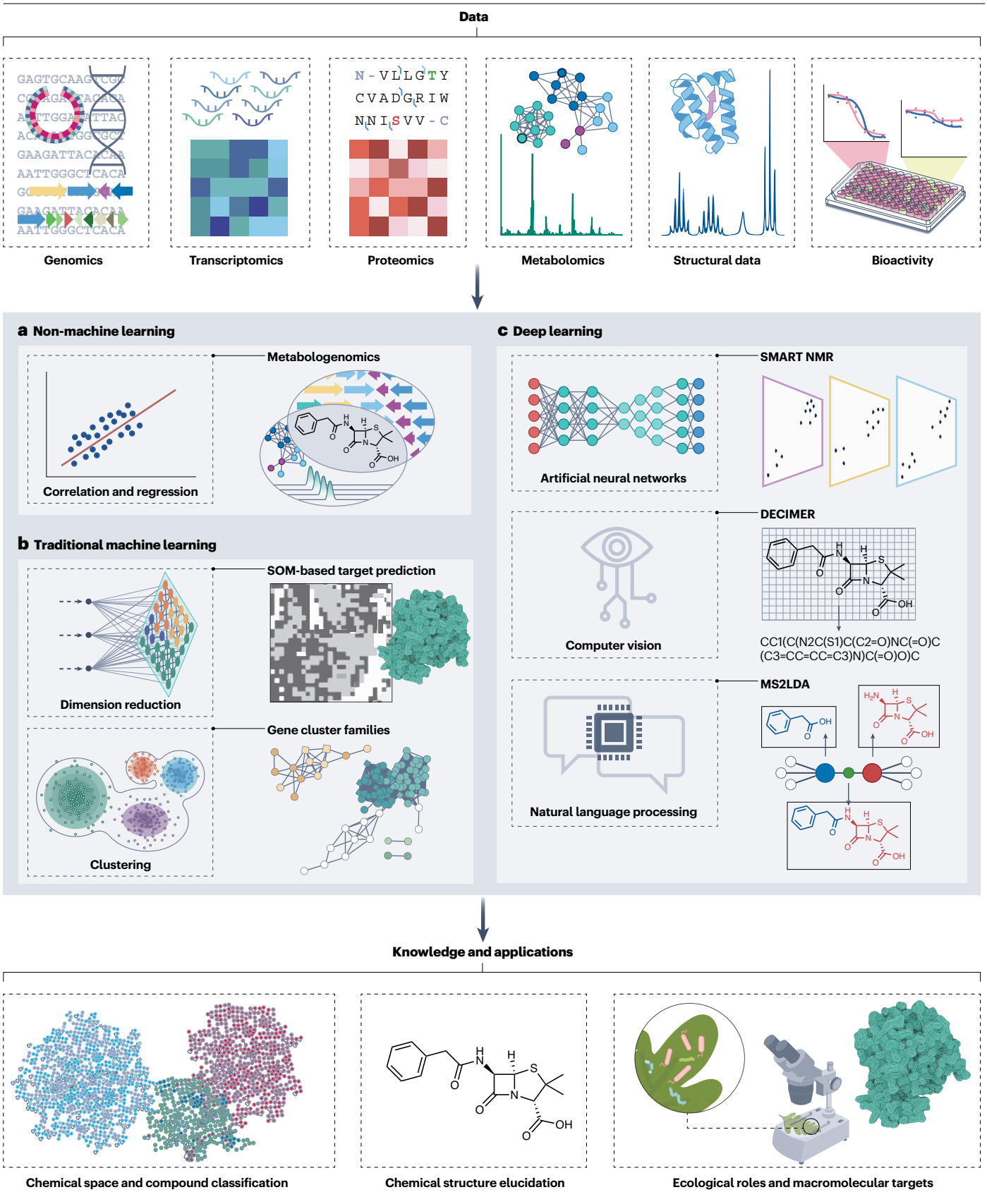

Data
Genomics
Transcriptomics
Proteomics
Metabolomics
Structural data
Bioactivity
a Non-machine learning
Correlation and regression
Metabologenomics
b Traditional machine learning
Dimension reduction
SOM-based target prediction
Clustering
Gene cluster families
c Deep learning
Artificial neural networks
SMART NMR
Computer vision
DECIMER
CC1(C(N2C(S1)C(C2=O)NC(=O)C(C3=CC=CC=C3)N)C(=O)O)C
Natural language processing
MS2LDA
Knowledge and applications
Chemical space and compound classification
Chemical structure elucidation
Ecological roles and macromolecular targets

Fig. 1 | Applications of artificial intelligence in natural product and drug discovery. Classical analyses typically use only a small fraction of datasets of various types, such as genomics, transcriptomics, proteomics, metabolomics, structural data and bioactivity data. Artificial intelligence (AI) methods can help to integrate different data types to learn complex feature relationships and develop meaningful hypotheses. AI methods that can have a key role in natural product drug discovery include, but are not limited to: non-machine learning methods (part **a**) such as correlation and regression (for example, linking metabolomic and genomic data[190]); traditional machine learning methods (part **b**), such as self-organizing maps (SOMs) (for example, for macromolecular target prediction[221]) and clustering (for example, grouping gene cluster families[222]); and deep learning (part **c**), such as convolutional neural networks (for example, for chemical structure elucidation[47]), computer vision (for example, automatic chemical image recognition[166]) and natural language processing (for example, topic modelling for chemical substructure exploration and annotation[223]).

of the ensemble-based popular prediction methods PASS[60] and SEA[61] to natural products have shown some successes. Given the distinct chemical structures and physicochemical properties of natural products[55,62], the most successful applications use additional preprocessing steps or rely on chemical descriptions and representations that are agnostic to the chemical differences between natural products and the training data of synthetic compounds. For example, the SPiDER method, based on self-organizing maps, was specifically developed to predict the bioactivities of molecules and has been successfully applied to predict the biological activity of macrocyclic natural products[55,62] and fragment-like natural products[57].

Other successful applications of bioactivity predictions have used representations such as 3D pharmacophore matching[57] of bioactivity signatures coupled to deep neural networks[63,64]. A notable approach consists of constructing learned representations using the deep learning-based chemprop message-passing neural network[65]. Such models capture essential properties of molecules without directly using classic chemical fingerprints and have enabled the prediction of the bactericidal activity of the synthetic chemical compounds halicin[64] and abaucin[66], as well as eight additional molecules with antibiotic properties structurally distinct from known antibiotic classes[64] (Fig. 2).

Molecular dynamics simulations and structure-based prediction of bioactivity. Structure-based approaches use spatial information about a protein target to predict a compound's binding mode. This information can be obtained from experimentally determined structures (for example, with X-ray crystallography) or via deep learning-based modelling approaches such as AlphaFold[67]. Then, potential binding modes can be enumerated via strategies such as molecular docking with protein dynamics accounted for via molecular dynamics approaches. These methods are computationally expensive, but have been taking advantage of both hardware (graphics processing unit (GPU) computing) and software improvements[68]. Structure-based methods can provide a wealth of information; for example, the applicability and use of the free-energy perturbation (FEP) method has recently increased substantially in academic and industrial drug discovery projects[69]. Molecular docking, molecular dynamics and FEP could be extended to study affinities of natural products.

Sequence- or BGC-based predictions of bioactivity. A growing number of approaches have been used to predict bioactivities based on DNA and/or protein sequence data from BGCs with machine learning[12,70,71], and other strategies have the potential to do so in the near future.

One approach that leverages knowledge of existing small molecules is to predict the final product of a BGC and infer its activity from this prediction directly, as exemplified by PRISM[12]. One issue with this method is the challenge faced in predicting activities for BGCs with poorly predicted structures, where even small mistakes in the final prediction could yield vastly different activities for the real compound. As substructure prediction is more robust, use of discrete substructural features such as β-lactam rings or specific amino acids may produce more accurate results for a broader range of BGCs.

Alternative approaches emerging for bioactivity prediction draw on the field of natural language processing (NLP). NLP-based methods such as word2vec[72], originally developed for context-aware embedding of words within sentences in text documents, have been extended to embed protein domains within BGCs using pfam2vec[16]. DeepBGC, a de novo BGC prediction tool[16], represents predicted BGCs using pfam2vec-derived features from protein domains; these features are then supplied to a random forest classifier to predict natural product activity. Building on the DeepBGC framework, Deep-BGCpred implements dual-model serial screening and a 'sliding window' strategy for more accurate BGC boundary detection[71]. Just as NLP has revolutionized other fields, we expect continued, rapid advances in applications of NLP for BGC and bioactivity prediction.

Of note, the sequence boundaries for BGCs predicted by mining tools are not precise, often missing portions of the BGC or fusing them with others. To use BGC sequence data as input for machine learning, it is generally necessary for an expert to manually update the BGC boundaries. Improvements in BGC prediction will therefore be vital for such bioactivity prediction methods and remain an area where further research is needed.

Bioactivity predictions based on self-resistance, regulatory or evolutionary features. Bacteria have long been known to harbour resistance genes that enable them to withstand the effects of antibiotic natural products that they themselves produce[73]. Numerous antimicrobial resistance determinant databases are available, such as the Comprehensive Antibiotic Resistance Database (CARD)[74], a National Database of Antibiotic Resistant Organisms (NDARO) and ResFinder[75]. To leverage resistance information, various algorithms have been created to attempt to link these resistance genes with BGCs, as the resistance genes are necessary to confer immunity in the host[76,77]. A recent study incorporated both general protein domains and resistance genes to create a more robust feature set; this method proved accurate when sufficient training data were available, such as for antibacterial prediction in bacterial BGCs[70].

As an additional layer of biological information, transcription factor networks and their cognate regulatory elements can be used to classify BGCs on the basis of how they are controlled and to which (environmental) signals they respond. The EvoMining framework[78] is based on the concept that streptomycetes adapt to their ecological niche by evolving their primary and secondary metabolism in response to their environment[79]. Regulatory networks that control BGCs and the cognate signals that unlock their biosynthesis may provide key information on the function of the natural products they specify. Regulatory networks have so far been largely ignored in genome mining approaches but may well be a key determinant for biological understanding and function

Box 1

Standard practices for evaluating a machine learning model

'Garbage in, garbage out' is a well-known concept in machine learning that is intuitive to understand, but without proper model validation it can be challenging to identify the true predictive power of a model. There are two key points to keep in mind when assessing a model: data balancing and model evaluation on an independent test set.

Data balancing

Datasets that are used for machine learning are usually not homogeneous. Imbalance can exist in multiple ways that lead to incorrect model evaluation.

- Over-representation of one or more data labels. Consider a binary classification problem for drug–target interaction with a dataset of 10,000 positive and 100 negative data points. Without addressing this imbalance before training, the model will likely always predict an interaction between drug and target regardless of the input. The model will be correct 99% of the time even though it has no predictive power.
- Over-representation of one or more data features. This is a very common imbalance in biological data: some species and molecule types have been researched far more extensively than others, leading to datasets with an over-representation of certain sequences or molecular structures. Models trained on such data without consideration for this type of imbalance usually seem to perform very well, as they make good predictions for sequences or molecules from over-represented phylogenetic branches or compound classes. Poor predictions on under-represented clades often go unnoticed: either the few mispredictions in the independent test set form such a small proportion of the total tested data points that they do not affect the average performance much; or worse, the under-represented clades do not appear in the test set at all.

These data imbalances have to be targeted at three stages of model development.

- Data selection for training and test sets before model training. For each type of data label and data feature, data points should first be filtered for duplicates or near-duplicates and subsequently be divided proportionally across training and test sets. For sequence data, pre-filtering could mean selecting one representative of a phylogenetic clade and excluding the rest; for compound data, one could cluster based on chemical similarity and include only one member for each cluster. This avoids (near)-duplicates in training and test sets that would yield an automatic correct prediction. Proportional division of the resulting data points across training and test sets based on class and feature labels (for example, 80% training and 20% test for each label) ensures that the model can be separately evaluated on each data subclass, resulting in more accurate model evaluation.
- Sampling and data weighting during model training. When a model is not instructed otherwise, it will prioritize overall accuracy. Often, this means that the model tolerates mispredictions for under-represented data classes. To prevent this, data can be weighted during model training: under-represented classes should receive higher weights or contribute more towards a model's loss function such that the model penalizes prediction errors for those classes more than prediction errors for over-represented classes. Alternatively, it is possible to undersample or oversample the dataset to artificially reduce or expand the dataset such that each data class is proportionally represented. Both approaches result in models that should be more generally applicable and less biased towards over-represented data labels or features.
- Class-specific model evaluation after model training. To evaluate how the model performs for each data subclass, regardless of how many data points belong to that class, it is important to assess predictive power for each class separately. This can be done for data labels with true or false positive or negative rates, and for data features by assessing performance for each sequence or compound cluster.

