

Review

A Brief Review of Machine Learning-Based Bioactive Compound Research

Jihye Park ^{1,†}, Bo Ram Beck ^{2,†}, Hoo Hyun Kim ¹, Sangbum Lee ^{3,*} and Keunsoo Kang ^{1,*} 

¹ Department of Microbiology, College of Science & Technology, Dankook University, Cheonan 31116, Korea; apjh1998@dankook.ac.kr (J.P.); hoo Hyun.kim@dankook.ac.kr (H.H.K.)

² Deargen Inc., 193, Munji-ro, Yuseong-gu, Daejeon 34051, Korea; brbr777@deargen.me

³ Department of Software, College of Software Convergence, Dankook University, Yongin 16890, Korea

* Correspondence: sblee@dankook.ac.kr (S.L.); kangk1204@dankook.ac.kr (K.K.); Tel.: +82-31-8005-3224 (S.L.); +82-41-550-3456 (K.K.)

† These authors contributed equally to this work.

Abstract: Bioactive compounds are often used as initial substances for many therapeutic agents. In recent years, both theoretical and practical innovations in hardware-assisted and fast-evolving machine learning (ML) have made it possible to identify desired bioactive compounds in chemical spaces, such as those in natural products (NPs). This review introduces how machine learning approaches can be used for the identification and evaluation of bioactive compounds. It also provides an overview of recent research trends in machine learning-based prediction and the evaluation of bioactive compounds by listing real-world examples along with various input data. In addition, several ML-based approaches to identify specific bioactive compounds for cardiovascular and metabolic diseases are described. Overall, these approaches are important for the discovery of novel bioactive compounds and provide new insights into the machine learning basis for various traditional applications of bioactive compound-related research.

Keywords: bioactive compound; natural product; machine learning; bioinformatics; cheminformatics; chemical space; cardiovascular disease; metabolic disease



Citation: Park, J.; Beck, B.R.; Kim, H.H.; Lee, S.; Kang, K. A Brief Review of Machine Learning-Based Bioactive Compound Research. *Appl. Sci.* **2022**, *12*, 2906. <https://doi.org/10.3390/app12062906>

Academic Editor: Leonel Pereira

Received: 23 December 2021

Accepted: 9 March 2022

Published: 11 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Bioactive compounds are chemicals that exist in trace amounts in natural products (NPs) from plants and animals. They are used as components of health supplements or as early substances in drug development [1–3]. Bioactive compounds are considered to have suitable scaffolds that can interact with various proteins in living organisms through evolutionary pressure. Therefore, they were frequently used as initial substances for a significant number of therapeutics [1,4–6]. For example, plants contain rapidly evolving, multi-functional metabolic enzymes that appear to influence the production of structurally diverse chemicals such as secondary metabolites [7]. In addition, many approved drugs are structurally similar to bioactive compounds [8], providing a clue that finding bioactive compound analogs may be the right strategy for novel drug development. For example, a clinically approved anti-mitotic drug, eribulin (Halaven), is derived from the natural product halichondrin B [9]. In addition, a survey estimated that 28% of all currently approved drugs are either NPs or their derivatives [10]. In fact, a major source of oral drugs is derived from NPs [11]. Therefore, the discovery of novel bioactive compounds in NPs or their derivatives has routinely been a way to develop valuable therapeutic agents.

Chemical space is defined as a virtual space, comprising all theoretically possible small organic molecules in excess of 10^{60} [12–14]. Previously, navigating the entire chemical space was thought to be impractical due to mostly technical limitations. However, with the theoretical and practical innovation of machine learning (ML) fields such as deep learning (DL), which is undergoing rapid development in support of hardware such as graphics

processing units (GPUs) and tensor processing units (TPUs) in recent years, it has become possible to navigate a chemical subspace with desired properties. There are many successful case studies of the application of ML for various purposes. For example, recent studies identified many NP-like small molecules through the exploration of the chemical space of NPs [15–17]. In addition, ML models can accurately discriminate wild from farmed salmon based on gas chromatography with flame ionization detector (GC-FID) fatty acid profiles [18]. Therefore, with the help of state-of-the-art machine learning techniques, a variety of tasks such as discovering novel bioactive compounds with desired properties by training data measured through various large-scale experiments have become possible.

Recently, many ML algorithms are being applied to drug discovery, including hit identification, lead optimization, and clinical development stages [19–21]. Their goal is to accelerate the discovery of drugs more quickly and efficiently. For example, the traditional hit identification process requires labor-intensive, massive high throughput screening (HTS) that consumes many resources. To overcome this process, ML- or DL-based drug-target interaction (DTI) algorithms are used to identify small molecules that can bind to the desired target proteins using either sequence-based or 3D structure-based data [22–25]. To evaluate the properties of a given compound, cheminformatics approaches, such as the quantitative structure–activity relationship (QSAR), a computational modeling method for predicting the relationship between structural properties and biological activities of chemical compounds, are used widely, along with the prediction of absorption, distribution, metabolism, excretion, and toxicity (ADMET) [26–28]. Collectively, these computational predictions are effective at reducing the time and cost required for drug development, and many appropriate applications can also be applied to NP-based drug or active molecule discovery, which is expected to accelerate such processes.

In this review, we avoid discussing in-depth characteristics of machine learning techniques and the discovery and/or development of novel bioactive compounds, but rather offer an overview of recent research trends toward the machine learning-driven prediction and evaluation of bioactive compounds based on actual cases.

