Artificial Intelligence For Natural Product Drug Discovery

Michael W. Mullowney[1#], Katherine R. Duncan[2#], Somayah S. Elsayed[3#], Neha Garg[4#], Justin J.J. van der Hooft[5,6#], Nathaniel I. Martin[7#], David Meijer[5#], Barbara R. Terlouw[5#], 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 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[53,34], William H. Gerwick[54], Hyun Uk Kim[55], Rolf Müller[32], Gilles P. van Wezel[3,56], Gerard van Westen[12]*, Anna K.H. Hirsch[32]*, Roger Linington[18]*, Serina L. Robinson[57]*, Marnix H. Medema[5,58]*

1. Duchossois Family Institute, The University of Chicago, Chicago, IL 60637, USA.
2. University of Strathclyde, Strathclyde Institute of Pharmacy and Biomedical Sciences, Glasgow, United Kingdom.
3. Department of Molecular Biotechnology, Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE, The Netherlands.
4. School of Chemistry and Biochemistry, Center for Microbial Dynamics and Infection, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA, 30332-2000, USA.
5. Bioinformatics Group, Wageningen University, Droevendaalseweg, 6708 PB, Wageningen, the Netherlands.
6. Department of Biochemistry, University of Johannesburg, Auckland Park, Johannesburg 2006, South Africa.
7. Biological Chemistry Group, Institute of Biology, Leiden University, Sylviusweg 72, 2333 BE, The Netherlands.
8. Institute of Molecular Bio Science, Goethe-University Frankfurt, D-60438 Frankfurt am Main, Germany.
9. LOEWE Center for Translational Biodiversity Genomics (TBG), D-60325 Frankfurt am Main, Germany.
10. The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs.Lyngby, Denmark.
11. Biozentrum, University of Basel, Spitalstrasse 41, 4056 Basel .
12. Drug Discovery and Safety, Leiden Academic Centre for Drug Research, Einsteinweg 55, 2333 CC, Leiden, The Netherlands.
13. ONCODE institute, The Netherlands.
14. Center for Digitalization and Digitality, Hochschule Düsseldorf, Münsterstraße 156, 40476 Düsseldorf, Germany.
15. Institut für Mikrobiologie, Eidgenössische Technische Hochschule (ETH) Zürich, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland.
16. Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07743, Jena, Germany.
17. School of Chemical Sciences, University of Auckland, Auckland, New Zealand.
18. Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.
19. Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller University, Lessing strasse 8, 07743,Jena, Germany..
20. Bayer AG, Pharmaceuticals R&D, Müller strasse 178, 13343 Berlin, Germany..
21. Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA.
22. Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan, USA.
23. Équipe “Chimie des Substances Naturelles” Université Paris-Saclay, CNRS, BioCIS, 17 avenue des Sciences, 91400 Orsay, France..
24. Structural and Computational Biology Unit, EMBL, Heidelberg, Germany.
25. Division of Pharmaceutical Sciences, School of Pharmacy, University of Wisconsin-Madison, 777 Highland Avenue, Madison, WI 53705, USA.

This is a peer-reviewed, accepted author manuscript of the following research article: Mullowney, M. W., et. al. (2023). Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery* . https://doi.org/10.1038/s41573-023-00774-7

26. Bayer AG, Machine Learning Research, Müllerstrasse 178, 13343 Berlin, Germany..
27. Adapsyn Bioscience, Hamilton, ON, Canada.
28. University of St Andrews, Chemistry Department.
29. Institute for Complex Molecular Systems, Dept. Biomedical Engineering, Eindhoven University of Technology, Groene Loper 7, 5612AZ, Eindhoven, Netherlands.
30. Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Princetonlaan 6, 3584 CB Utrecht, The Netherlands. .
31. Laboratory of Physical Chemistry, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland.
32. Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), 66123 Saarbrücken, Germany.
33. Drug Bioinformatics, Medical Faculty, Saarland University, 66421, Homburg, Germany.
34. Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany.
35. Department of Chemistry, Scripps Research, Florida, United States.
36. College of Pharmacy and Integrated Research Institute for Drug Development, Dongguk University Seoul, Goyang-si, Gyeonggi-do, Republic of Korea.
37. Center for Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, Brazil.
38. VIB-KU Leuven Center for Microbiology, Kasteelpark Arenberg 31, 3001 Heverlee, Belgium.
39. KU Leuven Department of Biology, Kasteelpark Arenberg 31, 3001 Heverlee, Belgium.
40. Institute of Pharmaceutical Biology and Biotechnology, University of 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 (LIAC), 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, BC, 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, 6229 ER Maastricht, The Netherlands.
50. European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
51. Interfaculty Institute for Microbiology and Infection Medicine Tuebingen (IMIT), Institute for Bioinformatics and Medical Informatics (IBMI), University of Tuebingen, Germany, DZIF @ partnersite Tuebingen.
52. Bonsai team, CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, Université de Lille, 59655, Villeneuve d'Ascq Cedex, France.
53. In silico Toxicology and Structural Bioinformatics, Institute of Physiology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117, 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. Netherlands Institute of Ecology, NIOO-KNAW, Droevendaalsesteeg 10, 6708 PB, Wageningen, The Netherlands.
57. Department of Environmental Microbiology, Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, CH-8600, Dübendorf, Switzerland.
58. Institute of Biology, Leiden University, Sylviusweg 72, 2333BE Leiden, The Netherlands..

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 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 discuss current and future synergies between these developments to effectively identify drug candidates out of the plethora of molecules produced by nature.

Introduction

Plants, fungi, animals, and bacteria produce a wide range of specialized metabolites, also known as natural products (NPs). Across the tree of life, these comprise hundreds of thousands of different chemical structures—including terpenes, polyketides, peptides, saccharides and alkaloids—that facilitate an organism's ability to thrive in a particular environment. They play critical 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 toward the benefit of society with remarkable success as crop protection agents, antibiotics, chemotherapeutics, immunosuppressants, food preservatives, pigments and ingredients for cosmetics. Natural products remain a promising source for the discovery of such drugs based on their relatively high degree of three-dimensionality–as opposed to the often 'flat' synthetic structures–and their origins as natural metabolites, making them likely substrates for transporter systems[1,2].

While natural product discovery programs diminished between roughly 1990 and 2010 to make space for 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 catalyzed by the availability of large-scale omics data, which allows significantly 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: over 2,000 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], with the prospect of revolutionizing drug discovery. Fueled by data on known biosynthetic pathways and their chemical products, which is increasingly standardized and stored in public databases, artificial intelligence (AI) approaches are now being developed to train machine learning algorithms that predict (parts of) chemical structures of BGC products based on DNA sequence alone. While this helps distinguish new from known chemistry (dereplication) and link molecules to their biosynthetic genes[6], there is an urgent need for more effective ways to prioritize the enormous predicted natural product biosynthetic diversity for concrete drug leads.

In the field of computational drug design, a range of AI strategies are being developed that may help address this challenge by better understanding structure-activity relationships and predicting macromolecular targets for molecules based on their chemical structures. Here, traditionally two main disciplines prevail. On the one hand, the statistical modeling field

focuses on finding correlations between chemical structure and biological activity, termed quantitative structure–activity-relationship modeling. On the other hand, the structure-based field attempts to fit three-dimensional chemical structures to protein targets (docking) and subsequently study their behavior on the nano to millisecond timescale (molecular dynamics). For both fields, introducing AI methods has 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. Based on chemical similarity, advanced machine learning can use these data to obtain models able to predict the potential activity of untested chemical structures within these extensive chemical collections. Moreover, these methods can also be used to systematically analyze large datasets routinely produced from extended molecular dynamics and identify hidden patterns in the protein dynamics. 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), 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).

There is thus great potential for cross-fertilization between the fields of omics-based natural product discovery and computational drug design (Figure 1). The use of AI leads to a rapid acceleration of scientific progress in these fields and to a convergence of their methods and directions. For example, suppose machine learning algorithms could predict structural features of molecules from sequence information. In that case, these predictions could then be used in AI approaches to predict their functions and mechanisms of action. However, to date, these fields have interacted very little, and scientists from these disciplines thus far mainly participate in conferences in their respective fields. Additionally, some challenges are apparent, for example, matching synthetic product chemical space (‘drug like’, consisting mostly of planar small molecules designed using Lipinski’s rule of five [7]) to natural-product-like chemical space for which no such rules exist. This matching is key to accurately estimate models’ applicability domains and therefore reliability. In this current paper, scientists from both areas have come together to write a joint perspective on new ways to connect these research areas and jointly leverage the power of AI to utilize the vast chemical diversity of the biosphere for the development of novel drugs that can protect humanity from the threats of antimicrobial resistance and future infectious-disease pandemics.