Cross-validation and independent test sets

Usually, machine learning algorithms are not trained just once: developers have to play around with input features, model parameters and model types before they find a model that works. A frequent inaccuracy in this process is that the same test set is often used for evaluation of these in-between models and for the evaluation of the final model. At this point, the test set is no longer truly independent, as decisions that influence model performance have been made based on the test set. Thus, overfitting of the model may remain unnoticed this way. Therefore, it is crucial to hold out an independent test set before any training and only use this test set to assess the model's performance at the very end of development. Monitoring model performance during development can be done by selecting a validation set from training data or by doing cross-validation with all training data. Optimally, multiple runs should be performed with a representative standard deviation to be able to statistically test observed improvements for significance. When selecting (cross-)validation sets, it is equally important to take into account data imbalance.

prediction. Whereas BGCs predict what types of metabolite may be produced, regulatory networks can be harnessed to estimate how BGCs are controlled and — notably — in response to which signals. This information may serve as a beacon to find BGCs or metabolites required for specific purposes, such as responses to stress or disease. This could, for example, be used to predict which gene clusters are expressed in mutualist microbes in response to pathogen invasion, which may help to prioritize BGCs for antibiotic discovery.

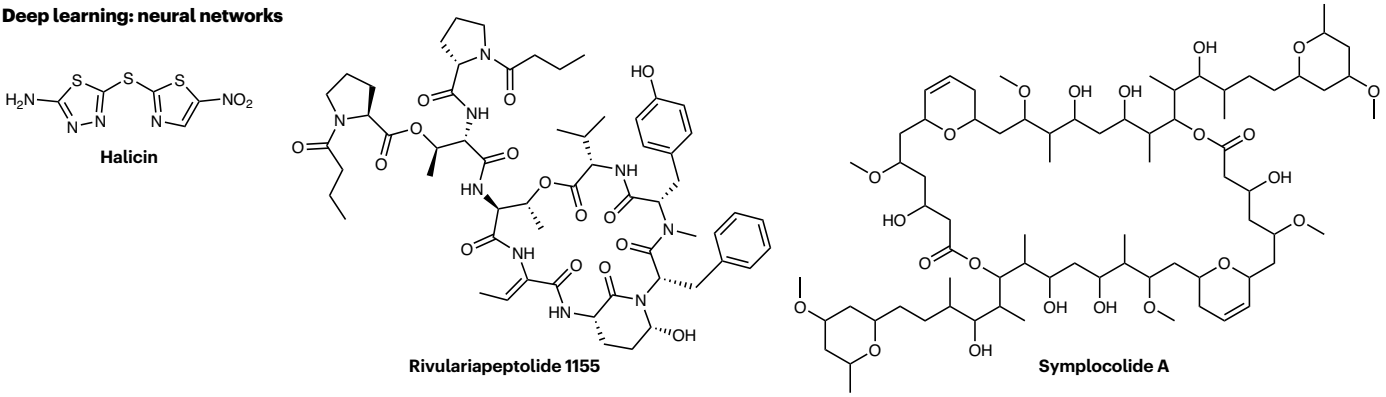
Emerging AI methods in natural product drug discovery

In all of the application areas mentioned above, AI technology is still in its infancy and suffers from a lack of (high-quality) standardized data.

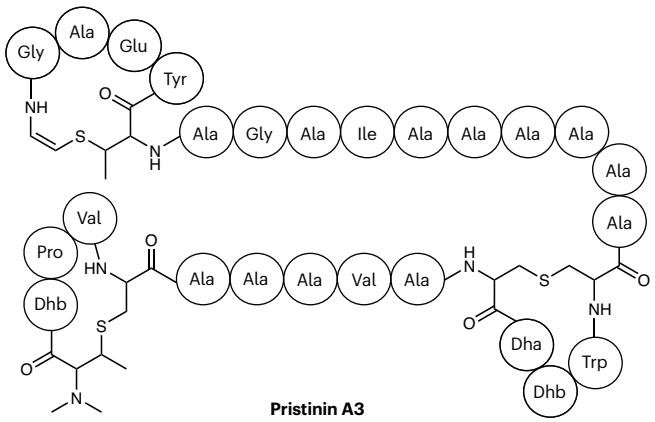
However, refined approaches for building machine learning models using sparse or variable training set data are being developed, and new (often community-driven) initiatives to curate or generate high-quality datasets are starting to emerge. Together, these advances suggest that major improvements in AI methodological accuracy are within reach. Below, we discuss algorithmic developments that could have a significant impact and then consider data generation and standardization challenges that will need to be addressed to exploit the full potential of these algorithms.

Molecular featurization methods. Complex molecular data are made machine readable through featurization, and the extent to which the most important information in a dataset can be captured concisely is crucial for the success of machine learning algorithms (Fig. 4). Simplification is inherent to featurization. In rare cases, this can lead to clashes whereby two or more molecules are represented by the same fingerprint. Hence, a featurization technique that aligns with the goal of the use should be carefully chosen.

Deep learning: neural networks



Supervised machine learning: support vector machine



Natural language processing

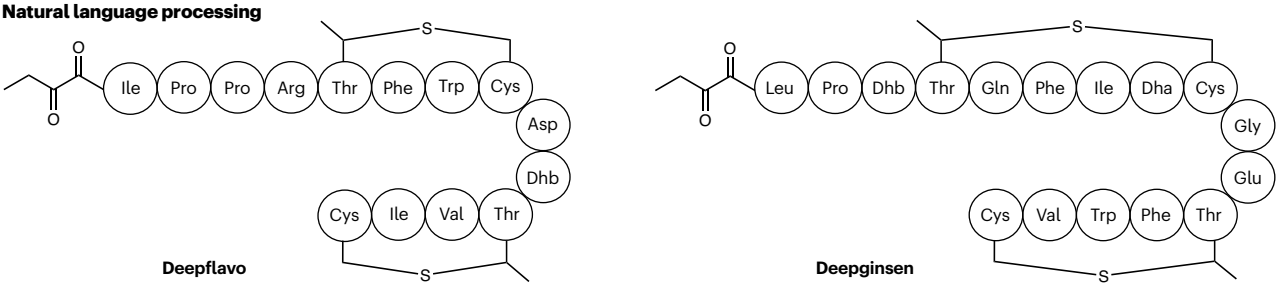


Fig. 2 | Example compounds discovered using artificial intelligence approaches. The synthetic compound halicin and related molecules were discovered using a deep neural network trained to predict antibiotic activity from chemical structure[64]. The structures of the rivulariapeptolides and symplocolide A were predicted from complex microbial extracts using a convolutional neural network[28,152]. Pristinin A3 was discovered using a support vector machine that mines pangenomes to prioritize novel ribosomally synthesized and post-translationally modified peptide (RiPP) precursors within operon-like structures in the accessory genome of a genus[19]. Deepflavo and deepginsen were discovered in part using natural language processing to predict their RiPP precursors and their cleavage patterns from genomes[21].

The most ubiquitous method for featurizing a molecule is to convert its molecular structure into a sequence of bits or counts[80]. Algorithms to create such fingerprints are readily implemented in cheminformatic software packages such as RDKit (see Related links) and the Chemistry Development Kit[81]; however, molecule features can be manually determined as well[82].

Circular fingerprints have enabled the most accurate identification of structurally related natural products[83–86]. However, circular fingerprints were found to be less useful than pharmacophore-based descriptors for scaffold hopping from natural products to synthetic mimetics[87]. Other recent examples are MAP4 fingerprints, which combine substructure and atom-pair concepts and can be used to distinguish bacterial from fungal natural products[88,89]. Also, features created from short molecular dynamics simulations can be used to accurately predict partition coefficients, solvation free energies or even ligand affinity[90–94]. Recent approaches to 'k-merize' 3D shapes[95], which can be sampled from molecule conformers, may also provide promise for fingerprinting, as they may take into account the 3D shape of molecules. Conversely, compound features that do not describe the compound structure at all can also be helpful, as exemplified by bioactivity fingerprints[63,96–99].

Deep learning. A diverse array of AI algorithms have been developed over the past decade, many of which have been successfully applied to natural product research (Fig. 1). One machine learning technology that has recently received considerable attention and application is deep learning. Deep learning has the flexibility to capture nonlinear relationships and to accept non-tabular input that extends the applicability of AI for natural product computational research to non-Euclidean domains[100,101]. Deep learning for molecular function prediction on molecular graphs sometimes outperforms simpler machine learning models on circular fingerprints[65], although this seems to vary between datasets and applications[102,103]. Furthermore, explainable AI methods have been shown to improve interpretability of such deep learning models[104,105]; for example, in the assessment of preclinical relevance[106] and for pharmacophore and toxicophore identification[107,108].

Box 2

Visualizing and navigating chemical space

Chemical space — typically defined by using multiple compound properties of interest, such as physicochemical properties — is vast and largely unexplored[224]. Just 'drug-like' chemical space, composed of all compounds that comply with Lipinski's 'rule-of-five' guidelines for oral bioavailability[225], has been estimated to encompass ~10^{60} compounds, and even the largest chemical libraries used for computational screening usually encompass only ~10^{10} compounds. Importantly for the context of this article, however, the study that underlies Lipinski's rule[225] identified natural products as common exceptions, and the chemical features of natural products and typical compounds in the screening libraries of pharmaceutical companies differ. These library compounds are often planar, synthetic small molecules that comply with Lipinski's rules, with mass <500 Da, whereas natural products typically have greater size and 3D complexity.

Exploring chemical space is a daunting task, not only because of the sheer quantity of compounds that can be (virtually) enumerated, but also because the description and labelling of compounds is by definition a multidimensional problem. For visualization purposes, a high-dimensional space will be reduced to only two or three dimensions. Also, depending on the properties of interest, the chemical space to be explored will be constructed differently. Still, given that most of chemical space is unexplored, taking the challenge of solving the multiparameter optimization problem to navigate chemical space is considered a promising strategy for identifying novel drug candidates[226–228].

A common way to reduce dimensionalities to navigate chemical space is via principal component analysis (PCA). PCA of chemical properties has revealed that both drug molecules and natural products occupy a very similar topological diversity distribution, which was not the case for combinatorial compounds[229]. Another method is *t*-distributed stochastic neighbour embedding (*t*-SNE), which has been used successfully for the design of new drug classes, for example, new kinase inhibitors[226]. A recent development to *t*-SNE is the uniform manifold approximation and projection (UMAP) algorithm, which is less computationally expensive than the previous approach and can therefore be applied to larger datasets[227]. More recently, a Tree MAP (TMAP) algorithm was developed to visualize data sets with sample sizes up to around 10^7 in a tree layout[228]. In this article, using TMAP, a tree of all the compounds in the ChEMBL database (1.13 million) with their associated biological assay data was constructed within 10 min.

The application of unsupervised learning approaches (such as PCA, *t*-SNE, UMAP and TMAP) to reduce dimensionalities in chemical space data can be used to infer the likely biological activity of compounds and ultimately identify new scaffolds. This approach has proved successful in the small-molecule discovery field, and we believe its application to natural products will open up new avenues to characterize and address, among others, biological activity and pharmacokinetic properties. It would be exciting to implement the newly developed dimensionality reduction tools, with their improved computational capabilities, in mapping both natural product and small molecules, to identify overlapping chemical space and ultimately transfer knowledge between the two fields.

A starting point could be the merging of the large Papyrus database on drug-like molecules with existing natural product databases[230]. Molecular standardization of Papyrus could be applied to the natural product databases to determine whether additional rules or procedures are required. The resulting database could be used as a dataset to apply existing visualization and dimensionality reduction methods. A subsequent challenge is that the application of these methods should be validated using known synthetic molecules and natural products.