2. Cheminformatics, Bioinformatics, and Databases for Machine Learning

Cheminformatics is a research field that achieves desired goals by utilizing various information on chemicals, including 2D and 3D structures of chemical species. In general, the structural information of chemicals can be standardized and converted to machine-readable formats such as a simplified molecular-input line-entry system (SMILES) [29]. SMILES strings can be stored electronically into a database and then utilized through a variety of computational algorithms to identify, evaluate, and predict properties of chemicals. For example, the PubChem database [30] currently contains more than 100 million unique chemical structures extracted from contributed PubChem substance records (<https://pubchemdocs.ncbi.nlm.nih.gov/statistics> (accessed on 10 December 2021)). There are other databases for chemical collections such as ChEMBL (<https://www.ebi.ac.uk/chembl/> (accessed on 10 December 2021)) [31] and ZINC15 (<https://zinc15.docking.org/> (accessed on 10 December 2021)) [32]. Therefore, integrating cheminformatics methods and databases into existing drug discovery procedures has become indispensable for successful drug development. Accordingly, it is not at all surprising that many initial hits (a hit is defined as a molecule that shows the desired type of activity in an experimental screening assay) and further optimization for lead compounds can be performed via computational methods, such as computer-aided drug design (CADD) approaches. There are many successful applications of CADD in the development of currently available commercial drugs. For example, zanamivir (Relenza), a neuraminidase inhibitor that is used as a prophylaxis to treat the symptoms of influenza, was approved by the United States Food and Drug Administration (FDA) in 1999 [33] and is one of the most successful outcomes of CADD.

The identification of new bioactive compounds for a given target protein requires not only chemical information, but also detailed molecular information about the target protein, such as amino acid sequence, domain, and three-dimensional structural infor-

mation. Fortunately, this kind of information has been accumulated experimentally and computationally over several decades [34]. Conceptually, genome (DNA sequences), RNA sequences, protein sequences, and chemical structure information that is fundamentally similar between individuals within the same species can be defined as static information. On the other hand, molecular information that can be dynamically changed between individuals and/or conditions, such as DNA sequence variants (DNA mutations), gene expressions, and protein amounts, can be defined as dynamic information (Figure 1). Rapid innovation in high-throughput screening methods and large-scale expansion into virtually all molecular and clinical domains has dramatically expanded the scale of this molecular data over the past decade [35]. Therefore, bioinformatics has emerged to analyze this vast amount of molecular data.

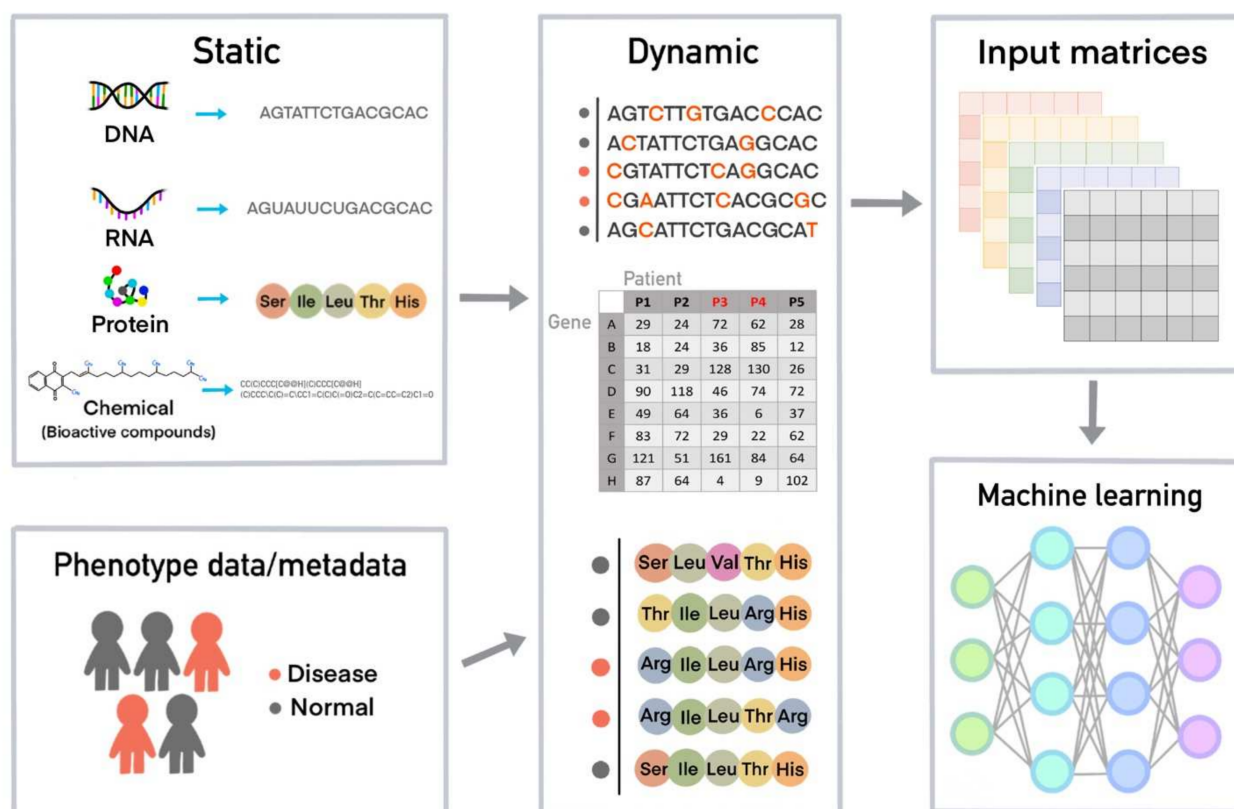


Figure 1. Various molecular and phenotypic data can be integrated to build a machine learning model. DNA, RNA, and protein, as well as chemical information, can be represented as strings (**top left**). This kind of fundamental molecular information can be defined as static information (deterministic), which shows little or no difference between individuals. On the other hand, phenotypic information (e.g., clinical data; **bottom left**) and high-throughput screening data (e.g., RNA-seq) produce dynamic information of molecules that can be integrated and represented as matrices (e.g., rows represent features and columns indicate samples, or vice versa) (**middle**). A collection of properly labelled matrices (**top right**) is used as input data to build a machine learning model for a particular purpose (**bottom right**).