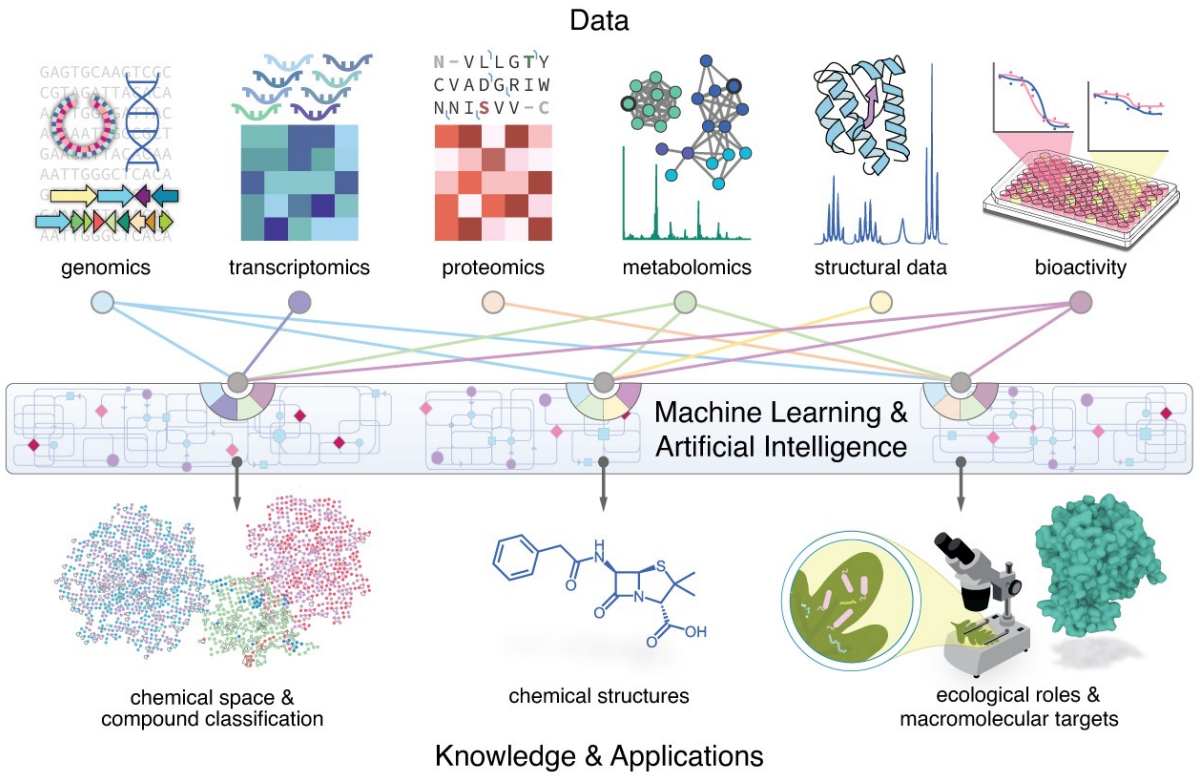

Figure 1. Applications of machine learning in natural-product and drug discovery. Classical analyses typically only use a small fraction of multifaceted datasets. AI methods can help to integrate different data types to learn complex feature relationships and develop meaningful hypotheses.

Advances in AI technologies

In recent years, scientists have started to apply machine learning for the discovery and structural characterization of NPs and to predict relationships between structure and pharmaceutical properties. Machine learning is a subfield of AI that generates insights by using algorithms to recognize patterns from data. In this article, AI is used as a broad term encompassing machine learning. As algorithms and featurization methods become more powerful and diverse, we expect AI technologies to play an increasingly important role in NP exploration.

Current applications of AI in natural product drug discovery

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 biosynthetic gene clusters (BGCs) still largely relies on rule-based methods such as those used in antiSMASH[8] and PRISM[9]. While these approaches are successful at detecting known BGC classes, they are less proficient at identifying novel types of BGCs or unclustered pathways[10,11]. In these more complex cases, ML algorithms have been shown to offer significant advantages over rule-based methods. For

example, the HMM-based method ClusterFinder[12], the deep-learning approaches DeepBGC[13] and GECCO [14], and several RiPP genome mining algorithms [15–18] each employ deep learning or support vector machines (SVMs) 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.[19,20] Although they still have a higher false positive rate than rule-based approaches and also suffer from false negatives for known types of BGCs, they have already demonstrated utility in identifying novel classes of natural product biosynthetic pathways[10]. For example, the decRiPPter algorithm, aimed at predicting novel RiPP families, identified pristinin, which belongs to a novel class of lanthipeptides[15]. 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[17].

While 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[21], including library matching/searching using mass spectral similarity metrics[22,23], molecular-formula annotation[24,25], molecular-class annotation[26,27] and retention time prediction[28]. The efficacy of these algorithms is still limited by the relatively small sets of MS/MS spectra annotated with the fragment ion chemical structures of their corresponding metabolites. However these algorithms can be enhanced by imputing missing data, *e.g.,* by predicting molecular fingerprints or simulated spectra from metabolite structures directly[27]. Similarly, NMR metabolome mining tasks are undergoing significant transformations[29], as deep learning provides new avenues towards improving NMR spectrum reconstruction, denoising[30], peak picking, *J*-coupling prediction[31] and spectral deconvolution[32].

Ultimately, AI algorithms that link genome-mined BGCs and GCFs 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[33]. Such algorithms will help 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 NP drug discovery studies require the ability to unambiguously solve the structures of isolated compounds[34]. This task is challenging due to the chemical complexity of metabolites existing in nature. Structure elucidation requires the collection, analysis and compilation of multiple data types, including NMR, HRMS, MS/MS, IR, UV, ECD, X-ray, and experimental and/or computational inspection of the encoded enzymes within the producing BGC[35,36]. Recently, the microcrystal electron diffraction (MicroED) technique was added to this arsenal, which has the potential to significantly accelerate structure elucidation by allowing analysis of submicron-sized crystals of chemical compounds[37,38].

In general, significant efforts have been made to improve the structural characterization of NPs 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[39,40]. Subsequently, AI has been used to predict molecular formulae from MS spectra[41], match MS spectra to compounds in molecular databases using deep neural networks[41,39], elucidate structures *de novo* as SMILES strings from MS/MS spectra[42], and predict chemical properties and identify small molecules from MS[1] and collisional cross section (CCS) data[43]. Similarly, AI has been used to augment NMR-based structure elucidation and annotation. Computer-Assisted Structure Elucidation (CASE) programs [44] 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 NPs including the new macrolide symplocolide A [45], SMART-Miner[46] and COLMAR[47], 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 [48,49]. 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[50] have been developed to address this issue by filtering force field-generated conformations and thus significantly minimizing the computational cost.

Structure-activity relationship prediction

The rapid deployment of NPs or NP-inspired compounds in medicine is often hampered by the fact that the targets of these NPs are rarely known. This caveat 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 due 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 [51]. Virtually all computational drug discovery approaches have been successfully applied to elucidate targets of NPs, including docking[52], clustering[53], bioactivity fingerprints [54], pharmacophores[55], and machine learning[56]. In some cases, this has also led to new insights regarding the mechanisms of action of NPs that were already in clinical trials [57]. 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 to predict the biological activities of NPs from their chemical structures.

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 ML models using sparse or variable training set data are under active development, 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 will discuss algorithmic developments that could have a significant impact. Subsequently (in section 3), we will consider data generation and standardization challenges that will need to be addressed to exploit the full potential of these algorithms.

New AI technologies to boost natural product drug discovery

New chemical featurization technologies

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 (Figure 2). Simplification is inherent to featurization. In rare cases this can lead to clashes where 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. The most ubiquitous method for featurizing a molecule is to convert its molecular structure into a sequence of bits or counts[58]. Algorithms to create such fingerprints are readily implemented in cheminformatic software packages such as RDKit[59] and the Chemistry Development Kit[60]; however, molecule features can be manually determined as well[61].

Circular fingerprints have enabled the most accurate identification of structurally related NPs[62–65]. However, circular fingerprints were found to be less useful than pharmacophore-based descriptors for scaffold hopping from NPs to synthetic mimetics [66]. Other recent examples are MAP4 fingerprints, which combine substructure and atom-pair concepts and can be used to distinguish bacterial from fungal NPs [67,68]. Also, features created from short molecular dynamic simulations can be used to accurately predict partition coefficients and solvation free energies[69–72]. Recent approaches to 'k-merize' 3D shapes[73], which can be sampled from molecule conformers, may also provide promise for fingerprinting, as they may take into account the three-dimensional shape of molecules. Conversely, compound features that do not describe the compound structure at all can also be helpful, as exemplified by bioactivity fingerprints[74–78].

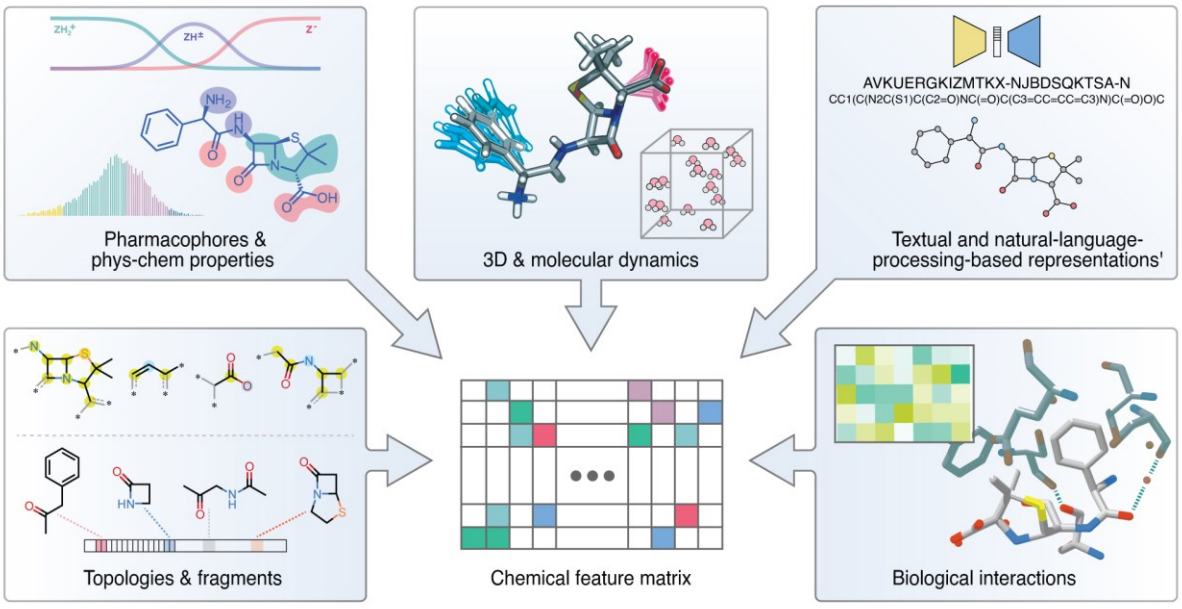

Figure 2. Chemical featurization techniques. Numerous featurization technologies are available to encode chemical information in a manner that machine learning techniques can process. These technologies range from simple physicochemical properties, via commonly used circular fingerprints, to advanced 3D and neural net-based encoders. Usage of an appropriate featurization method is key as the interpretation of a machine learning model is based on the features this model is trained on. Although possible, combinations of featurization techniques are not common.