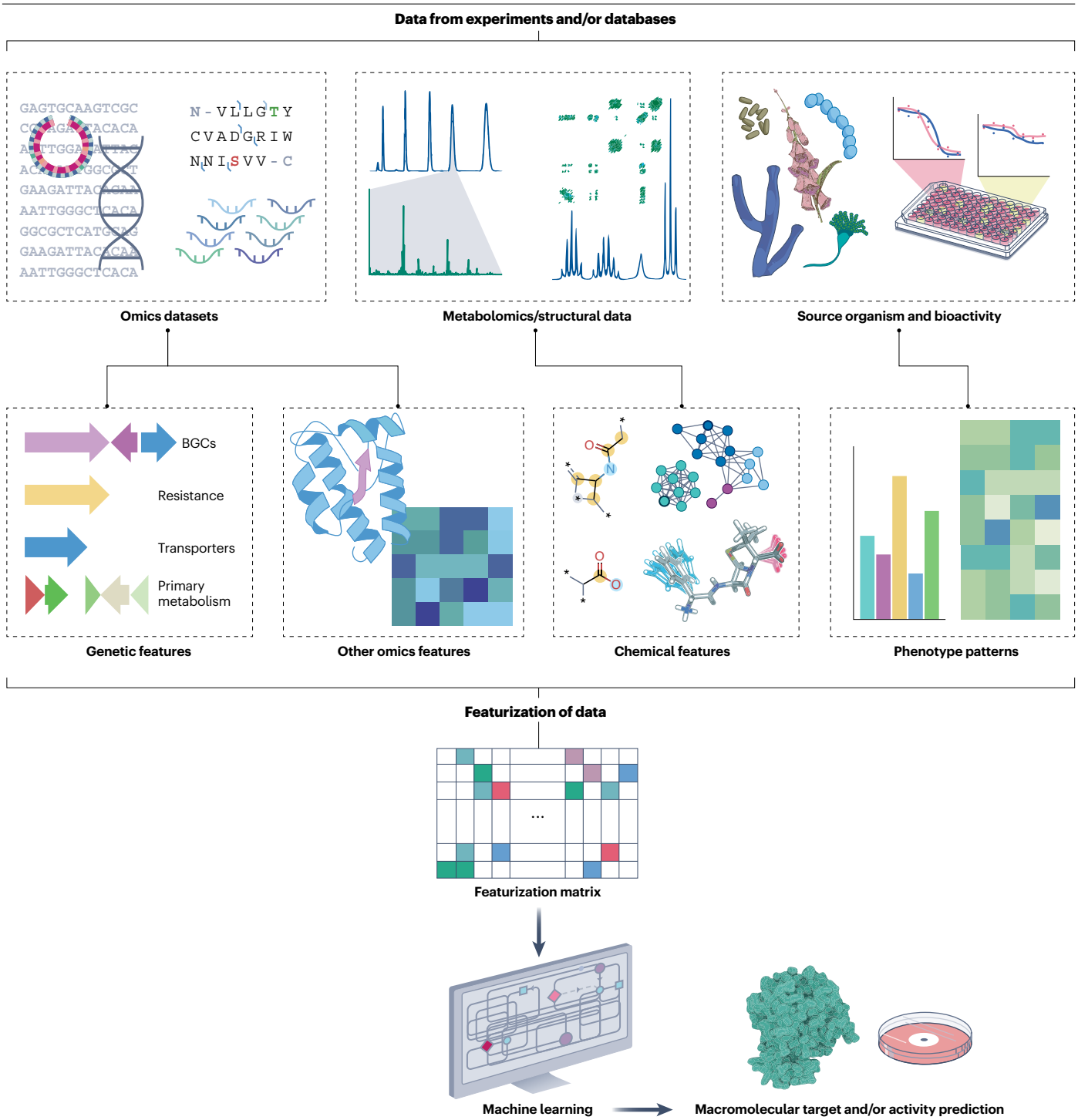


Fig. 3 | Predicting biological activities and macromolecular targets from genomic, metabolomic and phenotypic data. Omics datasets can be mined to identify genetic features of natural product biosynthetic pathways, such as resistance genes, transporters and links with primary metabolism, which are predictive of the biological activity or macromolecular target of the products of the pathway. Metabolomics and NMR (in concert with analysis of biosynthetic genes) can be used to identify chemical features of metabolites that are predictive of certain activities or targets. Finally, large-scale standardized phenotypic bioassays are key. There is considerable potential for artificial intelligence approaches to then predict targets and activities based on combined sets of genetic and chemical features of natural products and their biosynthetic pathways. BGC, biosynthetic gene cluster.

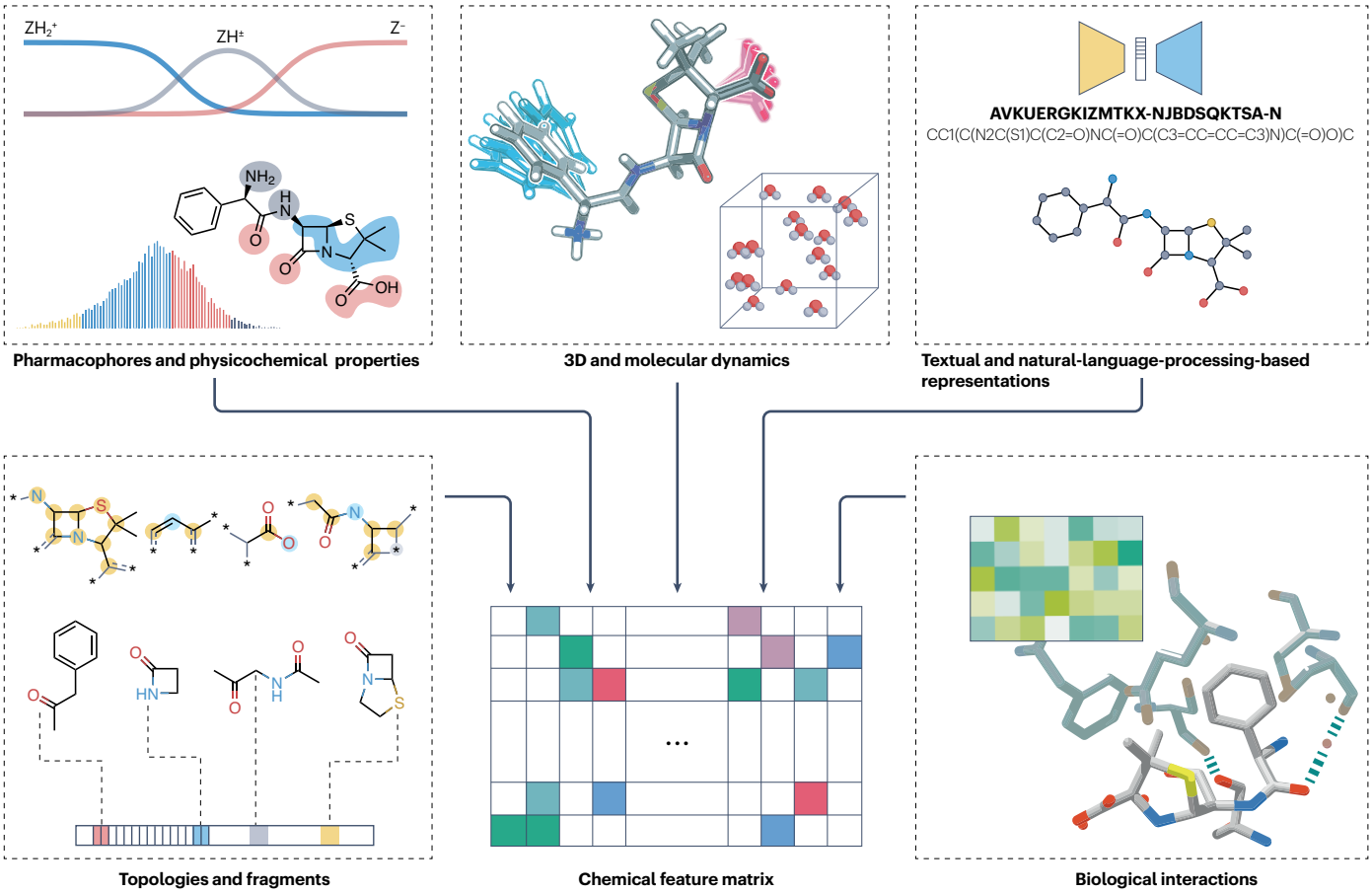


Fig. 4 | Chemical featurization techniques. Numerous featurization technologies are available to encode chemical information in a manner that machine learning techniques can process. These technologies range from simple physicochemical properties, via commonly used circular fingerprints, to advanced 3D and neural net-based encoders. Use of an appropriate featurization method is key, as the interpretation of a machine learning model is based on the features on which this model is trained. Although possible, combinations of featurization techniques are not common.

Applications of deep learning include molecular graph neural network approaches[109–112]; for instance, for predicting drug–target binding affinity[113], SMILES-based approaches for de novo drug-like molecule generation[114,115], graph-based de novo molecular generation[116], and property prediction[117,118] and surface mesh-based approaches for protein pocket-conditioned molecular representations[119]. Moreover, encoder–decoder architectures are used to featurize compounds for virtual screening from different input formats[120–122]. A comprehensive overview of deep learning molecular representations, which can be applied to molecular structure data in natural product research, is provided in ref. 123.

One of the most notable deep learning approaches of past years is AlphaFold[67], which can predict the 3D structure of proteins from their primary amino acid sequence by learning from the entire corpus of the Protein Data Bank. Since the landmark breakthrough by AlphaFold, accurate modelling approaches building on this work continue to raise the bar[124] by tackling challenges such as multimeric structure prediction[125]. For natural product research, structural prediction is highly relevant, as it can, for example, help to predict the substrate specificities across natural product biosynthetic enzyme families or help to predict the evolution of drug resistance by target modification. The precedent set by AlphaFold suggests that deep learning has the potential to solve long-standing problems in natural product computational research, although natural product data are currently much sparser.

As deep learning for natural product computational research is still in its infancy, caution should be applied to its predictions[126,127]. To build trust and use the full potential of deep learning, we believe a set of best practices needs to be established for using deep learning techniques in natural product research[128,129].

- Compare the performance of new deep learning models with simpler models to validate and motivate the trade-off between interpretability and prediction results[130–134].
- Clarify the scope in which the model optimally performs by defining its applicability domain and adding confidence estimates to predictions[135,136].
- Evaluate the model through cross-validation and use of a true hold-out set, avoiding a random splitting approach with a preference for chemical clustering or temporal splitting[131], and, if applicable, including prospective experiments. Owing to the practice of publishing synthetic compounds as chemical analogues with a structure–activity relationship, random splitting for validation overestimates the ability of models to generalize.

Therefore, chemical clustering or temporal splitting is essential to truly validate created models[131].

- Understand the results of a new model. If allowed by the chosen method, map what the algorithm learned back to input features and provide proper visualizations that allow interpretation of results for bench scientists[106,108,137].

Deep learning algorithms will definitely not always be the most suitable tools[138]. Nonetheless, we do expect that they will become increasingly useful to address challenges such as structure elucidation and activity prediction as datasets in compatible formats grow.

Approaches to address data limitations. One of the biggest challenges for deep learning in natural product research is open access to large curated datasets, which is discussed in the next section. 'Data-hungry' algorithms such as deep learning will only improve performance if training datasets are sufficient to support model complexity. One solution to reduce the number of effectively required data points is to use weights from pre-trained models on larger chemical datasets. Using pre-validated and pre-trained chemical models such as ChemBERTa[139] or MoleculeNet[102] reduces the computational load required to train new models from scratch. In many cases, pre-trained models will also yield higher prediction accuracies[140].

Although deep learning techniques can overcome issues of incomplete sample labelling and small datasets, semi-supervised learning (combining labelled with unlabelled data) can assist with learning on datasets with incomplete labelling[141,142]. This has been applied in the past, for example, to improve substrate specificity predictions of natural product biosynthetic enzymes using transductive support vector machines, where this helped to map the shape of unlabelled sequence space to better know how queries would relate to labelled data points[143]. An alternative is transfer learning[144], a strategy in which knowledge from a task learned on an extensive dataset can then be transferred to a related task for which fewer data are available. This can improve model efficiency and mitigate issues relating to low-data regimes[145], for example, in de novo molecular design[146–148].

Active learning techniques, which guide the selection of unlabelled data for labelling through experimentation, can also be deployed when labelled training data are limited[149]. This has been successfully applied to identify small molecules that inhibit the protein–protein interaction between the anticancer target CXC chemokine receptor 4 and its ligand by actively retrieving informative active compounds that continuously improved the adaptive structure–activity model[150]. Multiple practical challenges remain before active learning can be broadly deployed[149], many of which revolve around the time requirements and cost of standardized experimental data acquisition. This might explain why active learning has not yet been broadly deployed in natural product research, where experiments are commonly complex. For example, CANOPUS[28], a deep neural network-based structure class annotation tool that is based on MS spectra, uses other AI tools including ClassyFire[151] and NPClassifier[29] to label data and thus train the network. This enabled the structural elucidation of the novel rivulariapeptolide protease inhibitors from complex mixtures[28,152]. With increasing experimental resolution and automation, we believe that active learning will play a central part in future natural product research.

Similarly, reinforcement learning, which steers the output of a machine learning algorithm towards user-defined regions of optimality via a predefined (computational) reward function, has shown promise in de novo design towards attractive regions of chemical space[153–155], for rule-based organic chemistry and for retrosynthesis prediction[156–159].