The terms “cheminformatics” and “bioinformatics” are used interchangeably as the two fields deal primarily with computer-readable data from small chemical compounds, metabolites, RNAs, DNAs, and/or proteins. As mentioned above, cheminformatics mainly focuses on various chemical information, while bioinformatics primarily deals with DNA, RNA, and protein information. Many DNA and RNA nucleotides and/or protein databases are publicly available, including the nucleotide database maintained by the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/nucleotide/>) (accessed on 10 December 2021) [36], the UniProt database (<https://www.uniprot.org/>

(accessed on 10 December 2021)) [37], and the Protein Data Bank (PDB) database (<https://www.rcsb.org/> (accessed on 10 December 2021)) [38]. In addition, there are several web-based ready-to-use databases related to bioactive compounds, which have been well described in a review [39]. These databases mostly contain sequence (string) information of given molecules, obtained from various molecular experiments. For example, DNA and RNA can be represented as sequences of A, G, C, and T (or U for RNA), while protein can be denoted as strings of twenty letters where a single letter indicates an amino acid. In contrast to this static molecular information, there are big databases that contain massive amounts of dynamic molecular information generated by high-throughput screening approaches, such as next-generation sequencing (NGS). Compared to the static molecular information mentioned above, NGS-based high-throughput screening systems have been generating molecular dynamic data, such as RNA (transcription or gene expression), DNA (single nucleotide variant, SNV; insertion and deletion, INDEL; copy number variation, CNV; structural variation, SV), and DNA or RNA binding proteins in various cells, tissues, and even at the single cell level. Currently, more than 10 petabytes of raw sequencing data (<https://trace.ncbi.nlm.nih.gov/Traces/sra> (accessed on 10 December 2021)) can be downloaded from the Sequence Read Archive (SRA) website. The majority of these data is from gene expression profiling experiments, such as RNA-seq. For example, changes in the expression level of approximately 60,000 genes (or 230,000 transcripts) in human cells, with drug treatment, can be measured with an RNA-seq approach. With 100 samples, the total number of values would be 6,000,000 for genes (or 23,000,000 for transcripts), and this information can be represented as a matrix (Figure 1). However, since most genes are not normally expressed, the number of values will typically be reduced to less than one-third of that number. Then, the mode of action (MOA) prediction of the given drug can be achieved by analyzing this dynamic information using various bioinformatic algorithms and/or databases. In general, a typical MOA prediction can be conducted using gene ontology (GO) analysis and/or gene set enrichment analysis (GSEA). One of the most commonly used tools for interpreting transcriptome changes is the GSEA [40], which lists biological pathways that have been statistically significantly altered between different conditions. This knowledge-based information can be useful for narrowing down features (genes) for training because, in general, the number of features is much larger than the number of samples when building an ML model with gene expression profiling data. Another obstacle to applying machine learning technology to gene expression data is that individual gene expression data have been processed in different analysis pipelines, so unknown bias (noise) that is difficult to identify, such as batch effects, different sequencing devices, and/or various experimental conditions, is inherent to the data. To minimize this bias and provide vast transcriptome resources in a user-friendly manner, a recent study constructed a transcriptome database called ARCHS4 (<https://maayanlab.cloud/archs4/> (accessed on 10 December 2021)), comprising more than 200,000 human and mouse transcripts processed through a uniform analysis pipeline [41]. In the case of DNA variants in humans, gnomAD (<https://gnomad.broadinstitute.org/> (accessed on 10 December 2021)) provides an excellent and the largest source of DNA variant information identified in multiple populations via a wide variety of large-scale sequencing projects [42].

Overall, advances in cheminformatics, bioinformatics, and a variety of publicly accessible databases have emerged rapidly over the past decade, accelerating the process of interconnecting chemistry, biology, and drug development. This system will be used to expand our limited understanding of chemicals and ultimately build promising ML-based models that generate novel bioactive-like compounds, with high efficacy and low toxicity, against various diseases, including cardiovascular and metabolic diseases.

3. Chemical Space Where Unidentified Bioactive Compounds Exist

Chemical space, defined as the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created [43], is a multi-dimensional space in which the area to explore varies depending on what characteristics of chemicals

(e.g., molecular mass, lipophilicity, toxicity, drug-likeness, number of hydrogen donors and acceptors, fraction of rotatable bonds, and druggability) the coordinates are set. In order to reach out after a subspace (e.g., NP-like chemicals) in the chemical space, it is necessary to be able to calculate unique properties of a group of similar chemicals numerically. For example, Chen et al. successfully defined a chemical space where NP-like chemicals are located and found that this space also contains chemicals highly relevant for drug discovery [44]. Fortunately, most properties of chemicals can numerically be calculated through various algorithms with metrics. For instance, the quantitative estimate of drug-likeness (QED) metric [45] estimates the drug-likeness of chemicals, and the absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of given chemicals can be calculated through various algorithms developed by computational and medicinal chemistry fields [46–53]. These numeric values, if appropriately used, will be a guide to discover novel hits for a given target protein through exploring a desired NP-like chemical space.

4. How a Machine Learns from Data and Creates a Model for a Task Using Machine Learning Algorithms

Machine learning (ML) is a research area encompassing computational methods of learning relationships from data without specifying them. Deep learning (DL) is a subset of ML [54] and is used in almost all areas because of the recent rapid development of hardware. In general, these approaches utilize a large number of quality-assured samples with many features. There are two types of ML methods, depending on dependent variables (or labels). Supervised learning methods learn the relationships between input features and the given labels of samples from a training dataset, whereas unsupervised learning algorithms infer patterns of input features without labels. We focus only on the supervised learning method in this review.