Deep learning

A diverse array of AI algorithms have been developed over the past decade, many of which have been successfully applied to NP research (Figure 3A-F). 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 NP computational research to non-Euclidean domains[79,80]. Deep learning for molecular function prediction on molecular graphs sometimes outperforms simpler machine-learning models on circular fingerprints [81], although this seems to vary between datasets and applications[82,83]. Furthermore, explainable AI methods (XAI) have been shown to improve interpretability of such deep-learning models [84,85], for example in the assessment of preclinical relevance[86] and for pharmacophore and toxicophore identification[87,88].

Applications of deep learning include molecular graph neural network-approaches [89–92], for instance, for predicting drug-target binding affinity [93], SMILES-based approaches for de novo drug-like molecule generation [94,95] and property prediction [96,97] and surface mesh-based approaches for protein pocket-conditioned molecular representations [98]. Moreover, encoder-decoder architectures are used to featurize compounds for virtual screening from different input formats [99–101]. A comprehensive overview of deep learning molecular representations, which can be applied to molecular structure data in NP research, is found in ref. 102.

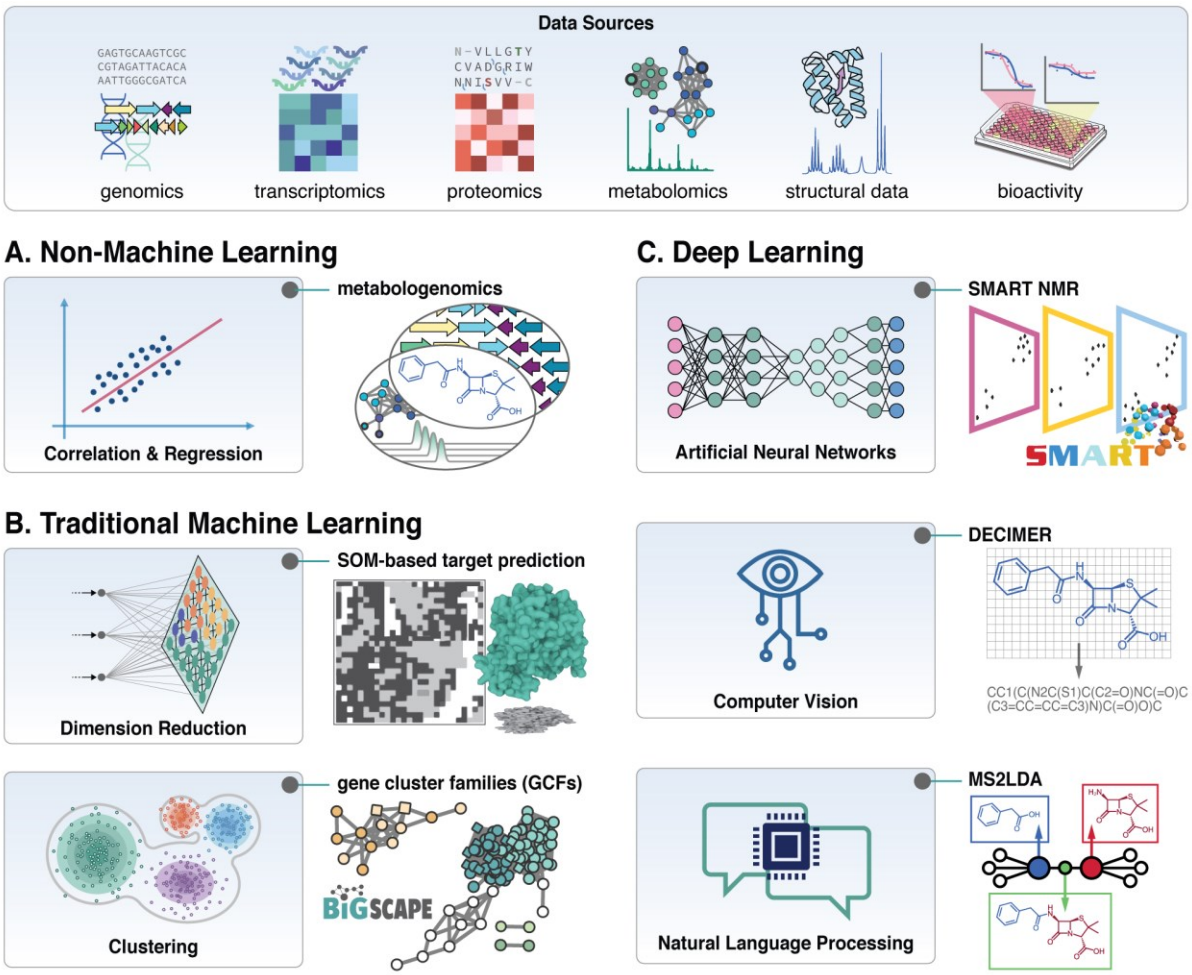

Figure 3. AI technologies key to natural product drug discovery and example applications. These techniques include, but are not limited to, A) non-machine learning methods such as correlation and regression e.g., linking metabolomic and genomic data103, B) traditional machine learning such as self-organizing maps (SOMs) e.g., for macromolecular target prediction[103] and clustering e.g., grouping gene cluster families[105]. C) deep learning such as convolutional neural networks e.g., for chemical structure elucidation[45] computer vision e.g., automatic chemical image recognition[106] and natural language processing e.g., topic modeling for chemical substructure exploration and annotation[107].

One of the most notable deep learning approaches of the past years is AlphaFold[108], which can predict the 3D structure of protein structures from their primary amino acid sequence by learning from the entire corpus of the Protein Data Bank. Since the landmark breakthrough by AlphaFold, accurate modeling approaches building on this work continue to raise the bar[109] by tackling challenges such as multimeric structure prediction[110]. For NP research, structural prediction is highly relevant as it can, *e.g.,* help predict the substrate specificities across NP biosynthetic enzyme families or help 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 NP computational research, although NP data are currently much sparser. As deep learning for NP computational research is still in its infancy, caution should be applied to its predictions[111,112]. To build trust and utilize the full potential of deep learning, we believe a set of best practices needs to be established for using deep learning techniques in NP research[113,114]:

1. Compare the performance of new deep learning models with simpler models to validate and motivate the trade-off between interpretability and prediction results[115–119].
2. Clarify the scope in which the model optimally performs by defining its applicability domain and adding confidence estimates to predictions[120,121].
3. Evaluate the model through cross-validation and using a true hold-out set, avoiding a random splitting approach with a preference for chemical clustering or temporal splitting[116], and, if applicable, including prospective experiments. Due to the practice to publish synthetic compounds as chemical analogs with a Structure-Activity Relationship, random splitting for validation overestimates models' abilities to generalize. Therefore, chemical clustering or temporal splitting is essential to truly validate created models. [116].
4. 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[88,122,123].

Deep learning algorithms will definitely not always be the most suitable tools[124]. Nonetheless, we do expect they will become increasingly useful to address challenges like structure elucidation and activity prediction as datasets in compatible formats grow.

Data, transfer, active, and reinforcement learning

One of the biggest challenges for deep learning in NP research is open access to big, curated data regimes. This would involve convincing more public funding agencies and joint consortia (*e.g.*, Melloddy for federated learning) to support the creation and maintenance of curated and FAIR datasets[125]. Data augmentation and synthetic data generation, while valuable techniques, should be done 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[126,127].

While deep learning techniques can overcome issues of incomplete sample labeling and small datasets, semi-supervised learning (combining labeled with unlabeled data) can assist with learning on datasets with incomplete labeling [128,129]. This has been applied in the past, for example, to improve substrate specificity predictions of NP biosynthetic enzymes using transductive support vector machines, where this helped map the shape of unlabeled sequence space to better know how queries would relate to labeled data points[130]. An alternative is transfer learning[131], 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[132], for example, in *de novo* molecular design[133–135].

Active learning techniques, which guide the selection of unlabeled data for labeling through experimentation, can also be deployed when labeled training data are limited[136]. This has been successfully applied for identifying small molecules that inhibit the protein-protein interaction between the anti-cancer target CXC chemokine receptor 4 and its ligand by actively retrieving informative active compounds that continuously improved the adaptive structure-activity model [137]. Multiple practical challenges remain before active learning can be broadly deployed [136], 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 NP research, where experiments are commonly complex. For example, CANOPUS[27], a deep neural network-based structure class annotation tool that is based on MS spectra, utilizes other AI tools including Classyfire[138] and NPClassifier[139] to label data and thus train the network. This enabled the structural elucidation of the novel rivulariapeptolide protease inhibitors from complex mixtures[27,140]. With increasing experimental resolution and automation, we believe that active learning will play a central role in future NP research.

Similarly, reinforcement learning, which steers the output of a machine learning algorithm toward user-defined regions of optimality *via* a predefined (computational) reward function, has shown promise in *de novo* design towards attractive regions of chemical space [141–143], for rule-based organic chemistry and for retrosynthesis prediction[144–147].