Data sources and data standardization

High-quality training datasets are crucial to the success of AI algorithms. Unstructured datasets (for example, unannotated MS data) can be used for unsupervised learning applications such as dimensionality reduction and bioactivity prediction. By contrast, supervised learning requires training data that are both accurately annotated and of sufficient scope to answer the question being addressed. This is a particular challenge for natural products applications in which the breadth of chemical space is high but the coverage of most published datasets is low. Data augmentation and synthetic data generation, although valuable techniques, should be carried out with care to avoid the accumulation of bias. In addition, data error is a challenge in the field. Heterogeneous biological public data generated in many labs tends to provide multiple sources of error that can hamper highly sensitive deep learning methods[160,161]. Integrating data from different datasets and ensuring that annotation methods are consistent is therefore a major bottleneck for the development of training sets for machine learning. In this section, we explore the characteristics and attributes necessary to create high-quality datasets to advance natural product discovery, including discussion of the current state of natural product databases (Table 1) and data dissemination, the need for data standardization, annotation and integration and the creation of training sets.

The natural product database landscape

The landscape of natural product databases is large and diverse, but is also highly fragmented, and it currently contains few comprehensive and well-curated data resources[162]. Unfortunately, natural product-related data are often under-represented or not annotated as natural products in large generalist databases (such as PubChem, ChEMBL, Reaxys and Scifinder); for example, as of January 2023, only 8,951 natural products have a ChEMBL identifier according to Wikidata (see Related links). Additionally, documentation of data sources, acquisition and changes — known as data provenance — is not well maintained in most natural product databases. For example, literature citations or information on source organisms and associated BGCs may be missing. Furthermore, although some databases (such as ChEMBL[163] and BindingDB[164]) include bioassay data for pure compounds, very few include bioassay data for natural product extracts and fractions. Finally, some natural product databases lack options for full data download, or are not licensed for open use by academic groups. Together, these issues severely limit the availability of amenable datasets to train AI models.

Challenges with natural product data dissemination

Literature curation. Scientific publication remains the dominant mechanism for disseminating new natural product information. Unfortunately, automated data extraction from natural product journals is often impossible because data are not in machine-readable formats, despite the existence of simple solutions such as compact identifiers[165]. Database completeness is also hampered by the broad spectrum of journals that feature natural product research, including many journals that are not natural products specific.

Consequently, database developers must manually curate articles to convert them into structured data formats. Curation difficulties include image-to-structure conversion, absence of core data

Table 1 | Databases for natural product data

Resource name	Chemical identifiers	Chemical structures	Documented entries	Is NP-specific	Has an API	Full dump available	Notable experimental data	Notable calculated data	Has version control and archive available to download	Has a user submission system for new data upload	Licence
Chemical-specific resources											
LOTUS[203]	Yes	Yes	Yes	All NPs	Yes	Yes	Producer taxonomy	Molecular descriptors, chemical classification, bioactivities	Yes	No	CC0
COCONUT[202]	Yes	Yes	Yes	All NPs	Yes	Yes	None	Molecular descriptors, chemical classification, bioactivities	Yes	No	CC BY-SA
Natural Products Atlas[170,171]	Yes	Yes	Yes	Microbial NPs	Yes	Yes	Producer taxonomy	Chemical classification	Yes	Yes	CC BY
BGC resources											
MIBiG[4,169]	Yes	Yes	Yes	Microbial NPs	Yes	Yes	BGC genomic coordinates and gene function annotation; compound produced by BGC	antiSMASH annotations	Yes	Yes	CC BY
antiSMASH database[211]	Yes	No	No	Microbial NPs	Yes	Yes	None	BGC genomic coordinates and gene function annotations; compounds produced by BGC	Yes	No	CC BY
PRISM gold standard BGCs[12]	No	No	No	Microbial NPs	No	Yes	Genomic coordinates and gene function annotation; compound produced by BGC	BGC genomic coordinates and gene function annotations; compounds produced by BGC	Yes	No	CC BY
Spectral resources											
GNPS[172]	No	Yes	Yes	Yes	Yes	Yes	–	–	No	Yes	CC0
MassBank[212]	Yes	Yes	Yes	No	No	Yes	MS and tandem MS spectra	None	No	Yes	CC BY-NC
NP-MRD[173]	Yes	Yes	Yes	Yes	No	Yes	NMR	No	Yes	Yes	CC BY
CH-NMR-NP	Yes[a]	Yes	Yes	All NPs	No	No	NMR, producer	Molecular weight	No	No	–
MetaboLights[213]	Yes	Yes	Yes	No	No	Yes	MS and tandem MS spectra; NMR	None	No	Yes	EMBL-EBI's Terms of use
Paired Omics Data Platform[191]	No	No	Yes	Yes	No	Yes	LC–MS, genomics	None	No	Yes	CC BY
nmrshiftdb2[214]	No	Yes	Yes	No	Yes	No	NMR	Calculated NMR	No	Yes	Modified CC BY

Table 1 (continued) | Databases for natural product data

Resource name	Chemical identifiers	Chemical structures	Documented entries	Is NP-specific	Has an API	Full dump available	Notable experimental data	Notable calculated data	Has version control and archive available to download	Has a user submission system for new data upload	Licence
NP-friendly useful resources											
ZINC20[215]	Yes	Yes	No	No	Yes	No	None	Molecular descriptors, bioactivities	No	Yes	CC0
ChEBI[216]	Yes	Yes	Yes	No	Yes	Yes	None	Chemical classification, bioactivities	Yes	Yes	CC BY
ChEMBL[163]	Yes	Yes	Yes	No	Yes	Yes	Bioactivities	Molecular descriptors	Yes	Yes	CC BY-SA
WikiPathways[217]	Yes	No	Yes	No	Yes	Yes	Metabolic networks	None	Yes	Yes	CC0
Reactome[218]	Yes	No	Yes	No	Yes	Yes	Metabolic networks	Metabolic networks	Yes	No	CC0
CO-ADD[219]	Yes	Yes	No	No	No	Yes	Bioactivities	None	No	Yes	The University of Queensland (2016)
Wikidata[220]	Yes	Yes	Yes	No	Yes	Yes	None	None	Yes	Yes	CC0

API, application programming interface; BCG, biosynthetic gene cluster; GNPS, Global Natural Product Social Molecular Networking; LC, liquid chromatography; MIBiG, Minimum Information about a Biosynthetic Gene cluster; MS, mass spectrometry; NP, natural product. [a]CAS registry number.

(for example, BGC sequence), resolution of name conflicts (multiple structures with the same name, or structures with multiple names) and extraction of data and metadata for biological assays. Improvements are underway for structure recognition from images using DECIMER 1.0 (refs. 166,167) and through new formats for reporting of chemical structure data[168]. Nevertheless, high-quality digitization of research data into structured open formats remains an unsolved challenge. This is further complicated by the byzantine and overly restrictive copyright rules that currently govern journal articles. Finally, because most natural product databases focus on only one feature of natural product data, there is presently high redundancy in curation efforts, as the existence of minor variations in the extracted data (for example, structure standardization methods or character encodings for compound names) may interfere with linking records between databases.

One solution to this issue would be to encourage authors to include a standardized machine-readable file for each compound described in the paper, similar to the cif file required for each X-ray structure. This machine-readable file could contain crucial information about each structure (for example, SMILES, compound name, availability and location of spectral data, source organism and BGC) and would offer a central point of reference for data dissemination and automated database importation by natural product-centric resources.

Data deposition. Several of the larger natural product data repositories, including Minimum Information about a Biosynthetic Gene cluster (MIBiG)[169], the Natural Products Atlas (NP Atlas)[170,171], Global Natural Product Social Molecular Networking (GNPS)[172], Natural Products Magnetic Resonance Database (NP-MRD)[173] and Norine[174] offer mechanisms to accept user-deposited data (Fig. 5). However, without clear incentives to deposit data, deposition rates are low. In addition, managing the infrastructure for data depositions (interactive web page construction, database version control, authentication management and database security) and curating and correcting errors is complicated and time consuming, and often beyond the capacity of academic database developers.

The extensive and often manual data entry requirements for journal article submission lead to 'deposition fatigue' for authors. The varied natural product-related data types (such as source organisms, MS, NMR, BGC and SMILES) amplify this, and increase the number of platforms that users must navigate to deposit raw data in open repositories. The community must therefore develop mechanisms to streamline, incentivize and reward data and metadata deposition, such as with the development of a centralized venue for pre-publication data deposition that can disseminate these data to speciality databases (Fig. 5).

Two principal avenues exist to incentivize data deposition to public repositories: 'value added' and 'requirements'. First, authorships during data 'curatathons', increased citations, opportunities for collaboration and facilitated automated re-analysis are very beneficial for depositors[175]. An example of added value is ReDU, which aids in rapid re-analysis of existing and future data[176] through subscription to one's own and public datasets[172]. Alternatively, repositories can offer validation reports, quality metrics, prevalence statistics (for example, the statistics page of MIBiG[4] that facilitates cross-species comparisons of biosynthetic potential) and other feedback on data to depositors that provides a tangible and immediate benefit to deposition. We acknowledge that, at present, data deposition is usually a long process that requires submitters to fill in as much metadata as possible following ontologies or controlled vocabularies. These

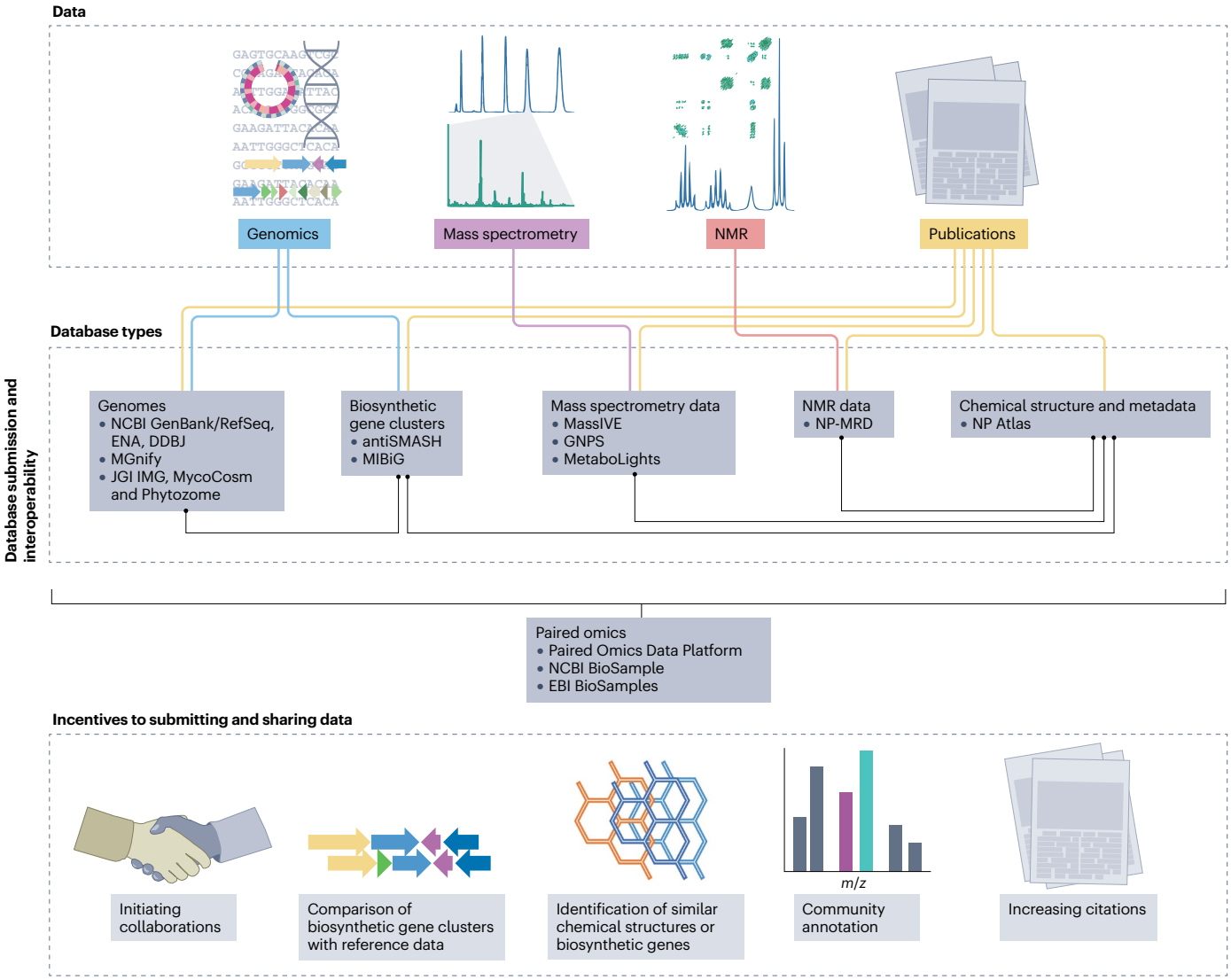


Fig. 5 | Depositing and sharing natural product data: infrastructure and incentives. Diverse types of data on the structures, biological activities and biosynthetic pathways of natural products can be deposited into dedicated community databases, allowing their reuse as well as providing training data for artificial intelligence (AI) algorithms. As standardized deposition of such data will be crucial for the future of AI-driven natural product drug discovery, it will be important to provide the scientific community with clear incentives and rewards to submit and share their data. This includes opportunities for collaboration, online (comparative) analysis capabilities linked to these databases, community-driven annotation and knowledge build-up and increasing impact through follow-up work and the citations that result from this. GNPS, Global Natural Product Social Molecular Networking; NP Atlas, Natural Products Atlas; NP-MRD, Natural Products Magnetic Resonance Database.

extended processes should become more user friendly; for example, by including an autofill during metadata reporting, using tools that automatically generate entries from well-defined ontologies and automated emails to authors with filtered web-crawled data that authors can complete and send into relevant repositories.