To build a model that can accurately classify samples into classes (classification) or predict values for samples (regression), a dataset, which contains a large number of correctly labelled-samples with many features, should be divided into three sub-datasets, called training, validation, and test datasets. The training dataset is used for training to create an optimal model among the various models to be tested. Therefore, only the training dataset is used to train the models. In contrast, the validation set is a dataset used for testing multiple models with different hyper-parameters to find the optimal values for the best model. Through extensive training steps using training and validation datasets, the best model, which has been trained and validated, is selected. Then, the selected model's classification (or regression) performance is evaluated using the test dataset. Therefore, the proper way to introduce a ML-based model is to explicitly describe all processes and parameters as much as possible [55]. Because the performance of a ML-based model is solely based on training, validation, and test datasets as well as ML algorithms with parameters (e.g., random initializations, data shuffling, dropout, etc.), it should be mandated to report all the information for the sake of transparency and reproducibility [56]. Details of machine learning have been well covered in recent reviews [57–59].

Numerous ML-based applications have been developed for a variety of purposes, including drug discovery. For example, Shin et al. introduced a transformer-based deep learning model named MT-DTI that can accurately predict binding affinities between drugs and a target protein [60] (Figure 2). Briefly, this model utilized PubChem's approximately 97 million compounds to train key knowledge such as the syntax and semantics of small molecules using their SMILES information. The trained molecular transformer (MT) information was transferred into a drug target interaction (DTI) model (so-called transfer learning), and the final DTI model, called MT-DTI, was constructed using two DTI datasets, KIBA [61] and Davis [62]. These datasets contain experimentally validated interactions between kinases and kinase inhibitors. The MT-DTI model takes two input sequences, the SMILESs of chemicals and the amino acid sequences of a target protein, to predict their

binding affinities, and it can therefore be used to identify potential target proteins of known bioactive compounds or vice versa.

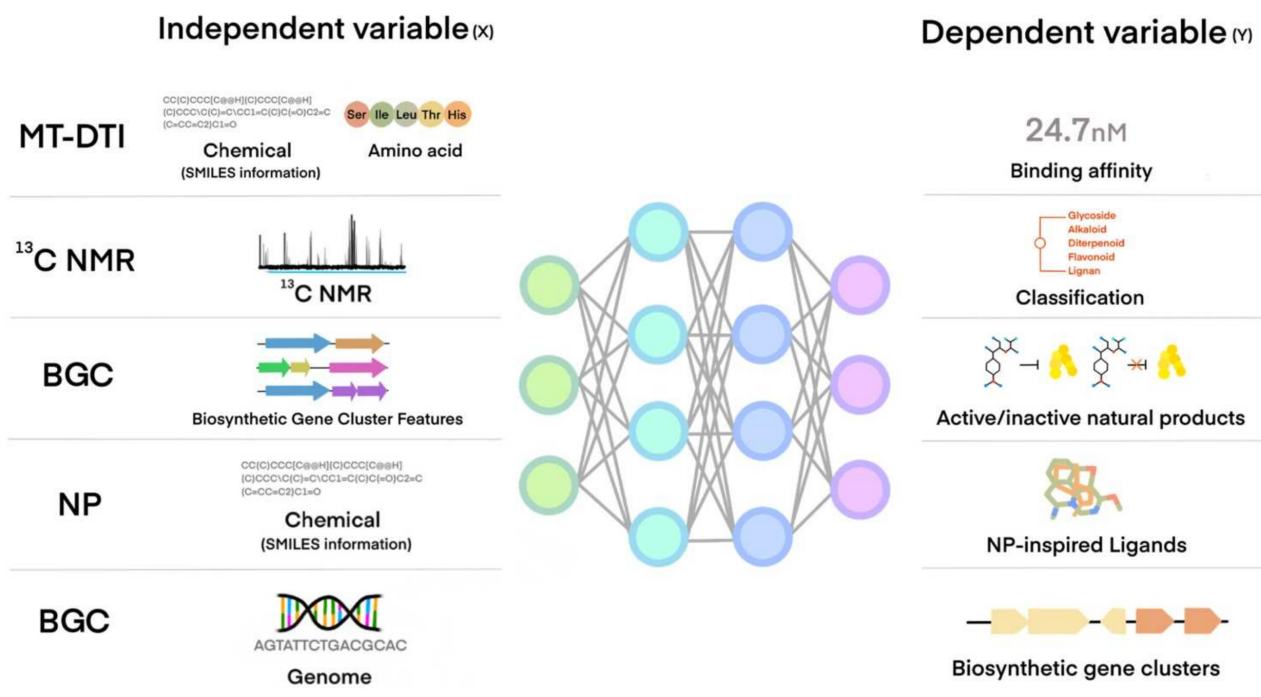


Figure 2. Machine learning-based models have been developed for different purposes. Various input data were used as independent variables (X, **left** panel) to predict dependent variables (Y, **right** panel). Five different examples are shown. The first example shows a model (called MT-DTI) that can predict the binding affinity (Y, the equilibrium dissociation constant known as K_D) between a given protein and ligand (small chemical compound). Strings of protein (amino acids) and ligand (SMILES) information (X) are used as input. The second model takes NMR data (X) as input and classifies them into known NP classes (Y). The third model analyzes genomic sequences called biosynthetic gene clusters (X) and predicts the likelihood that a natural product will have activity (Y) from those gene clusters. The fourth model takes a SMILE of a chemical of interest (X, e.g., natural product) and identifies similar small chemicals (Y) with desired multi-target profiles on disease-related targets. The last model detects biosynthetic gene clusters (Y) using a deep learning model called DeepBGC. Bacterial genome sequences are used as input (X).