Data sources and data standardization

High-quality training datasets underpin and are thus critical to the success of AI algorithms. Unstructured datasets (e.g. unannotated mass spectrometry 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 where the breadth of chemical space is high but the coverage of most published datasets is low. Integrating data from different datasets and ensuring that annotation methods are consistent is therefore a major bottleneck for ML training set development.

In this section, we explore the characteristics and attributes necessary for creating high-quality datasets to advance natural product discovery. In particular, we assess the prerequisites needed in relation to standardization, accuracy, and completeness, review the current state of

natural product data and databases, and identify key challenges and potential solutions for database development.

The NP database landscape

The NP database landscape is large and diverse, but is also highly fragmented and currently contains few comprehensive and well-curated data resources[148]. Unfortunately, NP-related data are often underrepresented or not annotated as NPs in large generalist databases (*e.g.*, PubChem, ChEMBL, Reaxys, Scifinder); for example, as of January 2023, only 8951 natural products have a ChEMBL identifier according to WikiData[149]. Additionally, documentation of data sources, acquisition, and changes–known as data provenance– is not well maintained in most NP databases. For example, literature citations or information on source organisms and associated biosynthetic gene clusters may be missing. Furthermore, although some databases include bioassay data for pure compounds (*e.g.*, ChEMBL[150], BindingDB[151]), very few include bioassay data for natural product extracts and fractions. Finally, some NP 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 NP 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 like compact identifiers. This presents a significant challenge for database development[152]. Consequently, database developers must manually curate articles to convert them into structured data formats. Database completeness is also hampered by the broad spectrum of journals that feature NP research. Curation difficulties include image-to-structure conversion, absence of core data (*e.g.,* BGC sequence), resolving name conflicts (multiple structures with the same name, or structures with multiple names) and extracting data and metadata for biological assays. Improvements are underway for structure recognition from images using DECIMER 1.0[106,153] and through new formats for reporting of chemical-structure data[154]. 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 currently governing journal articles. Finally, because most NP databases focus on only one feature of NP data, there is presently high redundancy in curation efforts, as the existence of minor variations in the extracted data (*e.g.*, 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 critical information about each structure (*e.g.,* SMILES, compound name, availability and location of spectral data, source organism, BGC) and would offer a central point of reference for data dissemination and automated database importation by NP-centric resources.

Data deposition

Several of the larger NP data repositories, including Minimum Information about a Biosynthetic Gene cluster (MIBiG)[155], the Natural Products Atlas[156,157], Global Natural Product Social Molecular Networking (GNPS)[158], Natural Products Magnetic Resonance Database (NP-MRD)[159] and Norine[160] offer mechanisms to accept user-deposited data (Figure 4). 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/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 NP-related data types (*e.g.*, source organisms, MS, NMR, BGC, SMILES, etc.) amplify this, and increase the number of platforms 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 specialty databases (Figure 4).

Two principal avenues exist to incentivize data deposition to public repositories: 'value added' and 'requirements'. Firstly, authorships during data-'curatathons', increased citations, opportunities for collaboration, and facilitated automated re-analysis are very beneficial for depositors[161]. An example of added value is ReDU, which aids in rapid re-analysis of existing and future data[162] through subscription to one's own and public datasets[158]. Alternatively, repositories can offer validation reports, quality metrics, prevalence statistics (*e.g.*, the statistics page of MIBiG that facilitates cross-species comparisons of biosynthetic potential), and other feedback on data to depositors that provides a tangible and immediate benefit to deposition[155]. 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 extended processes should become more user-friendly, for example by including an autofill during metadata reporting, employing 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. Secondly, 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 will require the deposition of raw NMR data starting in January 2023[163]. 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 select few laboratories. Typically, however, the value of these datasets is limited due 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 LCMS/MS analyses[164]. As a result, a global assemblage of data would be incredibly valuable; yet

challenges exist of poor interoperability (*i.e.*, connecting data between resources) and weak compatibility (*i.e.*, resources use different standards and ontologies to annotate and identify their contents). It is important to note that the quality of biologically-derived data (*e.g.*, 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[165], 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 (*e.g.*, clustering MS/MS spectra[166], binning metagenomes[167]). 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 NP databases. At present, most database managers communicate updates on an *ad hoc* basis. In addition, some databases like NP Atlas maintain interoperable 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 (*i.e.*, controlled vocabulary). Open standards play 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 NPs[168]. Furthermore, the adoption of Universal Spectrum Identifiers (USIs) for identifying mass spectra in proteomics[169] and metabolomics[170] showcase standardization tools, enabling data analysis across datasets. Such tools play a pivotal role in enabling large-scale studies by structuring omics data and represent an area of development that the NP community should consider. The implementation of semantic web approaches is also an essential step forward, which standardizes how we disseminate knowledge and data and integrate exchange formats, linking between resources, and ontological representation[171]. An overview of current NP ontologies is provided in Table 2.

The need for standardization is apparent in describing bioactivities of natural products and ensuring that experimental conditions are comparable between laboratories. While standards exist for reporting the biological activities of purified compounds (*e.g.* ChEMBL[150], PubChem[172], Supernatural II[173], NPASS[174]), 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 NP databases. Finally, as further discussed in Section 3.5, experimental conditions must be described as accurately as possible, with scientists preferably using the same growth conditions for their experiments. Overall, while 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 (e.g. sample taxonomy, extract preparation protocol, instrument parameters) the addition of contextual annotations can greatly increase the value of NP datasets. For example, accurate annotation of compound structures to metabolomics

datasets would provide many opportunities to build ML models that integrate structural and biological/ 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, while 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.

Integration of data

The value of linked/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 only be answered 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 NP 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 NP Linker platforms [175–182]. In the first case the objective is to define which data types exist for each sample, while in the second case the goal is to perform paired analyses where both data types are mined at the same time [183]. 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 [184]. Also, transcriptomics has been used as a constraints 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 where certain metabolite features are observed [185]

Methodology and opportunities for data integration

Data integration faces several current challenges that are mostly centered 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[104], GraphOmics[187] and anvi’o[188] 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 due 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 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.

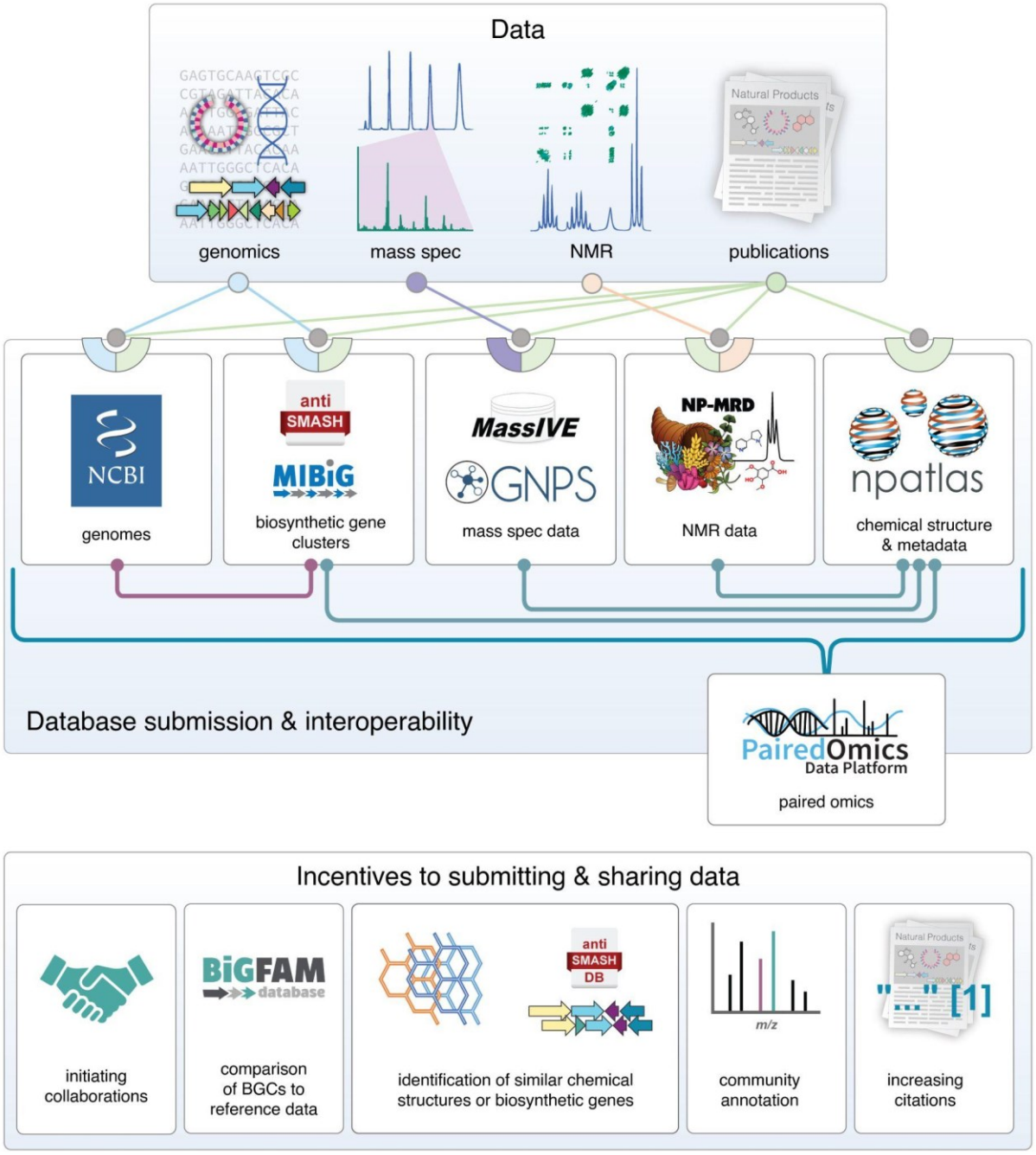

Figure 4. Depositing and sharing natural product data: infrastructure and incentives.