Second, journals and/or funding agencies can mandate data deposition, eliminating the need for incentives. An excellent example of this is a recent announcement that the *Journal of Natural Products* requires the deposition of raw NMR data starting in July 2023 (ref. 177). Regardless of the motivation, promoting community-driven data deposition is indispensable to making the natural products field AI compatible.

The need for data standardization

The foundation of high-quality datasets begins with experimental design and practice, the key being consistency. Currently, the most extensive, high-quality natural product-related datasets in the public domain have been generated by a few laboratories. Typically, however, the value of these datasets is limited owing to the lack of sample diversity and the limited number of data types available for a single study. Furthermore, even if appropriate controls and replication are used, there can be fundamental differences in the quality and quantity of detected features for the same sample set, as demonstrated for intra-laboratory liquid chromatography (LC)–MS/MS analyses[178]. As a result, a global

assemblage of data would be incredibly valuable; yet challenges exist of poor interoperability (that is, connecting data between resources) and weak compatibility (that is, resources use different standards and ontologies to annotate and identify their contents).

It is important to note that the quality of biologically derived data (for example, MS resolution and/or accuracy, gene-sequencing depth and/or error rate) should be defined in light of the desired outcome. The metabolomics field, for example, has initiated the Metabolomics Standards Initiative[179], which describes key parameters to report to facilitate quality assessment. Often, AI tasks rely on having a large corpus of data to train and/or search (for example, clustering MS/MS spectra[180] and binning metagenomes[181]). One challenge with this requirement is that experimental datasets may contain only a single or very few representatives in each class, limiting their value for model building. Dedicating the effort to creating comprehensive training sets is an essential step for the field as it looks to embrace AI technologies.

To achieve standardization, a key focus must be the interoperability between existing natural product databases. At present, most database managers communicate updates on an ad hoc basis. In addition, some databases such as NP Atlas maintain interoperable application programming interfaces (APIs) to enable regular, automatic data crawls between resources. However, this becomes exceedingly complex if databases operate in a continuously updating fashion, mainly if resources use varied data standardization strategies, such as PubChem versus ChEMBL structure standardization protocols.

Besides specific, persistent identifiers, data interoperability requires common languages (that is, controlled vocabulary). Open standards have an essential role here, defining exchange formats, vocabularies and ontologies, and experimental protocols. For example, they could facilitate accurate description and reporting of the structural characterization of natural products[182]. Furthermore, the adoption of universal spectrum identifiers (USIs) to identify mass spectra in proteomics[183] and metabolomics[184] showcases standardization tools, enabling data analysis across datasets. Such tools have a pivotal role in enabling large-scale studies by structuring omics data and represent an area of development that the natural product community should consider. The implementation of semantic web approaches is also an essential step forwards, which standardizes how we disseminate knowledge and data and integrate exchange formats, linking between resources and ontological representation[185]. An overview of current natural product ontologies is provided in Table 2.

The need for standardization in describing bioactivities of natural products and ensuring that experimental conditions are comparable between laboratories is apparent. Although standards exist for reporting the biological activities of purified compounds (for example, ChEMBL[163], PubChem[186], Supernatural II[187] and NPASS[188]), such standardization does not extend to microbial crude extracts and fractions. In addition, metadata such as extract preparation methods can substantially impact bioactivity data, yet they are rarely recorded in natural product databases. Finally, as further discussed below, experimental conditions must be described as accurately as possible, with scientists preferably using the same growth conditions for their experiments. Overall, although it is clear that the move towards FAIR (findable, accessible, interoperable and reusable) data and metadata is happening in natural product research, many depositions still fail to include all required components.

The need for data annotation

In addition to essential metadata (such as sample taxonomy, extract preparation protocol and instrument parameters), the addition of contextual annotations can greatly increase the value of natural product datasets. For example, accurate annotation of compound structures to metabolomics datasets would provide many opportunities to build machine learning models that integrate structural and biological and/or genomic data.

However, creation of annotated datasets faces two significant hurdles. The first is that most datasets can be annotated in many different ways, making it unrealistic to aggregate annotations from different studies into a single monolithic training set. Secondly, most annotation methods include elements of bias and false assignment that will influence model structure and accuracy. Therefore, although dataset annotation by subject experts is very valuable for AI developers, the creation and adoption of annotation standards for core information types should be seen as a priority for the field.

The need for data integration

The value of linked or paired data. As omics technologies mature, there is an increasing need for data integration between platforms. This is relevant to the development of AI models because some questions can be answered only by considering data from multiple data types. For example, large-scale integration of NMR spectra and MS fragmentation data could dramatically affect the accuracy and coverage of automated compound identification platforms.

Integration of natural product data involves two core activities: the pairing of datasets for analysis, such as that of the paired omics data platform, or the linking of raw or processed data across data types, such as the peptidogenomics, glycogenomics, metabologenomics or NPLinker platforms[189–196]. In the first case, the objective is to define which data types exist for each sample, whereas in the second case, the goal is to perform paired analyses whereby both data types are mined at the same time[197]. An example of this combined data approach is the integration of enzyme-constrained models and omics analysis of *Streptomyces coelicolor* to reveal metabolic and genetic changes that enhance heterologous production[198]. Transcriptomics has also been used as a constraint to improve the statistical association of BGCs from genome data to metabolites in metabolome data by identifying which BGCs are in fact expressed under the conditions in which certain metabolite features are observed[199].

Methodology and opportunities for data integration. Data integration faces several current challenges that are mostly centred around inter-dependencies of the data types and the various data formats that need to 'talk' to each other. Fortunately, early tools such as NPLinker[190], GraphOmics[200] and anvi'o[201] are starting to overcome some of these challenges. However, the number of tools available that facilitate and ease the analysis and interpretation of linked data is currently very limited, with users still needing considerable expertise to interpret the results. Furthermore, overparameterization of models is a risk when linking two or multiple datasets. For example, the same information can be present in more than one data type; it is then essential to effectively correct for that to avoid bias. Another bottleneck is getting the data in the appropriate format so it can be used by AI algorithms. Standardization remains the main issue here, particularly in areas such as metabolomics where the data are inherently heterogeneous owing to the nature of the samples.

The fields of genomics, proteomics and transcriptomics have all developed excellent community standards that have encouraged data standardization. Outstanding challenges with separating and identifying individual components from complex mixtures have hampered

Table 2 | Recommended ontologies and controlled vocabularies for natural product research

Ontology name	Focus	Description
Biology		
Plant Ontology (PO)	Controlled vocabulary, formats, standards	Structured description of terms to plant anatomy, morphology and growth and development to plant genomics data
BRENDA Tissue Ontology (BTO)	Controlled vocabulary, formats	Structured description for enzyme sources: tissues, cell lines, cell types and cell cultures
Gene Ontology (GO)	Controlled vocabulary, formats, standards	Framework and set of concepts for describing the functions of gene products
PIERO Enzyme Reaction Ontology	Controlled vocabulary, standards	Description of partial reaction characteristics of enzymatic reactions
Phenotype And Trait Ontology (PATO)	Controlled vocabulary, formats	Description of phenotypic qualities: properties, attributes and characteristics
NCBI Taxonomy (NCBITAXON)	Controlled vocabulary	NCBI organismal taxonomy
BioAssay Ontology (BAO)	Controlled vocabulary, formats, standards	Description of the biological screening assays
Chemistry		
ChEBI	Controlled vocabulary, chemical classes, standards	Structured classification of 'small' chemical compounds of biological interest
NPClassifier Ontology	Semantic vocabulary and categories in natural products	Structured description of terms for secondary metabolism in natural products
ChemOnt (from ClassyFire)	Controlled vocabulary, formats	Structured description of terms by extracting common or existing chemical classification category terms from the scientific literature and available chemical databases
Chemical Information Ontology (CHEMINF)	Controlled vocabulary, formats	Terminology for the descriptors commonly used in cheminformatics software applications and algorithms
Chemical Methods Ontology	Controlled vocabulary	Description of the methods and instruments used to collect data in chemical experiments
Reaction Ontology (RXNO)	Controlled vocabulary	Reaction-name ontology
Omics		
Experimental Factor Ontology (EFO)	Controlled vocabulary, formats	Systematic description of many experimental variables available in the EBI databases
Metabolomics Standards Initiative Ontology (MSIO)	Controlled vocabulary, formats, standards	Application ontology for supporting description and annotation of mass spectrometry and NMR spectroscopy-based metabolomics experiments and fluxomics studies
Sequence types and features ontology (SO)	Controlled vocabulary, formats	Structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases
The RNA Ontology (RNAO)	Controlled vocabulary	Controlled vocabulary pertaining to RNA function and based on RNA sequences, secondary and 3D structures
GENO ontology	Controlled vocabulary, formats, standards	OWL model for genotypes, their sequence components and links to corresponding biological and experimental entities
PRIDE Controlled Vocabulary	Controlled vocabulary, formats, standards	Ontology for PRIDE (proteomics identifications), a centralized, standards-compliant, public data repository for proteomics data
Medical/biomedical		
Ontology for Biomedical Investigations (OBI)	Controlled vocabulary, formats, standards	Description of biomedical investigations: study design, protocols, instrumentation, data and analyses
The Drug Ontology (DRON)	Controlled vocabulary	Ontology for drugs, containing ingredients, mechanisms of action, physiological effects and therapeutic intent
Antibiotic Resistance Ontology (ARO)	Controlled vocabulary	Description of antibiotic resistance genes and their mutations
Integration		
Semanticscience Integrated Ontology (SIO)	Controlled vocabulary	Integrated ontology of types and relations for rich description of objects, processes and their attributes
Unit Ontology (UO)	Controlled vocabulary	Standardized description of units of measurements
Citation Typing Ontology (CiTO)	Controlled vocabulary	Description of the nature of reference citations in scientific research articles and other scholarly works

similar standardization efforts in metabolomics. This is particularly true for the field of natural products, where the range of possible compounds from any source organism can number in the thousands and where many of the structures remain to be discovered. The wide range of sources, processing methods, chromatographic separation conditions and analytical approaches all combine to make data standardization particularly difficult in this area.

Training sets for AI models and benchmarking

Requirements for high-quality training sets. Machine-readable data are essential for the creation of training sets for AI models. Although the data have often already been collected, they are either converted into an unstandardized written form within publications or not reported at all. Furthermore, well-curated and consistent metadata are also key to training successful models. Indeed, data can be of variable quality owing to inherent differences, for example, in analytical equipment used; however, when this is documented well, researchers can select the relevant data for AI.

Examples of existing natural product-based training and benchmarking sets. Chemical structure and biosynthetic data for natural products are now reasonably well standardized and centralized. For example, the NP Atlas[170,171], COCONUT[202] and LOTUS[203] databases provide information about chemical structures, and the MIBiG database contains information on BGCs[169]. These resources have been applied as training datasets for a wide array of machine learning applications, including the prediction of natural product-likeness of molecules[204], de novo BGC predictions[16,17], matching of chemical structures to their mass spectra[205], automated chemical classification of natural product structures[29] and the identification of unknown metabolites from NMR spectral matching[47].

Using USIs for mass spectra will enable easy standardized access to the mass spectral data for natural products, including the underlying raw data. In this regard, spectral databases for natural products are under active development, such as the GNPS for MS and MS/MS data and the NP-MRD for NMR data. Importantly, entries in MIBiG, GNPS and the NP-MRD are now all cross-linked to the NP Atlas, creating a central hub that connects structural, spectroscopic and biosynthetic data for natural products.

By contrast, two areas that lack natural product database coverage are catalytic activities of biosynthetic tailoring enzymes (key to predicting natural product structures) and biological activities (key to understanding structure–activity and structure–property relationships). In the former case, the absence of well-curated data for tailoring enzymes limits our ability to predict core structures and their modifications from BGC data. In the second case, the absence of well-standardized bioactivity training sets prevents us from predicting potential target space for newly discovered natural products, or natural product structures predicted from bioinformatic tools. Together, these two issues limit our ability to deliver on the promise offered by massively parallel whole-genome sequencing and large-scale discovery and annotation of BGCs.