Nuclear magnetic resonance (NMR) spectroscopic data can also be used to build ML-based classification models. Martínez-Treviño et al. [63] built a model that can predict NP classes of chemicals using an ML-based approach with one-dimensional ^{13}C NMR spectroscopic data (Figure 2). In this study, the proposed approach does not rely on a typical peak matching task that requires available well-annotated NMR spectroscopic databases of chemicals, but they constructed a ML model that can predict the presence or absence of a particular NP family from the ^{13}C NMR spectra. To this end, they trained the model with continuous variables of NP spectra in the NAPROC-13 database [64], which contains ^{13}C spectral information of more than 18,000 unique NP structures and found that a XGBoost-based classifier accurately predicted a given NP into eight NP families (sesquiterpenoid, diterpenoid, triterpenoid, lignan, steroid, chroman, flavonoid, and alkaloid).

String information, such as DNA nucleotides and/or protein amino acid sequences, can be used to build a ML model for a particular purpose. Walker et al. [65] developed a machine learning bioinformatics method for predicting a natural product's antibiotic activity directly from the sequence of its biosynthetic gene cluster. In this study, they assembled a training dataset from the MiBIG (Minimum Information about Biosynthetic

Gene Cluster) database, comprising sequences of biosynthetic gene clusters (BGCs). They assembled a dataset of known BGCs paired with the activity of their products by representing the BGCs as vectors based on the number of times various gene annotations appeared in the cluster. Then, binary classification models for antibacterial, anti-Gram-positive, anti-Gram-negative, antifungal/antitumor/cytotoxic, antitumor/cytotoxic, and antifungal activities were constructed using this information as the training data set. The proposed binary classifiers can achieve up to 80% accuracy. Another study also built a ML model with the BGC information. Hannigan et al. [66] present a deep learning strategy (DeepBGC) that offers reduced false positive rates in BGC identification and an improved ability to extrapolate and identify novel BGC classes, compared to existing machine-learning tools. DeepBGC is a deep learning approach that uses a Bidirectional Long Short-Term Memory (BiLSTM) RNN [67] and a word-embedding skip-gram neural network (pfam2vec) such as word2vec [68]. The input consists of amino acid sequences of protein family (Pfam) domains [69] represented by vectors, and the output is a sequence of values between 0 and 1, representing the prediction score of whether a given domain is part of a BGC. This implementation improved the accuracy of BGC detection in genome sequences and the ability to identify BGCs capable of encoding natural products with novel biological activities.

The identification of novel NP-like chemicals with desired properties is of great interest in drug discovery. Grisoni et al. [70] present a machine-learning workflow for the identification of NP mimetics with the desired polypharmacology. More than 3.3 million commercially available compounds were computationally screened to identify chemicals similar to (–)-galantamine, a natural product-based drug approved by the US Food and Drug Administration (FDA) for the treatment of cognitive decline in mild to moderate Alzheimer’s disease (AD) [71]. To this end, weighted holistic atom localization and entity shape (WHALES) descriptors [72], which represent the partial charge distribution and three-dimensional shape of compounds, were used to identify initial compounds similar to (–)-galantamine. Then, the compounds most similar to (–)-galantamine were further evaluated using the SOM-based prediction of drug equivalence relationships (SPiDER) [73] and target inference generator (TIGER) [74] applications, which predict target proteins of a given compound. Because (–)-galantamine reversibly inhibits acetylcholinesterase (AChE), the purpose of this process was to identify AChE inhibitors. Finally, this computational pipeline identified eight compounds with bioactivity on at least one of the macromolecular targets of (–)-galantamine, showing different polypharmacological properties.

Inferences about the origin of NPs provide insight into our understanding of the physicochemical properties of NPs. Pereira [75] developed a machine learning-based classification model to classify NPs into three classes: marine natural products (MNPs), terrestrial natural products (TNPs), and NPs that appear in both the terrestrial and marine environments. Briefly, molecular structures were retrieved from the Reaxys[®] database (Elsevier Information Systems GmbH; <https://www.reaxys.com> (accessed on 10 December 2021)) and were filtered appropriately. A total of 22,398 NPs were defined, which comprised 10,790 MNPs, 10,857 TNPs, and 761 as both. The data set was randomly divided into a training set of 15,676 NPs and a test set of 6722 NPs. Then, classification models were constructed using random forest, support vector machine, and multilayer perceptron networks with an overall predictive accuracy of 81% for the test set. Although all these examples only represent limited use cases of machine learning algorithms, these examples show how machine learning techniques can be utilized for NP-related research.

5. Machine Learning Application of NP or NP-Like Chemical Compounds Discovery for Cardiovascular and Metabolic Diseases

Machine learning approaches may identify active molecules derived from NPs to improve human health in numerous areas of interest through methodologies described in previous sections. From a metabolic disease perspective, there are many attempts to discover NP mimetics that target specific proteins. For example, PPAR γ is a master regulator protein of metabolism which is identified to be involved in metabolic diseases such as

chronic inflammation, obesity, and diabetes when its function is lost or dysregulated [76,77]. Rupp et al. identified several active compound structures toward PPAR γ that are related to cinnamic acid and truxillic acid through non-linear Bayesian regression with Gaussian processes and virtual screening [78]. Another study suggests caffeine as an anti-diabetic NP through an iterative stochastic elimination (ISE) model built on in-house data and confirmation with wet experiments [79]. Recently, a deep learning based model using latent knowledge, molecule to target interactions, and compound property as heterogeneous input features predicted several NPs or NP-derivatives for various cardiovascular diseases (CVD): ergosterol and arginine for heart failure; reserpine, norepinephrine, octopamine, and digitoxin for hypertension; resveratrol for myocardial infarction; and aspirin and agmatine for stroke [80].