Toward the future: 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. While 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 various quality

due 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 NP-based training and benchmarking sets

Chemical structure and biosynthetic data for natural products are now reasonably well-standardized and centralized. For example, the Natural Products Atlas[189,190], COCONUT[191] and LOTUS[192] databases provide information about chemical structures, while the MIBiG database contains information on BGCs[155]. 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[193], *de novo* BGC predictions [13,14], matching of chemical structures to their mass spectra[194], automated chemical classification of NP structures[139], and the identification of unknown metabolites from NMR spectral matching[45].

Using Universal Spectrum Identifiers (USIs) for mass spectra will enable standardized access to the mass spectral data of NPs enabling the community to easily retrieve and visualize the underlying raw data. In this regard, spectral databases for natural products are under active development, such as the Global Natural Products Social Molecular Network (GNPS) for MS and MS/MS data and the Natural Products Magnetic Resonance Database (NP-MRD) for NMR data. Importantly, entries in MIBiG, GNPS, and the NP-MRD are now all cross-linked to the Natural Products Atlas, creating a central hub that connects structural, spectroscopic, and biosynthetic data for natural products.

By contrast, two areas lacking NP database coverage are catalytic activities of biosynthetic tailoring enzymes (key to predicting NP structures) and biological activities (key to understanding structure-activity and structure-property relationships). In the former case, 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 NPs, or NP 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.

While 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 mass spectrometry 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 critical type of data to guide future natural product drug discovery. At the same time, these data are currently the least standardized and systematically documented. While databases such as ChEMBL[150] can host such data, stored using standardized ontologies[195,196], the vast majority of natural product activity data is never deposited. It 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 culturing conditions would be highly beneficial. Some efforts already do exist that generate specific types of data. For example, the NCI60 panel of tumor 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[197]. Similarly, CO-ADD constitutes a community-driven approach to antibiotic discovery[198], allowing compounds to be sent to a central location to test their activities according to standardized protocols.

The future of NP data repositories

Because of the vast array of data types associated with NP research it is unlikely that a single monolithic repository will serve the needs of the NP community. Instead, specialized repositories that focus on different aspects of NP data (e.g., structures, BGCs, spectral data, biological activities, *etc.*) must focus on improving interoperability to develop a distributed network of data resources. This interoperability must not only involve the connection of entries between databases but must also 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 (PDB) for structural biology and the Cambridge Structural Database (CSD) for X-ray crystallography. The NP community must prioritize and promote these efforts if they are to benefit from the new and exciting applications being offered by AI-based technologies.

Biological activity and target prediction

One of the most important application areas for AI in natural product drug discovery is predicting associated biological activities, macromolecular targets, and possible toxicities. Accurate predictions of these characteristics will provide direct clues as to which areas of molecular structure and sequence space are most promising for drug discovery. This will be key to the potential success of genome mining, which currently suffers from yielding too large lists of candidate BGCs with few strategies available to target efforts towards parts of NP space with actual pharmaceutical potential. Below, as a case study, we will discuss how AI techniques (combined with other technologies) can make a difference in addressing this challenge (Figure 5).

Visualizing and navigating chemical space

One strategy to successfully develop new drugs is to target regions of chemical space that are structurally novel and have favorable drug-like properties. However, it should be noted that there is a mismatch between the typical chemical features of natural products and drug-like properties[199,200]. The success of finding such molecules largely depends on a) the nature of the dataset that is used for screening and b) the underlying search algorithm (usually based on some form of similarity). In contrast to more conservative approaches, which try to identify activities based on analog series, screening virtual libraries sampled from an estimated 10^{63}

drug-like molecules has become feasible thanks to the latest improvements in AI-related technologies.

The concept of chemical space itself is somewhat unsatisfactorily defined in cheminformatics, when it comes to the space spanned by chemicals based on their properties. For visualization purposes, a high-dimensional space will be reduced to only two or three dimensions. Also, depending on the properties of interest (physicochemical properties, target profiles, toxicity, *etc*.), the chemical space to be explored will be constructed differently. Still, given the impressive number of underexplored possibilities, taking the challenge of solving the multiparameter optimization problem to navigate chemical space is a very promising strategy[201–203].

Advances in technologies to navigate chemical space

Chemical space is vast. Exploring it is a daunting task, not only because of the sheer quantity of compounds that can be (virtually) enumerated but also because the description and labeling of compounds is by definition a multidimensional problem. To navigate chemical space visually we therefore need dimensionality-reduction approaches. A common way to reduce dimensionalities is via principal component analysis (PCA). PCA of chemical properties has revealed that both drug molecules and NPs occupy a very similar topological diversity distribution, which was not the case for combinatorial compounds. Another method is t-distributed stochastic neighbor embedding (t-SNE), which has been used successfully for the design of new drug classes, for example new kinase inhibitors[201]. 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 data sets[202]. More recently, a TMAP (Tree MAP) algorithm was developed to visualize data sets with sample sizes up to around 10^7 in a tree layout [203]. Using TMAP, a tree of all the compounds in the ChEMBL database (1.13 million) with their associated biological-assay data was constructed within a mere 10 minutes.

The application of unsupervised learning approaches (*e.g.*, 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 proven successful in the small-molecule discovery field and we believe its application to NPs will open up new avenues to characterize and address, amongst 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 NP and small molecules, to identify overlapping chemical space and ultimately transfer knowledge between the two fields.

Predicting bioactivities from chemical and protein-structure data

Classical cheminformatics and pharmacophore-based predictions

Methods relying on the use of classical cheminformatics and computer-assisted drug-discovery tools for predicting bioactivities for natural products are plentiful[51]. For example, the direct application of the popular prediction methods PASS[204] and SEA[205] to natural products have shown some successes. Given the distinct chemical structures and physicochemical properties of natural products[53,206], 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 was specifically developed to predict the bioactivities of molecules and has been successfully applied to predict the biological activity of macrocyclic natural products[53,206] and fragment-like natural products [55]. Other successful applications of bioactivity predictions have used representations such as 3D pharmacophores[55], bioactivity signatures[78,207], and learned representations[207], which capture essential properties of natural products without directly employing classic chemical fingerprints. Notably, learned representations aided in the prediction of the bactericidal activity of halicin and eight additional molecules with antibiotic properties structurally-distinct from known antibiotic classes [208]."

Molecular-dynamics simulations and structure-based predictions

Structure-based approaches leverage a protein target's spatial information to predict a compound's binding mode. This information can be obtained from experimentally determined structures or modeling approaches such as AlphaFold.[108] 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 (GPU computing) and software improvements [209]. Structure-based methods can provide a wealth of information, an example is the free-energy perturbation (FEP) method, which recently grew in popularity and saw a substantial increase in the applicability of the method in drug discovery in academic and industrial settings[210] Molecular docking, molecular-dynamics, and FEP could be extended to study affinities of NPs. Besides binding affinities, the computational prediction of enzyme kinetics, for instance using the empirical valence bond (EVB) approach, holds promise for the design of artificial biocatalysts[211].