Although well-curated training sets for chemical structures and BGCs increasingly meet the demands for creating AI models, almost no high-quality datasets exist for benchmarking the performance of AI models in genome mining (sequence quality dependent) or MS data (instrument parameter dependent). As a result, various datasets are currently used for performance comparisons, making it difficult to reliably establish how well a novel algorithm truly outperforms its predecessor.

Opportunities for generating standardized data sets: the case of biological activities. Data on biological activities and modes of action of natural products perhaps constitute the most crucial type of data to guide future natural product drug discovery. At the same time, these data are currently the least standardized and systematically documented. Although databases such as ChEMBL[163] can host such data, stored using standardized ontologies[206,207], the vast majority of natural product activity data are never deposited and can only be found in the text or supplementary materials of manuscripts. Additionally, the protocols by which activity data have been generated are highly diverse, which further frustrates the direct comparison of datasets generated in different laboratories. A unified effort for data standardization also calls for using standardized growth media and culturing conditions. For example, the International *Streptomyces* Project (ISP) media have been designed with this in mind. The media can be ordered from the same source, allowing direct comparison of growth conditions. Negative data for molecules not showing activity (equally important for machine learning purposes) are mostly not reported at all, leading to large biases in the primary literature. Populating biological activity databases with targeted standardized datasets and culture conditions would be highly beneficial. Some efforts already do exist that generate specific types of data. For example, the NCI60 panel of tumour cell lines for anticancer drug screening has existed for years, and molecules can be sent to the US National Cancer Institute to be subjected to this panel[208]. Similarly, CO-ADD constitutes a community-driven approach to antibiotic discovery[209], allowing compounds to be sent to a central location to test their activities according to standardized protocols.

Conclusions and outlook

In summary, progress in AI for natural product drug discovery is primarily limited by a shortage of large, high-quality datasets rather than a lack of innovative algorithms. As a general recommendation for the field, we caution against using new algorithms solely for their 'hype' factor. Instead of jumping on the bandwagon of the latest AI trend, we advise carefully considering which algorithms are best suited for the type and quantity of data available; the fact that natural product datasets are generally considerably smaller than generic computer vision-related datasets, for example, may mean that simpler models with fewer parameters may be more successful and less likely to suffer from overfitting; also in AI, Occam's razor is more relevant than ever.

That said, breakthroughs in the field have been made by crossing disciplinary boundaries to draw on algorithms from other fields, such as NLP. Algorithmic advances are especially needed to extract meaningful features from heterogeneous data sources with multiple inputs, including chemical spectra, DNA sequences, structures and bioactivity information. Another opportunity for the field is to adopt an 'active learning' approach towards dataset generation. By this, we mean characterizing underexplored areas of sequence, chemical, structural or bioactivity space in which gold-standard datasets are lacking to increase the number of effective data points. It is also important to recognize that AI approaches will generally not be able to predict entirely novel chemistry, mechanisms of actions that have never been observed before or completely new catalytic activities of enzymes. Investments in fundamental biochemical research are needed, to shed light on those parts of biochemical space for which AI currently does not yet provide meaningful insights[210].

New data-driven AI discoveries depend on underlying databases being preserved and maintained over time. Ironically, although AI is entirely reliant on high-quality data, longitudinal and stable financial

support for the maintenance of databases is challenging to obtain. Therefore, for future AI advances, we feel that continued support for database maintenance and interoperability should be a priority for international and national funding agencies. Because of the vast array of data types associated with natural product research it is unlikely that a single monolithic repository will serve the needs of the natural product community. Instead, specialized repositories that focus on different aspects of natural product data (such as structures, BGCs, spectral data and biological activities) must focus on improving interoperability to develop a distributed network of data resources. This interoperability not only must involve the connection of entries between databases but also must consider integrated data deposition and the adoption of common standardization protocols for core data types. There is much to learn about repository structure and governance strategies from other areas of science, such as the Protein Data Bank for structural biology and the Cambridge Structural Database for X-ray crystallography. The natural product community must prioritize and promote these efforts if they are to benefit from the new and exciting applications being offered by AI-based technologies.

Finally, we emphasize that the collective resources of our global scientific community far outweigh the capacity of any single lab. If appropriate incentives and guidelines are available, community-generated and curated datasets can have enormous potential to advance the field of AI-driven natural product drug discovery.

Published online: 11 September 2023

References

1. Dobson, P. D., Patel, Y. & Kell, D. B. 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov. Today* **14**, 31–40 (2009).
2. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).
3. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nat. Rev. Drug. Discov.* **4**, 206–220 (2005).
4. Terlouw, B. R. et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2023).
5. Gavriilidou, A. et al. Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nat. Microbiol.* **7**, 726–735 (2022).
6. van der Hooft, J. J. J. et al. Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).
7. Doerr, S. et al. TorchMD: a deep learning framework for molecular simulations. *J. Chem. Theory Comput.* **17**, 2355–2363 (2021).
8. Rodríguez-Espigares, I. et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
9. Liu, X., IJzerman, A. P. & van Westen, G. J. P. Computational approaches for de novo drug design: past, present, and future. *Methods Mol. Biol.* **2190**, 139–165 (2021).
10. Choudhury, C., Arul Murugan, N. & Priyakumar, U. D. Structure-based drug repurposing: traditional and advanced AI/ML-aided methods. *Drug Discov. Today* **27**, 1847–1861 (2022).
11. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
12. Skinnider, M. A. et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
13. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
14. Medema, M. H., de Rond, T. & Moore, B. S. Mining genomes to illuminate the specialized chemistry of life. *Nat. Rev. Genet.* **22**, 553–571 (2021).
15. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
16. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
17. Carroll, L. M. et al. Accurate de novo identification of biosynthetic gene clusters with GECCO. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.03.442509 (2021).
18. Sanchez, S. et al. Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.23.540769 (2023).
19. Kloosterman, A. M. et al. Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lanthipeptides. *PLoS Biol.* **18**, e3001026 (2020).
20. de Los Santos, E. L. C. NeuRiPP: neural network identification of RiPP precursor peptides. *Sci. Rep.* **9**, 13406 (2019).
21. Merwin, N. J. et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl Acad. Sci. USA* **117**, 371–380 (2020).
22. Tietz, J. I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
23. Louwen, J. J. R. & van der Hooft, J. J. J. Comprehensive large-scale integrative analysis of omics data to accelerate specialized metabolite discovery. *mSystems* **6**, e0072621 (2021).
24. Huber, F. et al. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **17**, e1008724 (2021).
25. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminform.* **13**, 84 (2021).
26. Ludwig, M. et al. Databse-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nat. Mach. Intell.* **2**, 629–641 (2020).
27. Hoffmann, M. A. et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat. Biotechnol.* **40**, 411–421 (2022).
28. Dührkop, K. et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2021).
29. Kim, H. W. et al. NPClassifier: a deep neural network-based structural classification tool for natural products. *J. Nat. Prod.* **84**, 2795–2807 (2021).
30. Aalizadeh, R., Nika, M.-C. & Thomaidis, N. S. Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. *J. Hazard. Mater.* **363**, 277–285 (2019).
31. Chen, D., Wang, Z., Guo, D., Orekhov, V. & Qu, X. Review and prospect: deep learning in nuclear magnetic resonance spectroscopy. *Chemistry* **26**, 10391–10401 (2020).
32. Wu, K. et al. Improvement in signal-to-noise ratio of liquid-state NMR spectroscopy via a deep neural network DN-unet. *Anal. Chem.* **93**, 1377–1382 (2021).
33. Ito, K., Xu, X. & Kikuchi, J. Improved prediction of carbonless NMR spectra by the machine learning of theoretical and fragment descriptors for environmental mixture analysis. *Anal. Chem.* **93**, 6901–6906 (2021).
34. Li, D.-W., Hansen, A. L., Yuan, C., Bruschweiler-Li, L. & Brüschweiler, R. DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.* **12**, 5229 (2021).
35. Zheng, S. et al. Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat. Commun.* **13**, 3342 (2022).
36. Milanowski, D. J. et al. Unequivocal determination of caulamidines A and B: application and validation of new tools in the structure elucidation tool box. *Chem. Sci.* **9**, 307–314 (2018).
37. Audoin, C. et al. Metabolome consistency: additional parazoanthines from the mediterranean zoanthid parazoanthus axinellae. *Metabolites* **4**, 421–432 (2014).
38. Fox Ramos, A. E. et al. CANPA: computer-assisted natural products anticipation. *Anal. Chem.* **91**, 11247–11252 (2019).
39. Jones, C. G. et al. The CryoEM method MicroED as a powerful tool for small molecule structure determination. *ACS Cent. Sci.* **4**, 1587–1592 (2018).
40. Kim, L. J. et al. Prospecting for natural products by genome mining and microcrystal electron diffraction. *Nat. Chem. Biol.* **17**, 872–877 (2021).
41. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:fingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
42. Lindsay, R. K. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project* (McGraw-Hill, 1980).
43. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
44. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **19**, 865–870 (2022).
45. Colby, S. M., Nuñez, J. R., Hodas, N. O., Corley, C. D. & Renslow, R. R. Deep learning to generate chemical property libraries and candidate molecules for small molecule identification in complex samples. *Anal. Chem.* **92**, 1720–1729 (2020).
46. Burns, D. C., Mazzola, E. P. & Reynolds, W. F. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. *Nat. Prod. Rep.* **36**, 919–933 (2019).
47. Reher, R. et al. A convolutional neural network-based approach for the rapid annotation of molecularly diverse natural products. *J. Am. Chem. Soc.* **142**, 4114–4120 (2020).
48. Kim, H. W., Zhang, C., Cottrell, G. W. & Gerwick, W. H. SMART-Miner: a convolutional neural network-based metabolite identification from ^{1}H-^{13}C HSQC spectra. *Magn. Reson. Chem.* **60**, 1070–1075 (2022).
49. Wang, C. et al. COLMAR lipids web server and ultrahigh-resolution methods for two-dimensional nuclear magnetic resonance- and mass spectrometry-based lipidomics. *J. Proteome Res.* **19**, 1674–1683 (2020).
50. Smith, S. G. & Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J. Am. Chem. Soc.* **132**, 12946–12959 (2010).
51. Howarth, A., Ermanis, K. & Goodman, J. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem. Sci.* **11**, 4351–4359 (2020).
52. Das, S., Edison, A. S. & Merz, K. M. Jr. Metabolite structure assignment using in silico NMR techniques. *Anal. Chem.* **92**, 10412–10419 (2020).
53. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).