In an extensive perspective of metabolic syndrome and CVD, inflammation is a biological process that is critical in the first line of defense of the host immune system, and dysregulatory inflammatory responses can lead to cardiovascular and metabolic diseases [81,82]. A recent study [83] identified capsaicin, hypaphorine, and moupinamide among NPs as anti-inflammatory drug candidates using the ISE model, and the results were further supported by the literature. Another study employed linear discriminant analysis to identify two NP derivatives, alizarin-3-methylimino-*N,N*-diacetic acid and (+)-dibenzyl-L-tartrate, as anti-inflammatory agents [84]. Chagas-Paula et al. demonstrated the application of a decision tree model and a back propagation multilayer perceptron model to identify new natural products with cyclooxygenase-1 and 5-lipoxygenase dual inhibition activity from Asteraceae species extracts [85].

Although case studies for specific usage, covering predictions for wet experiments of ML methods to predict bioactive NPs, are scarce, the studies discussed above demonstrate the potential power and opportunities of ML methods in this field, supported by either literature or experimental evidence.

6. Conclusions

Although ML-based approaches typically achieve high performance for classification or regression tasks, there are still some limitations to be widely applied in NP-related research. First, the number of natural product-related data in which bioactivity or a target have been experimentally verified is insufficient to build an ML model. Second, while ML-based models generally outperform conventional approaches for many tasks, the reason why the model makes such a decision cannot be interpreted directly (so called “black box”). Third, an ML model is a mathematical function whose tunable parameters (values may change during training) and hyperparameters (not updated during training) must be set appropriately, but these values are iteratively tested during training to achieve the best performance, which is a time-consuming operation. Nevertheless, recent advances in machine learning have made a breakthrough in resolving these issues. For example, scientific interest in the field of explainable artificial intelligence (XAI) is emerging to explain and interpret ML models [86], while automated machine learning (AutoML) seeks to automate the entire pipeline of ML model building: automatically select, compose, and parametrize ML models [87]. Therefore, machine learning-based bioactive compound research will benefit from these state-of-the-art ML technologies. To this end, it is most important for researchers who conduct bioactive compound-related research to track newly developed ML applications and collaborate with experts in the field.

7. Future Perspective

As discussed in previous sections, ML technology can establish strategically faster studies and new opportunities to develop bioactive NPs. Despite the usefulness of ML methods, there are a few more topics that need to be considered for practical application.

NPs are mostly generally recognized as safe (GRAS, <https://www.fda.gov/food/food-ingredients-packaging/generally-recognized-safe-gras> (accessed on 10 December 2021)) and food-grade substances and used as ingredients for a variety of health supple-

ments, functional food products, and cosmetic products. GRAS-listed NPs (i.e., berberine, canthaxanthin, eugenol, L-Glutamine, or limonene) can be directly applied to such products without any complex legal regulation, and this accessibility is a strength of NPs. In contrast, a downside of NPs is that their therapeutic efficacy is often low and limited compared to synthetic compounds. For example, phytoalexin coumarin is a plant-derived NP to defend plants from microbial infection and is studied as a bioactive NP for numerous disease indications [88]. However, coumarin itself needs to be better improved, or so-called optimized, as a derivative to achieve better effectiveness against bacterial infection, hypertension, or inflammation [88,89]. Therefore, the weight between using NPs of nature by themselves or further developing NP derivatives must be considered prior to using ML models.

To process the lead optimization task more efficiently, many machine learning approaches have been studied recently: (i) atom modification reinforcement learning models that add or delete atoms or bonds [90,91], (ii) generative reinforcement learning which generates similar but modified structure [92], (iii) generative machine learning with controlled chemical properties that also generates similar modified structure with preserved predictive properties [93], and (iv) a 3D structure-based ligand design model that uses a 3D crystal structure of protein and ligand to generate novel molecules [94]. These applications demonstrated the possibility of lead optimization or de novo small molecule-focused machine learning methods. Their goals are to apply listed models to drug discovery, but the basis of these models is not limited to de novo synthesis or the lead optimization of new chemical entities. By adopting such methods, NPs can serve as “seed” or “starting” molecules and NP derivatives as resulting modified products with preserved chemical structure similarity.

Author Contributions: Conceptualization, K.K. and S.L.; investigation, J.P., B.R.B. and H.H.K.; writing—original draft preparation, K.K., J.P., B.R.B. and S.L.; writing—review and editing, K.K. and S.L.; visualization, J.P. and H.H.K.; supervision, K.K.; funding acquisition, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by a research grant from the Korea Food Research Institute (Project Number: E0210601-01). This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (NRF-2019R1C1C1010385) and by a grant from the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (1720100). This research was also supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1A6A1A03043283). The Department of Microbiology was supported through the Research-Focused Department Promotion Project as a part of the University Innovation Support Program for Dankook University in 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: B.R.B. is employed by Deargen Inc., and K.K. is a shareholder of Deargen Inc.