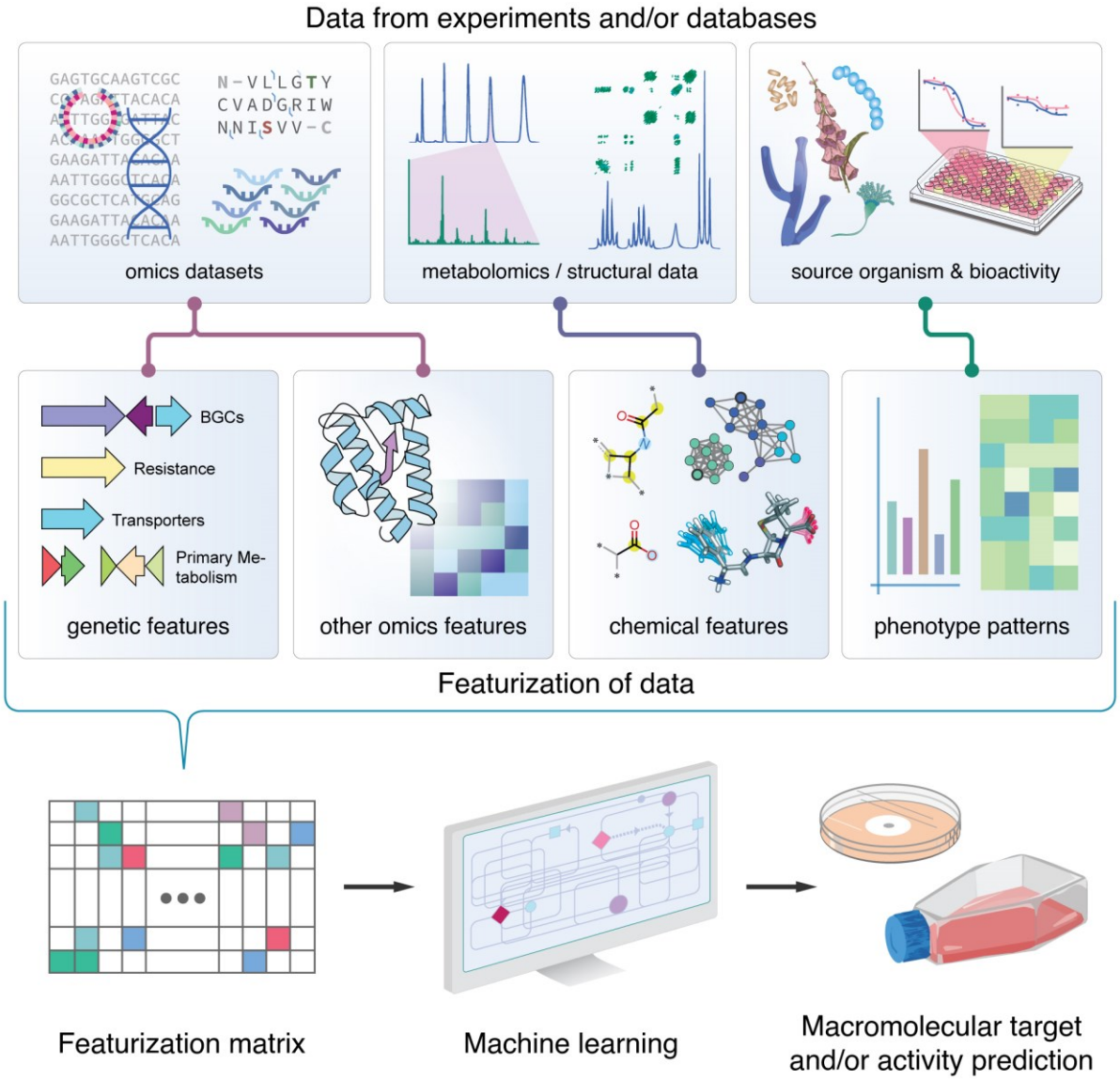

Figure 5. Predicting biological activities and macromolecular targets from genomic, metabolomic and phenotypic data. Omics datasets can be mined to identify genetic features of NP biosynthetic pathways, such as resistance genes, transporters and links with primary metabolism, which are predictive of 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 for certain activities or targets. Finally, large-scale standardized phenotypic bioassays are key. There is considerable potential for AI approaches to then predict targets and activities based on combined sets of genetic/chemical features of NPs and their biosynthetic pathways.

Sequence- or BGC-based predictions and natural language processing

In recent years, a growing number of approaches have been used to predict bioactivities based on DNA/protein sequence data with machine learning, while other strategies have the potential to do so in the near future. The sequence boundaries for BGCs predicted by mining tools are not precise, often missing portions of the BGC or fusing with others. Often it is necessary for an expert to manually update the BGC boundaries. Improvement in BGC prediction is vital for bioactivity prediction methods and remains an active area where further research is needed.

One approach that leverages knowledge of existing small molecules is to predict the final product of a BGC and infer its activity directly, as exemplified by PRISM [9]. 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, using of substructural features such as β-lactam rings or specific amino acids may produce more accurate results for a broader range of BGCs.

Alternate approaches emerging for bioactivity prediction draw on the field of natural language processing (NLP). NLP-based methods like word2vec[212], originally developed for context-aware embedding of words within sentences in text documents, have been extended to embed protein domains within BGCs using pfam2vec[213]. DeepBGC, a *de novo* BGC prediction tool, 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 [214]. Just as NLP has revolutionized other fields, we expect continued, rapid advances in applications of NLP for BGC and bioactivity prediction.

Bioactivity predictions based on self-resistance, regulatory, or evolutionary features

Bacteria have long been known to harbor resistance to their own antibacterials *via* resistance genes [213], and numerous antimicrobial resistance determinant databases are available (*e.g.*, the comprehensive antibiotic resistance database [CARD][215], a national database of antibiotic resistant organisms [NDARO][216], and ResFinder[217]. To leverage resistance information, various algorithms were created to attempt to link these resistance genes with BGCs, as the resistance genes are necessary to confer immunity in the host[218,219]. Walker and Clardy recently 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 [220].

As an additional layer of biological information, transcription factor networks and their cognate regulatory elements can be used to cluster BGCs based on how they are controlled and to which (environmental) signals they respond. The EvoMining framework[221] is based on the concept that streptomycetes adapt to their ecological niche by evolving their primary and secondary metabolism in response to their environment[222]. 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 prediction. While BGCs predict what types of metabolites may be produced, regulatory networks can be harnessed to cluster BGCs on how they are controlled and – notably – in response to which signals. This information may serve as a beacon to find BGCs/metabolites required for specific purposes (stress, 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 prioritize BGCs for antibiotic discovery.

Algorithmic innovations for activity prediction

Algorithmic innovations will only improve performance if training datasets are sufficient to support model complexity. One solution to reduce the number of effective 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[223] or MoleculeNet[82] reduces the computational load required to train new models from scratch. In many cases, pre-trained models will also yield higher prediction accuracies[224].

Conclusions and future 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. Rather than 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, e.g., generic computer-vision-related datasets 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, many breakthroughs in the field have been made by crossing disciplinary boundaries to draw on algorithms from other fields, such as natural language processing. 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 adopting an 'active-learning' approach toward dataset generation. By this, we mean characterizing underexplored areas of sequence, chemical, structural, or bioactivity space where gold-standard datasets are lacking to increase the number of effective data points. New data-driven AI discoveries depend on underlying databases being preserved and maintained over time. Ironically, while 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. In addition, it is important to realize 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; therefore, investments in fundamental biochemical research needs to be continued as well to shed light on those parts of biochemical space on which AI currently does not yet provide meaningful insights [225]. As a final outlook, 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.

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 Discovery 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. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research* **48,** D454-D458 (2019).
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. 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).
8. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **49,** W29–W35 (2021).
9. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
10. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).
11. 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).
12. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
13. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).

14. Carroll, L. M. *et al.* Accurate de novo identification of biosynthetic gene clusters with GECCO. *BioRxiv*, https://doi.org/10.1101/2021.05.03.442509 (2021).
15. 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).
16. de Los Santos, E. L. C. NeuRiPP: Neural network identification of RiPP precursor peptides. *Sci. Rep.* **9**, 13406 (2019).
17. Merwin, N. J. *et al.* DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 371–380 (2020).
18. Tietz, J. I. *et al.* A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **13**, 470–478 (2017).
19. de Los Santos, E. L. C. NeuRiPP: Neural network identification of RiPP precursor peptides. *Sci. Rep.* **9**, 13406 (2019).
20. Merwin, N. J. *et al.* DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 371–380 (2020).
21. 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).
22. Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **17**, e1008724 (2021).
23. 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. *Bioinformatics* **13,** 84 (2021).
24. Hoffmann, M. A. *et al.* Assigning confidence to structural annotations from mass spectra with COSMIC. *BioRxiv*, https://doi.org/10.1101/2021.03.18.435634 (2021).
25. Ludwig, M. *et al.* Database-independent molecular formula annotation using Gibbs

sampling through ZODIAC. *Nature Machine Intelligence* **2,** 629–641 (2020).

26. Kim, H. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **84,** 2795-2807 (2021).
27. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2021).
28. 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).
29. 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).
30. 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).
31. 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).
32. 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).
33. Zheng, S. *et al.* Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat. Commun.* **13**, 1–9 (2022).
34. Milanowski, D. J. *et al.* Unequivocal determination of caulamidines A and B: application and validation of new tools in the structure elucidation tool box. *Chemical Science* **9,** 307–314 (2018).
35. Audoin, C. *et al.* Metabolome consistency: additional parazoanthines from the mediterranean zoanthid parazoanthus axinellae. *Metabolites* **4**, 421–432 (2014).
36. Fox Ramos, A. E. *et al.* CANPA: Computer-Assisted Natural Products Anticipation. *Anal. Chem.* **91**, 11247–11252 (2019).

37. 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).
38. Kim, L. J. *et al.* Prospecting for natural products by genome mining and microcrystal electron diffraction. *Nat. Chem. Biol.* **17**, 872–877 (2021).
39. 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. U. S. A.* **112**, 12580–12585 (2015).
40. Lindsay, R. K. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*. (McGraw-Hill Companies, 1980).
41. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
42. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: De novo structure generation from mass spectra. *Nat. Methods* **19,** 865-870 (2022).
43. 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).
44. 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).
45. 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).
46. 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. *Magnetic Resonance in Chemistry* **60,** 1070–1075 (2022).
47. 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).

48. 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).
49. Howarth, A., Ermanis, K. & Goodman, J. DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure. *Chem. Sci.* **11,** 4351-4359 (2020).
50. Das, S., Edison, A. S. & Merz, K. M., Jr. Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal. Chem.* **92**, 10412–10419 (2020).
51. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).
52. 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).
53. Reker, D. *et al.* Revealing the macromolecular targets of complex natural products. *Nat. Chem.* **6**, 1072–1078 (2014).
54. 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).
55. 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).
56. Reker, D. *et al.* Machine Learning Uncovers Food- and Excipient-Drug Interactions. *Cell Rep.* **30**, 3710–3716.e4 (2020).
57. Conde, J. *et al.* Allosteric Antagonist Modulation of TRPV2 by Piperlongumine Impairs Glioblastoma Progression. *ACS Cent Sci* **7**, 868–881 (2021).
58. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
59. Landrum, G. & Others. RDKit: Open-Source Cheminformatics Software. *URL: http://www. rdkit. org/, https://github. com/rdkit/rdkit* (2016).
60. Willighagen, E. L. *et al.* The Chemistry Development Kit (CDK) v2.0: atom typing,

depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics* **9,** 33 (2017).

61. Todeschini, R. & Consonni, V. *Handbook of Molecular Descriptors*. (John Wiley & Sons, 2008).
62. 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. *Journal of Cheminformatics* **9,** 46 (2017).
63. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
64. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* **5**, 26 (2013).
65. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **8**, 36 (2016).
66. Grisoni, F. *et al.* Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Communications Chemistry* **1**, 1–9 (2018).
67. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
68. Capecchi, A. & Reymond, J.-L. Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning. *Biomolecules* **10**, (2020).
69. Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **57**, 726–741 (2017).
70. 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).
71. Bannan, C. C. *et al.* Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design* **30,** 927–944 (2016).
72. 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).