54. Lanz, J. & Riedl, R. Merging allosteric and active site binding motifs: de novo generation of target selectivity and potency via natural-product-derived fragments. *ChemMedChem* **10**, 451–454 (2015).
55. Reker, D. et al. Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **6**, 1072–1078 (2014).
56. Wassermann, A. M. et al. A screening pattern recognition method finds new and divergent targets for drugs and natural products. *ACS Chem. Biol.* **9**, 1622–1631 (2014).
57. Rollinger, J. M., Hornick, A., Langer, T., Stuppner, H. & Prast, H. Acetylcholinesterase inhibitory activity of scopolin and scopoletin discovered by virtual screening of natural products. *J. Med. Chem.* **47**, 6248–6254 (2004).
58. Reker, D. et al. Machine learning uncovers food- and excipient-drug interactions. *Cell Rep.* **30**, 3710–3716.e4 (2020).
59. Conde, J. et al. Allosteric antagonist modulation of TRPV2 by piperlongumine impairs glioblastoma progression. *ACS Cent. Sci.* **7**, 868–881 (2021).
60. Lagunin, A., Filimonov, D. & Poroikov, V. Multi-targeted natural products evaluation based on biological activity prediction with PASS. *Curr. Pharm. Des.* **16**, 1703–1717 (2010).
61. Sá, M. S. et al. Antimalarial activity of physalins B, D, F, and G. *J. Nat. Prod.* **74**, 2269–2272 (2011).
62. Schneider, G. et al. Deorphaning the macromolecular targets of the natural anticancer compound doliculide. *Angew. Chem. Int. Ed. Engl.* **55**, 12408–12411 (2016).
63. Bertoni, M. et al. Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **12**, 3932 (2021).
64. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **181**, 475–483 (2020).
65. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
66. Liu, G. et al. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-023-01349-8 (2023).
67. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
68. Pandey, M. et al. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.* **4**, 211–221 (2022).
69. Schindler, C. E. M. et al. Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **60**, 5457–5474 (2020).
70. Walker, A. S. & Clardy, J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.* **61**, 2560–2571 (2021).
71. Yang, Z. et al. Deep-BGCpred: a unified deep learning genome-mining framework for biosynthetic gene cluster prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.15.468547 (2021).
72. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv*. https://doi.org/10.48550/ARXIV.1301.3781 (2013).
73. Thaker, M. N. et al. Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nat. Biotechnol.* **31**, 922–927 (2013).
74. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).
75. Bortolaia, V. et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **75**, 3491–3500 (2020).
76. Mungan, M. D. et al. ARTS 2.0: feature updates and expansion of the antibiotic resistant target seeker for comparative genome mining. *Nucleic Acids Res.* **48**, W546–W552 (2020).
77. Jia, B. et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).
78. Sélem-Mojica, N., Aguilar, C., Gutiérrez-García, K., Martínez-Guerrero, C. E. & Barona-Gómez, F. EvoMining reveals the origin and fate of natural product biosynthetic enzymes. *Microb. Genom.* **5**, e000260 (2019).
79. Chevrette, M. G. et al. Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).
80. Cereto-Massagué, A. et al. Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
81. Willighagen, E. L. et al. The chemistry development kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **9**, 33 (2017).
82. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors* (John Wiley & Sons, 2008).
83. Skinnider, M. A., Dejong, C. A., Franczak, B. C., McNicholas, P. D. & Magarvey, N. A. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *J. Cheminform.* **9**, 46 (2017).
84. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
85. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5**, 26 (2013).
86. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **8**, 36 (2016).
87. Grisoni, F. et al. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **1**, 44 (2018).
88. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
89. Capecchi, A. & Reymond, J.-L. Assigning the origin of microbial natural products by chemical space map and machine learning. *Biomolecules* **10**, 1385 (2020).
90. Riniker, S. Molecular dynamics fingerprints (MDFP): machine learning from MD data to predict free-energy differences. *J. Chem. Inf. Model.* **57**, 726–741 (2017).
91. Esposito, C., Wang, S., Lange, U. E. W., Oellien, F. & Riniker, S. Combining machine learning and molecular dynamics to predict p-glycoprotein substrates. *J. Chem. Inf. Model.* **60**, 4730–4749 (2020).
92. Bannan, C. C. et al. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **30**, 927–944 (2016).
93. Wang, S. & Riniker, S. Use of molecular dynamics fingerprints (MDFPs) in SAMPL6 octanol-water log P blind challenge. *J. Comput. Aided Mol. Des.* **34**, 393–403 (2020).
94. Gorostiola González, M. et al. 3DDPDs: describing protein dynamics for proteochemometric bioactivity prediction. A case for (mutant) G protein-coupled receptors. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv-2023-90082 (2023).
95. Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics* **36**, i718–i725 (2020).
96. Paull, K. D. 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.* **81**, 1088–1092 (1989).
97. Kauvar, L. M. et al. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2**, 107–118 (1995).
98. Petrone, P. M. et al. Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* **7**, 1399–1409 (2012).
99. Norinder, U., Spjuth, O. & Svensson, F. Using predicted bioactivity profiles to improve predictive modeling. *J. Chem. Inf. Model.* **60**, 2830–2837 (2020).
100. Mater, A. C. & Coote, M. L. Deep learning in chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559 (2019).
101. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric deep learning: grids, groups, graphs, geodesics, and gauges. Preprint at *arXiv*. https://doi.org/10.48550/arXiv.2104.13478 (2021).
102. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
103. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
104. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature, 2019).
105. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**, 573–584 (2020).
106. Jiménez-Luna, J., Skalic, M., Weskamp, N. & Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *J. Chem. Inf. Model.* **61**, 1083–1094 (2021).
107. Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S. & Unterthiner, T. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (eds Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) 331–345 (Springer International Publishing, 2019).
108. Webel, H. E. et al. Revealing cytotoxic substructures in molecules using deep learning. *J. Comput. Aided Mol. Des.* **34**, 731–746 (2020).
109. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
110. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* **57**, 1757–1772 (2017).
111. Duvenaud, D. et al. Convolutional networks on graphs for learning molecular fingerprints. in *Advances in Neural Information Processing Systems* 28 (NIPS 015).
112. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. in *Proceedings of the 34th International Conference on Machine Learning* 1263–1272 (2017).
113. Nguyen, T. et al. GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2021).
114. Yuan, W. et al. Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **57**, 875–882 (2017).
115. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
116. Liu, X., Ye, K., van Vlijmen, H. W. T., IJzerman, A. P. & van Westen, G. J. P. DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning. *J. Cheminform.* **15**, 24 (2023).
117. Li, X. & Fourches, D. Inductive transfer learning for molecular activity prediction: next-gen QSAR models with MolPMoFiT. *J. Cheminform.* **12**, 27 (2020).
118. Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminform.* **12**, 17 (2020).
119. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
120. Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).
121. Bjerrum, E. J. & Sattarov, B. Improving chemical autoencoder latent space and molecular generation diversity with heteroencoders. *Biomolecules* **8**, 131 (2018).
122. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
123. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**, 1023–1032 (2021).
124. Callaway, E. After AlphaFold: protein-folding contest seeks next big breakthrough. *Nature* **613**, 13–14 (2023).

125. Wallner, B. AFsample: improving multimer prediction with alphafold using aggressive sampling. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.20.521205 (2022).
126. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: ways to make an impact, and why we are not there yet. *Drug Discov. Today* **26**, 511–524 (2021).
127. Bender, A. & Cortés-Ciriano, I. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data. *Drug Discov. Today* **26**, 1040–1052 (2021).
128. Sydow, D., Rodríguez-Guerra, J. & Volkamer, A. in *Teaching Programming across the Chemistry Curriculum* 135–158 ACS Symposium Series vol. 1387 (American Chemical Society, 2021).
129. Korshunova, M., Ginsburg, B., Tropsha, A. & Isayev, O. OpenChem: a deep learning toolkit for computational chemistry and drug design. *J. Chem. Inf. Model.* **61**, 7–13 (2021).
130. Sieg, J., Flachsenberg, F. & Rarey, M. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *J. Chem. Inf. Model.* **59**, 947–961 (2019).
131. Lenselink, E. B. et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 45 (2017).
132. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
133. Topçuoğlu, B. D., Lesniak, N. A., Ruffin, M. T. 4th, Wiens, J. & Schloss, P. D. A framework for effective application of machine learning to microbiome-based classification problems. *MBio* **11**, e00434-20 (2020).
134. Quinn, T. P. & Erb, I. Examining microbe–metabolite correlations by linear methods. *Nat. Methods* **18**, 37–39 (2021).
135. Morger, A. et al. KnowTox: pipeline and case study for confident prediction of potential toxic effects of compounds in early phases of development. *J. Cheminform.* **12**, 24 (2020).
136. Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
137. Manica, M. et al. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).
138. Grinsztajn, L., Oyallon, E. & Varoquaux, G. in *Advances in Neural Information Processing Systems 35* (NeurIPS 2022) 507–520 (2022).
139. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at https://doi.org/10.48550/arXiv.2010.09885 (2020).
140. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**, 015022 (2022).
141. Chapelle, O., Zien, A. & Schölkopf, B. (Eds) *Semi-Supervised Learning* (MIT, 2006).
142. Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
143. Röttig, M. et al. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
144. Torrey, L. & Shavlik, J. in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* 242–264 (IGI Global, 2010).
145. Cai, C. et al. Transfer learning for drug discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).
146. Moret, M., Helmstädter, M., Grisoni, F., Schneider, G. & Merk, D. Beam search for automated design and scoring of novel ROR ligands with machine intelligence. *Angew. Chem. Int. Ed. Engl.* **60**, 19477–19482 (2021).
147. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).
148. Moret, M. et al. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 114 (2023).
149. Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discov. Today Technol.* **32–33**, 73–79 (2019).
150. Reker, D., Schneider, P. & Schneider, G. Multi-objective active machine learning rapidly improves structure-activity models and reveals new protein-protein interaction inhibitors. *Chem. Sci.* **7**, 3919–3927 (2016).
151. Djoumbou Feunang, Y. et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).
152. Reher, R. et al. Native metabolomics identifies the rivulariapeptolide family of protease inhibitors. *Nat. Commun.* **13**, 4619 (2022).
153. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
154. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
155. Liu, X., Ye, K., van Vlijmen, H. W. T., IJzerman, A. P. & van Westen, G. J. P. An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor. *J. Cheminform.* **11**, 35 (2019).
156. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
157. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
158. Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O. & Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **11**, 154–168 (2020).
159. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement learning for bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).
160. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public ki data. *J. Med. Chem.* **55**, 5165–5173 (2012).
161. Tiikkainen, P., Bellis, L., Light, Y. & Franke, L. Estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* **53**, 2499–2505 (2013).
162. Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **12**, 1–51 (2020).
163. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
164. Liu, T., Lin, Y., Wen, X., Jorissen, R. N. & Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35**, D198–D201 (2007).
165. Wimalaratne, S. M. et al. Uniform resolution of compact identifiers for biomedical data. *Sci. Data* **5**, 180029 (2018).
166. Rajan, K., Zielesny, A. & Steinbeck, C. DECIMER 1.0: deep learning for chemical image recognition using transformers. *J. Cheminformatics* **13**, 61 (2021).
167. Rajan, K., Brinkhaus, H. O., Sorokina, M., Zielesny, A. & Steinbeck, C. DECIMER-segmentation: automated extraction of chemical structure depictions from scientific literature. *J. Cheminform.* **13**, 20 (2021).
168. Schymanski, E. L. & Bolton, E. E. FAIR chemical structures in the *Journal of Cheminformatics*. *J. Cheminform.* **13**, 50 (2021).
169. Kautsar, S. A. et al. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.* **48**, D454–D458 (2020).
170. van Santen, J. A. et al. The natural products atlas: an open access knowledge base for microbial natural products discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
171. van Santen, J. A. et al. The natural products atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.* **50**, D1317–D1323 (2021).
172. Wang, M. et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
173. Wishart, D. S. et al. NP-MRD: the natural products magnetic resonance database. *Nucleic Acids Res.* **50**, D665–D677 (2022).
174. Flissi, A. et al. Norine: update of the nonribosomal peptide resource. *Nucleic Acids Res.* **48**, D465–D469 (2020).
175. Jarmusch, S. A., van der Hooft, J. J. J., Dorrestein, P. C. & Jarmusch, A. K. Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Nat. Prod. Rep.* **38**, 2066–2082 (2021).
176. Jarmusch, A. K. et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
177. Proteau, P. J. *Journal of Natural Products* 2022: perspectives, monthly cover art, and more. *J. Nat. Products* **85**, 1–2 (2022).
178. Clark, T. N. et al. Interlaboratory comparison of untargeted mass spectrometry data uncovers underlying causes for variability. *J. Nat. Prod.* **84**, 824–835 (2021).
179. Fiehn, O. et al. The metabolomics standards initiative (MSI). *Metabolomics* **3**, 175–178 (2007).
180. Frank, A. M. et al. Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
181. Miller, I. J. et al. Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic Acids Res.* **47**, e57 (2019).
182. Schymanski, E. L. et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
183. Deutsch, E. W. et al. Universal spectrum identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).
184. Bittremieux, W. et al. Universal MS/MS visualization and retrieval with the metabolomics spectrum resolver web service. Preprint at *BioRxiv* https://doi.org/10.1101/2020.05.09.086066 (2020).
185. Gordon, J. E. Chemical inference. 2. formalization of the language of organic chemistry: generic systematic nomenclature. *J. Chem. Inf. Comput. Sci.* **24**, 81–92 (1984).
186. Wang, Y. et al. PubChem's bioassay database. *Nucleic Acids Res.* **40**, D400–D412 (2012).
187. Banerjee, P. et al. Super Natural II—a database of natural products. *Nucleic Acids Res.* **43**, D935–D939 (2015).
188. Zeng, X. et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, D1217–D1222 (2018).
189. van der Hooft, J. J. J. A community-driven paired data platform to accelerate natural product mining by combining structural information from genomes and metabolomes. Preprint at https://doi.org/10.18174/fairdata2018.16286 (2018).
190. Eldjárn, G. H. et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput. Biol.* **17**, e1008920 (2021).
191. Schorn, M. A. et al. A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).
192. Doroghazi, J. R. et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–968 (2014).
193. McClure, R. A. et al. Elucidating the rimosamide-detoxin natural product families and their biosynthesis using metabolite/gene cluster correlations. *ACS Chem. Biol.* **11**, 3452–3460 (2016).
194. Goering, A. W. et al. Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent. Sci.* **2**, 99–108 (2016).
195. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).