References

1. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661. [[CrossRef](#)]
2. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **2012**, *75*, 311–335. [[CrossRef](#)]
3. Thomford, N.E.; Senthane, D.A.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. *Int. J. Mol. Sci.* **2018**, *19*, 1578. [[CrossRef](#)]
4. Atanasov, A.G.; Waltenberger, B.; Pferschy-Wenzig, E.-M.; Linder, T.; Wawrosch, C.; Uhrin, P.; Temml, V.; Wang, L.; Schwaiger, S.; Heiss, E.H.; et al. Discovery and resupply of pharmacologically active plant-derived natural products: A review. *Biotechnol. Adv.* **2015**, *33*, 1582–1614. [[CrossRef](#)]
5. Harvey, A.L.; Edrada-Ebel, R.; Quinn, R.J. The re-emergence of natural products for drug discovery in the genomics era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129. [[CrossRef](#)] [[PubMed](#)]

6. Waltenberger, B.; Mocan, A.; Šmejkal, K.; Heiss, E.H.; Atanasov, A.G. Natural products to counteract the epidemic of cardiovascular and metabolic disorders. *Molecules* **2016**, *21*, 807. [[CrossRef](#)] [[PubMed](#)]
7. Fraenkel, G.S. The raison d'être of secondary plant substances; These odd chemicals arose as a means of protecting plants from insects and now guide insects to food. *Science* **1959**, *129*, 1466–1470. [[CrossRef](#)] [[PubMed](#)]
8. Hu, Y.; Lounkine, E.; Bajorath, J. Many approved drugs have bioactive analogs with different target annotations. *AAPS J.* **2014**, *16*, 847–859. [[CrossRef](#)]
9. Yu, M.J.; Zheng, W.; Seletsky, B.M. From micrograms to grams: Scale-up synthesis of eribulin mesylate. *Nat. Prod. Rep.* **2013**, *30*, 1158–1164. [[CrossRef](#)]
10. Eder, J.; Sedrani, R.; Wiesmann, C. The discovery of first-in-class drugs: Origins and evolution. *Nat. Rev. Drug Discov.* **2014**, *13*, 577–587. [[CrossRef](#)]
11. Doak, B.C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral druggable space beyond the rule of 5: Insights from drugs and clinical candidates. *Chem. Biol.* **2014**, *21*, 1115–1142. [[CrossRef](#)] [[PubMed](#)]
12. Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861. [[CrossRef](#)] [[PubMed](#)]
13. 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.* **2001**, *46*, 3–26. [[CrossRef](#)]
14. Bohacek, R.S.; McMartin, C.; Guida, W.C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50. [[CrossRef](#)]
15. Zabolotna, Y.; Ertl, P.; Horvath, D.; Bonachera, F.; Marcou, G.; Varnek, A. NP navigator: A new look at the natural product chemical space. *Mol. Inform.* **2021**, *40*, e2100068. [[CrossRef](#)]
16. Nelson, A.; Karageorgis, G. Natural product-informed exploration of chemical space to enable bioactive molecular discovery. *RSC Med. Chem.* **2021**, *12*, 353–362. [[CrossRef](#)]
17. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* **2007**, *70*, 789–794. [[CrossRef](#)]
18. Grazina, L.; Rodrigues, P.J.; Igrejas, G.; Nunes, M.A.; Mafra, I.; Arlorio, M.; Oliveira, M.B.P.P.; Amaral, J.S. Machine learning approaches applied to GC-FID fatty acid profiles to discriminate wild from farmed salmon. *Foods* **2020**, *9*, 1622. [[CrossRef](#)]
19. Paul, D.; Sanap, G.; Shenoy, S.; Kalyane, D.; Kalia, K.; Tekade, R.K. Artificial intelligence in drug discovery and development. *Drug Discov. Today* **2021**, *26*, 80–93. [[CrossRef](#)]
20. Mak, K.-K.; Balijepalli, M.K.; Pichika, M.R. Success stories of AI in drug discovery—Where do things stand? *Expert Opin. Drug Discov.* **2022**, *17*, 79–92. [[CrossRef](#)]
21. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial intelligence in drug discovery: Recent advances and future perspectives. *Expert Opin. Drug Discov.* **2021**, *16*, 949–959. [[CrossRef](#)] [[PubMed](#)]
22. Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **2018**, *34*, i821–i829. [[CrossRef](#)] [[PubMed](#)]
23. Moon, S.; Zhung, W.; Yang, S.; Lim, J.; Kim, W.Y. PIGNet: A physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem. Sci.* **2022**, *in press*. [[CrossRef](#)]
24. Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M.E.; Cherkasov, A. Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Cent. Sci.* **2020**, *6*, 939–949. [[CrossRef](#)]
25. Francoeur, P.G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R.B.; Snyder, I.; Koes, D.R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *J. Chem. Inf. Model.* **2020**, *60*, 4200–4215. [[CrossRef](#)]
26. Venkatraman, V. FP-ADMET: A compendium of fingerprint-based ADMET prediction models. *J. Cheminform.* **2021**, *13*, 75. [[CrossRef](#)]
27. Ferreira, L.L.G.; Andricopulo, A.D. ADMET modeling approaches in drug discovery. *Drug Discov. Today* **2019**, *24*, 1157–1165. [[CrossRef](#)]
28. Göller, A.H.; Kuhnke, L.; Montanari, F.; Bonin, A.; Schneckener, S.; ter Laak, A.; Wichard, J.; Lobell, M.; Hillisch, A. Bayer's in silico ADMET platform: A journey of machine learning over the past two decades. *Drug Discov. Today* **2020**, *25*, 1702–1709. [[CrossRef](#)]
29. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [[CrossRef](#)]
30. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [[CrossRef](#)]
31. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [[CrossRef](#)] [[PubMed](#)]
32. Sterling, T.; Irwin, J.J. ZINC 15-ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [[CrossRef](#)] [[PubMed](#)]
33. Woods, J.M.; Bethell, R.C.; Coates, J.A.; Healy, N.; Hiscox, S.A.; Pearson, B.A.; Ryan, D.M.; Ticehurst, J.; Tilling, J.; Walcott, S.M. 4-guanidino-2,4-dideoxy-2,3-dehydro-N-acetylneuraminic acid is a highly effective inhibitor both of the sialidase (neuraminidase) and of growth of a wide range of influenza A and B viruses in vitro. *Antimicrob. Agents Chemother.* **1993**, *37*, 1473–1479. [[CrossRef](#)] [[PubMed](#)]
34. Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Biol.* **2003**, *10*, 980. [[CrossRef](#)]