73. 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).
74. 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).
75. Kauvar, L. M. *et al.* Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2**, 107–118 (1995).
76. Petrone, P. M. *et al.* Rethinking molecular similarity: comparing compounds on the basis of biological activity. *ACS Chem. Biol.* **7**, 1399–1409 (2012).
77. Norinder, U., Spjuth, O. & Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J. Chem. Inf. Model.* **60**, 2830–2837 (2020).
78. Bertoni, M. *et al.* Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **12**, 3932 (2021).
79. Mater, A. C. & Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **59**, 2545–2559 (2019).
80. Bronstein, M. M., Bruna, J., Cohen, T. & Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv*, https://doi.org/10.48550/arXiv.2104.13478 (2021).
81. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
82. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
83. 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).
84. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. (Springer Nature, 2019).

85. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2**, 573–584 (2020).

86. 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).

87. Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S. & Unterthiner, T. Interpretable Deep Learning in Drug Discovery. 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).

88. Webel, H. E. *et al.* Revealing cytotoxic substructures in molecules using deep learning. *J. Comput. Aided Mol. Des.* **34**, 731–746 (2020).

89. 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).

90. 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).

91. Duvenaud, D. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* 28, https://doi.org/10.48550/arXiv.1509.09292 (2015).

92. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc Mach Learn Res.* **70,** 1263-1272 (2017).

93. Nguyen, T. *et al.* GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2021).

94. Yuan, W. *et al.* Chemical Space Mimicry for Drug Discovery. *J. Chem. Inf. Model.* **57**, 875–882 (2017).

95. 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).

96. Li, X. & Fourches, D. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *J. Cheminform.* **12**, 27 (2020).
97. Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminform.* **12**, 17 (2020).
98. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
99. 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).
100. Bjerrum, E. J. & Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular Generation Diversity with Heteroencoders. *Biomolecules* **8**, 131 (2018).
101. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4**, 268–276 (2018).
102. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nature Machine Intelligence* **3**, 1023–1032 (2021).
103. 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).
104. 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).
105. Navarro-Muñoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
106. Rajan, K., Zielesny, A. & Steinbeck, C. DECIMER 1.0: Deep Learning for Chemical Image Recognition using Transformers. *J. Cheminform.* **13,** 61 (2021).
107. 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.*

U. S. A. **113**, 13738–13743 (2016).

108. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

109. Callaway, E. After AlphaFold: protein-folding contest seeks next big breakthrough. *Nature* **613**, 13–14 (2023).

110. Wallner, B. AFsample: Improving Multimer Prediction with AlphaFold using Aggressive Sampling. *BioRxiv*, https://doi.org/10.1101/2022.12.20.521205 (2022).

111. 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).

112. 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).

113. Sydow, D., Rodríguez-Guerra, J. & Volkamer, A. Teaching Computer-Aided Drug Design Using TeachOpenCADD. in *Teaching Programming across the Chemistry Curriculum* **1387**, 135–158 (American Chemical Society, 2021).

114. 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).

115. 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).

116. 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).

117. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1,** 206–215 (2019).

118. 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).

119. Quinn, T. P. & Erb, I. Examining microbe–metabolite correlations by linear methods. *Nature Methods* **18,** 37–39 (2021).

120. 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).

121. Soleimany, A. P. *et al.* Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Cent Sci* **7**, 1356–1367 (2021).

122. Manica, M. *et al.* Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. *Mol. Pharm.* **16**, 4797–4806 (2019).

123. Jimenez-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).

124. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? (2022) doi:10.48550/arXiv.2207.08815.

125. Rothfritz, L. The FAIR data principles. Preprint at https://doi.org/10.14293/s2199-1006.1.sor-compsci.clnbrup.v1 (2019).

126. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem.* **55**, 5165–5173 (2012).

127. Tiikkainen, P., Bellis, L., Light, Y. & Franke, L. Estimating error rates in bioactivity databases. *J. Chem. Inf. Model.* **53**, 2499–2505 (2013).

128. Chapelle, O., Scholkopf, B. & Zien, A., Eds. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Trans. Neural Netw.* **20**, 542–542 (2009).

129. 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).

130. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain

specificity. *Nucleic Acids Research* **39,** W362–W367 (2011).

131. Torrey, L. & Shavlik, J. Transfer Learning. in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* 242–264 (IGI Global, 2010).

132. Cai, C. *et al.* Transfer Learning for Drug Discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).

133. 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).

134. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nature Machine Intelligence* **2**, 171–180 (2020).

135. Moret, M. *et al.* Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nat. Commun.* **14**, 1–12 (2023).

136. Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discov. Today Technol.* **32-33**, 73–79 (2019).

137. 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).

138. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 61 (2016).

139. Kim, H. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **84,** 2795-2807 (2021).

140. Reher, R. *et al.* Native metabolomics identifies the rivulariapeptolide family of protease inhibitors. *Nat. Commun.* **13**, 4619 (2022).

141. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).

142. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci Adv* **4**, eaap7885 (2018).

143. 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).

144. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).

145. Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, (2019).

146. 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).

147. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).

148. Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *J. Cheminform.* **12**, 1–51 (2020).

149. Wikidata Query. URL: https://w.wiki/5bpq.

150. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

151. 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**, D198–201 (2007).

152. Wimalaratne, S. M. *et al.* Uniform resolution of compact identifiers for biomedical data. *Sci Data* **5**, 180029 (2018).

153. 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).

154. Schymanski, E. L. & Bolton, E. E. FAIR chemical structures in the Journal of Cheminformatics. *J. Cheminform.* **13**, 50 (2021).

155. Kautsar, S. A. *et al.* MIBiG 2.0: a repository for biosynthetic gene clusters of known

function. *Nucleic Acids Res.* **48**, D454–D458 (2020).

156. 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).

157. 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).

158. 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).

159. Wishart, D. S. *et al.* NP-MRD: the Natural Products Magnetic Resonance Database. *Nucleic Acids Res.* **50,** D665-D677 (2021).

160. Flissi, A. *et al.* Norine: update of the nonribosomal peptide resource. *Nucleic Acids Res.* **48**, D465–D469 (2020).

161. 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).

162. Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).

163. Proteau, P. J. *Journal of Natural Products* 2022: Perspectives, Monthly Cover Art, and More. *Journal of Natural Products* **85,** 1–2 (2022).

164. Clark, T. N. *et al.* Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability. *Journal of Natural Products* **84,** 824–835 (2021).

165. Fiehn, O. *et al.* The metabolomics standards initiative (MSI). *Metabolomics* **3**, 175–178 (2007).

166. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).

167. Miller, I. J. *et al.* Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic Acids Res.* **47**, e57 (2019).

168. Schymanski, E. L. *et al.* Identifying small molecules via high resolution mass

spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).

169. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).

170. Bittremieux, W. *et al.* Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. *BioRxiv*, https://doi.org/10.1101/2020.05.09.086066 (2020).

171. Gordon, J. E. Chemical inference. 2. Formalization of the language of organic chemistry: generic systematic nomenclature. *Journal of Chemical Information and Computer Sciences* **24,** 81–92 (1984).

172. Wang, Y. *et al.* PubChem's BioAssay Database. *Nucleic Acids Res.* **40**, D400–12 (2012).

173. Banerjee, P. *et al.* Super Natural II--a database of natural products. *Nucleic Acids Res.* **43**, D935–9 (2015).

174. 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).

175. Hooft, J. J. J. van der A community-driven paired data platform to accelerate natural product mining by combining structural information from genomes and metabolomes. *Scientific Symposium FAIR Data Sciences for Green Life Sciences* (2018).

176. 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).

177. Schorn, M. A. *et al.* A community resource for paired genomic and metabolomic data mining. *Nat. Chem. Biol.* **17**, 363–368 (2021).

178. 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).

179. 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).

180. 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).

181. 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).

182. Caesar, L. K. *et al.* Correlative metabologenomics of 110 fungi reveals metabolite-gene cluster pairs. *Nat. Chem. Biol.* (2023) doi:10.1038/s41589-023-01276-8.

183. Soldatou, S. *et al.* Comparative Metabologenomics Analysis of Polar Actinomycetes. *Mar. Drugs* **19**, 103 (2021).

184. Sulheim, S. *et al.* Enzyme-Constrained Models and Omics Analysis of Streptomyces coelicolor Reveal Metabolic Changes that Enhance Heterologous Production. *iScience* **23**, 101525 (2020).

185. Amos, G. C. A. *et al.* Comparative transcriptomics as a guide to natural product discovery and biosynthetic gene cluster functionality. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E11121–E11130 (2017).

186. Pascal Andreu, V. *et al.* BiG-MAP: an Automated Pipeline To Profile Metabolic Gene Cluster Abundance and Expression in Microbiomes. *mSystems* **6,** e0093721 (2021).

187. Wandy, J. & Daly, R. GraphOmics: An Interactive Platform To Explore And Integrate Multi-Omics Data. *BMC Bioinformatics* **22,** 603 (2021).

188. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology* **6**, 3–6 (2020).

189. 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).

190. 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).

191. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT

online: Collection of Open Natural Products database. *J. Cheminform.* **13**, 2 (2021).

192. Rutz, A. *et al.* The LOTUS initiative for open knowledge management in natural products research. *Elife* **11**, e70780 (2022).

193. 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).

194. Cao, L. *et al.* MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nat. Commun.* **12**, 3718 (2021).

195. Visser, U. *et al.* BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics* **12,** 257 (2011).

196. Sarntivijai, S. *et al.* CLO: The cell line ontology. *J. Biomed. Semantics* **5**, 37 (2014).

197. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).

198. Cooper, M. A. A community-based approach to new antibiotic discovery. *Nat. Rev. Drug Discov.* **14**, 587–588 (2015).