196. Caesar, L. K. et al. Correlative metabologenomics of 110 fungi reveals metabolite-gene cluster pairs. *Nat. Chem. Biol.* **19**, 846–854 (2023).
197. Soldatou, S. et al. Comparative metabologenomics analysis of polar actinomycetes. *Mar. Drugs* **19**, 103 (2021).
198. Sulheim, S. et al. Enzyme-constrained models and omics analysis of streptomyces coelicolor reveal metabolic changes that enhance heterologous production. *iScience* **23**, 101525 (2020).
199. Amos, G. C. A. et al. Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl Acad. Sci. USA* **114**, E11121–E11130 (2017).
200. Wandy, J. & Daly, R. GraphOmics: an interactive platform to explore and integrate multi-omics data. *BMC Bioinform.* **22**, 603 (2021).
201. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2020).
202. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: collection of open natural products database. *J. Cheminform.* **13**, 2 (2021).
203. Rutz, A. et al. The LOTUS initiative for open knowledge management in natural products research. *eLife* **11**, e70780 (2022).
204. Chen, Y., Stork, C., Hirte, S. & Kirchmair, J. NP-scout: machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules* **9**, 43 (2019).
205. Cao, L. et al. MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nat. Commun.* **12**, 3718 (2021).
206. Visser, U. et al. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinform.* **12**, 257 (2011).
207. Sarntivijai, S. et al. CLO: the cell line ontology. *J. Biomed. Semant.* **5**, 37 (2014).
208. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).
209. Cooper, M. A. A community-based approach to new antibiotic discovery. *Nat. Rev. Drug. Discov.* **14**, 587–588 (2015).
210. Cech, N. B., Medema, M. H. & Clardy, J. Benefiting from big data in natural products: importance of preserving foundational skills and prioritizing data quality. *Nat. Prod. Rep.* **38**, 1947–1953 (2021).
211. Blin, K., Shaw, S., Kautsar, S. A., Medema, M. H. & Weber, T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.* **49**, D639–D643 (2021).
212. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass. Spectrom.* **45**, 703–714 (2010).
213. Haug, K. et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444 (2020).
214. Kuhn, S. & Schlörer, N. E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2–a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Reson. Chem.* **53**, 582–589 (2015).
215. Irwin, J. J. et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
216. Hastings, J. et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, D1214–D1219 (2016).
217. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).
218. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
219. Blaskovich, M. A. T., Zuegg, J., Elliott, A. G. & Cooper, M. A. Helping chemists discover new antibiotics. *ACS Infect. Dis.* **1**, 285–287 (2015).
220. Waagmeester, A. et al. Wikidata as a knowledge graph for the life sciences. *eLife* **9**, e52614 (2020).
221. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Target prediction by cascaded self-organizing maps for ligand de-orphaning and side-effect investigation. *J. Cheminform.* **6**, P47 (2014).
222. Navarro-Muñoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
223. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci. USA* **113**, 13738–13743 (2016).
224. Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **48**, 722–730 (2015).
225. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **46**, 3–26 (2001).
226. Janssen, A. P. A. et al. Drug discovery maps, a machine learning model that visualizes and predicts kinome–inhibitor interaction landscapes. *J. Chem. Inf. Model.* **59**, 1221–1229 (2019).
227. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open. Source Softw.* **3**, 861 (2018).
228. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).
229. Feher, M. & Schmidt, J. M. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **43**, 218–227 (2003).
230. Béquignon, O. J. M. et al. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminform.* **15**, 3 (2023).

Acknowledgements

All authors thank the Lorentz Center and Leiden University for funding the Lorentz Workshop 'Artificial Intelligence for Natural Product Drug Discovery' that laid the foundation for this Review. M.W.M. was supported by funds from the Duchossois Family Institute at the University of Chicago. K.R.D. was supported by the UK Research and Innovation Biotechnology and Biological Sciences Research Council (BB/R022054/1). N.G. was supported by an NSF CAREER award (award number 2047235). J.J.J.v.d.H. was supported by an ASDI eScience grant from the Netherlands eScience Center (award number ASDI.2017.030). N.I.M. is supported by funding from the European Research Council (ERC consolidator grant agreement no. 725523). K.B. was supported by a Novo Nordisk Foundation grant NNF20CC0035580. M.G.G. was supported by ONCODE funding. E.J.N.H. was supported by the LOEWE Center for Translational Biodiversity Genomics and the Funds of the Chemical Industry Germany. M.S. was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 239748522, SFB 1127 ChemBioSys. M.A.B. was supported by the National French Agency (ANR grants 15-CE29-0001 and 20-CE43-0010). C.M.C. was supported by a National Library of Medicine training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359). S.F. was supported by MASTS/IbioIC/Xanthella. O.V.K. was funded by the Klaus Faber Foundation. H.K. was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (grants NRF 2018R1A5A2023127 and NRF 2022R1F1A1074623 11). J.M. was supported by a grant from the Research Foundation – Flanders (G061821N). E.R.R. was supported by the US National Science Foundation (DBI-1845890). D.R. was supported, in part, by a Flash Grant from NC Biotech (2021-FLG-3819), a UNC CGIBD Pilot Award (NIH NIDDK DK034987), a Duke Cancer Institute and Duke Microbiome Center Pilot Award (NIH NCI CA014236), the Engineering Research Center for Precision Microbiome Engineering (NSF EEC-2133504), and the Duke Science and Technology Initiative. P.S. acknowledges support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. N.Z. was supported by Germany's Excellence Strategy – EXC 2124-390838134. H.U.K. was supported by the KAIST Key Research Institute (Interdisciplinary Research Group) Project. R.G.L. was supported by the US NIH (U41-AT008718 and U24-AT010811). S.L.R. was supported by Eawag discretionary funding. M.H.M. was supported by the Leiden University 'van der Klaauw' chair for theoretical biology and an ERC Starting Grant (DECIPHER-948770). We thank A. R. Leach for discussion of the content of this manuscript. We thank all participants of the Lorentz Workshop 'Artificial Intelligence for Natural Product Drug Discovery' who did not participate in the review writing for providing inspiration through their talks and/or discussions.

Author contributions

M.W.M., S.S.E., D.M., B.R.T., M.G.G., S.L.-M., K.R., T.d.R., J.A.v.S., M.S., S.F., A.K.H.H., R.G.L., S.L.R. and M.H.M. researched data for the article. M.W.M., K.R.D., S.S.E., N.G., J.J.J.v.d.H., N.I.M., D.M., B.R.T., F.B., J.D., E.J.N.H., F.H., T.d.R., M.S., M.J.B., D.A.v.B., L.M.C., C.M.C., C.A.D., C.D., F.G., A.H., W.J., O.V.K., H.K., T.F.L., J.M., E.R.R., R.R., D.R., P.S., M.S., M.A.S., A.S.W., N.Z., R.J.M.G., A.V., W.H.G., R.M., G.P.v.W., G.J.P.v.W., A.K.H.H., R.L., S.L.R. and M.H.M. contributed substantially to discussion of the content. M.W.M., K.R.D., N.G., J.J.J.v.d.H., B.R.T., F.B., J.D., M.G.G., M.S., M.J.B., M.A.B., L.M.C., C.M.C., C.A.D., S.F., A.H., W.J., O.V.K., S.A.K., T.F.L., J.M., D.R., M.A.S., A.S.W., B.Z., N.Z., R.J.M.G., P.G., A.V., W.H.G., G.J.P.v.W., A.K.H.H., R.G.L., S.L.R. and M.H.M. wrote the article. M.W.M., K.R.D., S.S.E., N.G., J.J.J.v.d.H., N.I.M., D.M., B.R.T., F.B., K.B., E.J.N.H., F.H., T.d.R., M.J.B., L.M.C., C.M.C., D.-A.C., C.A.D., F.G., S.A.K., H.K., E.R.R., R.R., P.S., M.A.S., E.L.W., B.Z., W.H.G., H.U.K., R.M., G.P.v.W., G.J.P.v.W., A.K.H.H., R.G.L., S.L.R. and M.H.M. reviewed and/or edited the manuscript before submission.

Competing interests

J.J.J.v.d.H. is a member of the scientific advisory board of NAICONS Srl., Milan, Italy. C.A.D. is a founding member of Adapsyn Bioscience. M.A.S. is a consultant to Adapsyn Bioscience. M.H.M. is on the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals. The other authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Drug Discovery* thanks Hosein Mohimani and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Related links

National Database of Antibiotic Resistant Organisms (NDARO): https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/
RDKit: Open-Source Cheminformatics Software: http://www.rdkit.org/
Wikidata Query Service: https://w.wiki/5bpq

[1]Duchossois Family Institute, The University of Chicago, Chicago, IL, USA. [2]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK. [3]Department of Molecular Biotechnology, Institute of Biology, Leiden University, Leiden, The Netherlands. [4]School of Chemistry and Biochemistry, Center for Microbial Dynamics and Infection, Georgia Institute of Technology, Atlanta, GA, USA. [5]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. [6]Department of Biochemistry, University of Johannesburg, Johannesburg, South Africa. [7]Biological Chemistry Group, Institute of Biology, Leiden University, Leiden, The Netherlands. [8]Institute of Molecular Bio Science, Goethe-University Frankfurt, Frankfurt am Main, Germany. [9]LOEWE Center for Translational Biodiversity Genomics (TBG), Frankfurt am Main, Germany. [10]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kongens Lyngby, Denmark. [11]Biozentrum, University of Basel, Basel, Switzerland. [12]Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Leiden, The Netherlands. [13]ONCODE institute, Leiden, The Netherlands. [14]Center for Digitalization and Digitality, Hochschule Düsseldorf, Düsseldorf, Germany. [15]Institut für Mikrobiologie, Eidgenössische Technische Hochschule (ETH) Zürich, Zürich, Switzerland. [16]Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany. [17]School of Chemical Sciences, University of Auckland, Auckland, New Zealand. [18]Department of Chemistry, Simon Fraser University, Burnaby, British Columbia, Canada. [19]Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Jena, Germany. [20]Pharmaceuticals R&D, Bayer AG, Berlin, Germany. [21]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA. [22]Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA. [23]Équipe "Chimie des Substances Naturelles", Université Paris-Saclay, CNRS, BioCIS, Orsay, France. [24]Structural and Computational Biology Unit, EMBL, Heidelberg, Germany. [25]Division of Pharmaceutical Sciences, School of Pharmacy, University of Wisconsin-Madison, Madison, WI, USA. [26]WRDM - Machine Learning Research, Pfizer, Berlin, Germany. [27]Adapsyn Bioscience, Hamilton, Ontario, Canada. [28]Chemistry Department, University of St Andrews, St Andrews, UK. [29]Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands. [30]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht, The Netherlands. [31]Laboratory of Physical Chemistry, ETH Zürich, Zürich, Switzerland. [32]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Saarbrücken, Germany. [33]Drug Bioinformatics, Medical Faculty, Saarland University, Homburg, Germany. [34]Center for Bioinformatics, Saarland University, Saarbrücken, Germany. [35]Department of Chemistry, Scripps Research, FL, USA. [36]College of Pharmacy and Integrated Research Institute for Drug Development, Dongguk University Seoul, Goyang-si, Republic of Korea. [37]Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil. [38]Center for Microbiology, VIB-KU Leuven, Heverlee, Belgium. [39]Department of Biology, KU Leuven, Heverlee, Belgium. [40]Institute of Pharmaceutical Biology and Biotechnology, University of Marburg, Marburg, Germany. [41]Institute of Pharmacy, Martin-Luther-University Halle-Wittenberg, Halle (Saale), Germany. [42]Department of Biomedical Engineering, Duke University, Durham, NC, USA. [43]Duke Microbiome Center, Duke University, Durham, NC, USA. [44]Laboratory of Artificial Chemical Intelligence, Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [45]Microsoft Research, Cambridge, UK. [46]Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia, Canada. [47]Department of Chemistry, Vanderbilt University, Nashville, TN, USA. [48]Department of Biological Sciences, Vanderbilt University, Nashville, TN, USA. [49]Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, The Netherlands. [50]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridgeshire, UK. [51]Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), Institute for Bioinformatics and Medical Informatics (IBMI), University of Tuebingen, Tuebingen, Germany. [52]Bonsai team, CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, Université de Lille, Villeneuve d'Ascq Cedex, France. [53]In silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité - Universitätsmedizin Berlin, Berlin, Germany. [54]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. [55]Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. [56]Department of Pharmacy, Saarland University, Saarbrücken, Germany. [57]German Center for infection research (DZIF), Braunschweig, Germany. [58]Helmholtz International Lab for Anti-Infectives, Saarbrücken, Germany. [59]Netherlands Institute of Ecology, NIOO-KNAW, Wageningen, The Netherlands. [60]Department of Environmental Microbiology, Eawag: Swiss Federal Institute for Aquatic Science and Technology, Dübendorf, Switzerland. [61]Institute of Biology, Leiden University, Leiden, The Netherlands. [62]These authors contributed equally: Michael W. Mullowney, Katherine R. Duncan, Somayah S. Elsayed, Neha Garg, Justin J. J. van der Hooft, Nathaniel I. Martin, David Meijer, Barbara R. Terlouw.