199. O'Shea, R. & Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem.* **51**, 2871–2878 (2008).

200. Reck, F., Jansen, J. M. & Moser, H. E. Challenges of antibacterial drug discovery. *Arkivoc* **2019 part iv**, 227–244 (2019).

201. Janssen, A. P. A. *et al.* Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome–Inhibitor Interaction Landscapes. *Journal of Chemical Information and Modeling* **59,** 1221–1229 (2019).

202. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3,** 861 (2018).

203. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).

204. 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).

205. Sá, M. S. *et al.* Antimalarial Activity of Physalins B, D, F, and G. *Journal of Natural Products* **74,** 2269–2272 (2011).

206. Schneider, G. *et al.* Deorphaning the Macromolecular Targets of the Natural Anticancer Compound Doliculide. *Angew. Chem. Int. Ed Engl.* **55**, 12408–12411 (2016).

207. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **181**, 475–483 (2020).

208. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702.e13 (2020).

209. Pandey, M. *et al.* The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence* **4**, 211–221 (2022).

210. 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).

211. Duarte, F., Pabis, A. & Kamerlin, S. C. L. Introduction to the Empirical Valence Bond Approach. *Theory and Applications of the Empirical Valence Bond Approach* 27–61 Preprint at https://doi.org/10.1002/9781119245544.ch2 (2017).

212. Mikolov, T. *et al.* word2vec. *URL https://code. google. com/p/word2vec* **22**, (2013).

213. Flemming, A. Resistance-guided discovery of new antibiotics. *Nat. Rev. Drug Discov.* **12**, 826–826 (2013).

214. Yang, Z. *et al.* Deep-BGCpred: A unified deep learning genome-mining framework for biosynthetic gene cluster prediction. *bioRxiv* (2021) doi:10.1101/2021.11.15.468547.

215. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).

216. National Database of Antibiotic Resistant Organisms (NDARO). https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/.

217. Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy* **75,** 3491–3500 (2020).

218. Mungan, M. D. *et al.* ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Research* **48,** W546–W552 (2020).

219. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45**, D566–D573 (2017).

220. 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).

221. 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).

222. Chevrette, M. G. *et al.* Evolutionary dynamics of natural product biosynthesis in bacteria. *Nat. Prod. Rep.* **37**, 566–599 (2020).

223. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv [cs.LG]* (2020).

224. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: A pre-trained Transformer for computational chemistry. *Mach. Learn.: Sci. Technol.* (2021) doi:10.1088/2632-2153/ac3ffb.

225. 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).

Funding

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. MWM was supported by funds from the Duchossois Family Institute at the University of Chicago. Katherine R. Duncan was supported by the United Kingdom Research and Innovation Biotechnology and Biological Sciences Research Council (BB/R022054/1). NG was supported by an NSF CAREER award (award number 2047235). JJJvdH was supported by an ASDI eScience grant from the Netherlands eScience Center (award number ASDI.2017.030). NIM is supported by funding from the European Research Council (ERC consolidator grant agreement no. 725523). KB was supported by a Novo Nordisk Foundation grant NNF20CC0035580. MGG was supported by ONCODE funding. EJNH was supported by the LOEWE Center for Translational Biodiversity Genomics and the Funds of the Chemical Industry Germany. MS was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 239748522, SFB 1127 ChemBioSys. MAB was supported by the National French Agency (ANR grants 15-CE29-0001 and 20-CE43-0010). CMC was supported by a National Library of Medicine training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM 5T15LM007359) SF was supported by MASTS/IbioIC/Xanthella. OVK was funded by the Klaus Faber Foundation. HK was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) [Grants NRF 2018R1A5A2023127 and NRF 2022R1F1A107462311]. JM was supported by a grant from the Research Foundation - Flanders (G061821N). ERR was supported by the U.S. National Science Foundation (DBI-1845890). DR was supported, in part, by a grant from NC Biotech (2021-FLG-3819), the NIH (P30 DK034987), the Duke Cancer Institute, the Duke Science and Technology Initiative and the UNC CGIBD. PS acknowledges support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. NZ was supported by Germany’s Excellence Strategy – EXC 2124–390838134. HUK was supported by the KAIST Key Research Institute (Interdisciplinary Research Group) Project. SLR was supported by Eawag discretionary funding. MHM was supported by the Leiden University 'van der Klaauw' chair for theoretical biology and an ERC Starting Grant (DECIPHER-948770).

Competing interests

JJJvdH is a member of the Scientific Advisory Board of NAICONS Srl., Milan, Italy. CAD is a founding member of Adapsyn Bioscience. MAS is a consultant to Adapsyn Bioscience. MHM is on the scientific advisory board of Hexagon Bio and co-founder of Design Pharmaceuticals.

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:

1. Overrepresentation 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 prior to 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.
2. Overrepresentation 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 overrepresentation 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 overrepresented phylogenetic branches or compound classes. Poor predictions on underrepresented 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 underrepresented clades do not appear in the test set at all.

These data imbalanced have to be targeted at three stages of model development:

1. Data selection for training and test sets prior to 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 which would yield an automatic correct prediction. Proportional division of the resulting data points across training and test sets based on class and feature labels (e.g. 80% training 20% test for each label) ensures that the model can be separately evaluated on each data subclass, resulting in more accurate model evaluation.
2. Sampling and data weighting during model training. When a model is not instructed otherwise, it will prioritise overall accuracy. Often, this means that the model tolerates mispredictions for underrepresented data classes. To prevent this, data can be weighted during model training: underrepresented classes should receive higher weights or contribute more towards a model's loss function such that the model penalises prediction errors for those classes more than prediction errors for overrepresented 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 overrepresented data labels or features.

3. 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/false positive/negative rates, and for data features by assessing performance for each sequence/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 prior to 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.

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 & archive available to download	Has a user submission system for new data upload	Link	Primary reference	Licence
Chemical NP-specific resources													
LOTUS	yes	yes	yes	all NP	yes	yes	producer taxonomy	molecular descriptors, chemical classification, bioactivities	yes	no	LOTUS	10.1101/2021.02.28.433265	CC0
COCONUT	yes	yes	yes	all NP	yes	yes	none	molecular descriptors, chemical classification, bioactivities	yes	no	COCONUT	10.1186/s13321-020-00478-9	CC BY-SA
NP Atlas	yes	yes	yes	microbial NP	yes	yes	producer taxonomy	chemical classification	yes	yes	Natural Products Atlas	10.1021/acscentsci.9b00806	CC BY
BGC resources													
MIBiG	yes	yes	yes	microbial NP	yes	yes	BGC genomic co-ordinates and gene function annotation; compound produced by BGC	antiSMASH annotations	yes	yes	MIBiG:	10.1093/nar/gkz882	CC BY
antiSMASH DB	yes	no	no	microbial NP	yes	yes	none	BGC genomic co-ordinates and gene function annotations; compounds produced by BGC	yes	no	antiSMASH database	10.1093/nar/gkaa978	CC BY
PRISM	no	no	no	microbial NP	no	yes	genomic co-ordinates and gene function annotation; compound produced by BGC	BGC genomic co-ordinates and gene function annotations; compounds produced by BGC	yes	no	PRISM	10.1038/s41467-020-19986-1	CC BY
Spectral resources													
GNPS	no	yes	yes	yes	yes	yes			no	yes	GNPS	10.1038/nbt.3597	CC0
MassBank	yes	yes	yes	no	no	yes	MS and tandem MS spectra	none	no	yes	MassBank	10.1002/jms.1777	CC BY-NC
NP-MRD	yes	yes	yes	yes	no	yes	NMR	no	yes	yes	NP-MRD	10.1093/nar/gkab1052	CC BY
CH-NMR-NP	Yes (CAS Registry No.)	Yes	Yes	All NP	No	No	NMR Producer	Molecular weight	No	No	CH-NMR-NP	-	-
MetaboLights	yes	yes	yes	no	no	yes	MS and tandem MS spectra; NMR	none	no	yes	MetaboLights	10.1093/nar/gkz1019	EMBL-EBI's Terms of use
Paired Omics Data Platform	no	no	yes	yes	no	yes	LC MS, genomics	none	no	yes	Paired Omics Data Platform	10.1038/s41589-020-00724-z	CC BY
NMRshiftDB	no	yes	yes	no	yes	no	NMR	calculated NMR	no	yes	nmrshiftdb2	10.1002/mrc.4263	Modified CC

													BY
NP-friendly useful resources													
ZINC	yes	yes	no	no	yes	no	none	molecular descriptors, bioactivities	no	yes	ZINC20	10.1021/acs.jcim.0c00675	CC0
ChEBI	yes	yes	yes	no	yes	yes	none	chemical classification, bioactivities	yes	yes	ChEBI	10.1093/nar/gkv1031	CC BY
ChEMBL	yes	yes	yes	no	yes	yes	bioactivities	molecular descriptors	yes	yes	ChEMBL	10.1093/nar/gkw1074	CC BY-SA
WikiPathways	yes	no	yes	no	yes	yes	metabolic networks	none	yes	yes	WikiPathways	10.1093/nar/gkaa1024	CC0
Reactome	yes	no	yes	no	yes	yes	metabolic networks	metabolic networks	yes	no	Reactome	10.1093/nar/gkz1031	CC0
CO-ADD	yes	yes	no	no	no	yes	bioactivities	none	no	yes	CO-ADD	10.1021/acsinfecdis.5b00044	Copyright (c) 2016 The University of Queensland
Wikidata	yes	yes	yes	no	yes	yes	none	none	yes	yes	Wikidata:WikiProject Chemistry/Natural products	10.7554/eLife.52614	CC0

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 NP	Structured description of terms to 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 three-dimensional 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.