35. Sheynkman, G.M.; Shortreed, M.R.; Cesnik, A.J.; Smith, L.M. Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu. Rev. Anal. Chem.* **2016**, *9*, 521–545. [[CrossRef](#)]
36. NCBI. Resource coordinators database resources of the national center for biotechnology information. *Nucleic Acids Res.* **2018**, *46*, D8–D13. [[CrossRef](#)]
37. Apweiler, R.; Bairoch, A.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699. [[CrossRef](#)]
38. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)]
39. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20. [[CrossRef](#)]
40. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [[CrossRef](#)]
41. Lachmann, A.; Torre, D.; Keenan, A.B.; Jagodnik, K.M.; Lee, H.J.; Wang, L.; Silverstein, M.C.; Ma'ayan, A. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **2018**, *9*, 1366. [[CrossRef](#)] [[PubMed](#)]
42. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alfoldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **2020**, *581*, 434–443. [[CrossRef](#)] [[PubMed](#)]
43. Dobson, C.M. Chemical space and biology. *Nature* **2004**, *432*, 824–828. [[CrossRef](#)]
44. Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the chemical space of known and readily obtainable natural products. *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532. [[CrossRef](#)] [[PubMed](#)]
45. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [[CrossRef](#)]
46. Durán-Iturbide, N.A.; Díaz-Eufracio, B.I.; Medina-Franco, J.L. In silico ADME/tox profiling of natural products: A focus on BIOFACQUIM. *ACS Omega* **2020**, *5*, 16076–16084. [[CrossRef](#)]
47. Dong, J.; Wang, N.-N.; Yao, Z.-J.; Zhang, L.; Cheng, Y.; Ouyang, D.; Lu, A.-P.; Cao, D.-S. ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. *J. Cheminform.* **2018**, *10*, 29. [[CrossRef](#)]
48. Bocci, G.; Carosati, E.; Vayer, P.; Arrault, A.; Lozano, S.; Cruciani, G. ADME-space: A new tool for medicinal chemists to explore ADME properties. *Sci. Rep.* **2017**, *7*, 6359. [[CrossRef](#)]
49. Banerjee, P.; Eckert, A.O.; Schrey, A.K.; Preissner, R. ProTox-II: A webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res.* **2018**, *46*, W257–W263. [[CrossRef](#)]
50. Van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204. [[CrossRef](#)]
51. Feinberg, E.N.; Joshi, E.; Pande, V.S.; Cheng, A.C. Improvement in ADMET prediction with multitask deep featurization. *J. Med. Chem.* **2020**, *63*, 8835–8848. [[CrossRef](#)] [[PubMed](#)]
52. Daina, A.; Michielin, O.; Zoete, V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, 42717. [[CrossRef](#)] [[PubMed](#)]
53. Cáceres, E.L.; Tudor, M.; Cheng, A.C. Deep learning approaches in predicting ADMET properties. *Future Med. Chem.* **2020**, *12*, 1995–1999. [[CrossRef](#)] [[PubMed](#)]
54. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2021**, *23*, 40–55. [[CrossRef](#)]
55. Artrith, N.; Butler, K.T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nat. Chem.* **2021**, *13*, 505–508. [[CrossRef](#)]
56. Moving towards reproducible machine learning. *Nat. Comput. Sci.* **2021**, *1*, 629–630. [[CrossRef](#)]
57. Patel, L.; Shukla, T.; Huang, X.; Ussery, D.W.; Wang, S. Machine learning methods in drug discovery. *Molecules* **2020**, *25*, 5277. [[CrossRef](#)]
58. Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N.C.; Ping, P. Machine learning and integrative analysis of biomedical big data. *Genes* **2019**, *10*, 87. [[CrossRef](#)]
59. Eraslan, G.; Avsec, Ž.; Gagneur, J.; Theis, F.J. Deep learning: New computational modelling techniques for genomics. *Nat. Rev. Genet.* **2019**, *20*, 389–403. [[CrossRef](#)]
60. Shin, B.; Park, S.; Kang, K.; Ho, J.C. Self-attention based molecule representation for predicting drug-target interaction. *arXiv* **2019**, arXiv:1908.06760.
61. Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; Aittokallio, T. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *J. Chem. Inf. Model.* **2014**, *54*, 735–743. [[CrossRef](#)] [[PubMed](#)]
62. Davis, M.I.; Hunt, J.P.; Herrgard, S.; Ciceri, P.; Wodicka, L.M.; Pallares, G.; Hocker, M.; Treiber, D.K.; Zarrinkar, P.P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011**, *29*, 1046–1051. [[CrossRef](#)] [[PubMed](#)]
63. Martínez-Treviño, S.H.; Uc-Cetina, V.; Fernández-Herrera, M.A.; Merino, G. Prediction of natural product classes using machine learning and ¹³C NMR spectroscopic data. *J. Chem. Inf. Model.* **2020**, *60*, 3376–3386. [[CrossRef](#)] [[PubMed](#)]

64. López-Pérez, J.L.; Therón, R.; del Olmo, E.; Díaz, D. NAPROC-13: A database for the dereplication of natural product mixtures in bioassay-guided protocols. *Bioinformatics* **2007**, *23*, 3256–3257. [[CrossRef](#)] [[PubMed](#)]
65. Walker, A.S.; Clardy, J. A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. *J. Chem. Inf. Model.* **2021**, *61*, 2560–2571. [[CrossRef](#)]
66. Hannigan, G.D.; Prihoda, D.; Palicka, A.; Soukup, J.; Klempir, O.; Rampula, L.; Durcak, J.; Wurst, M.; Kotowski, J.; Chang, D.; et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **2019**, *47*, e110. [[CrossRef](#)]
67. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
68. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 9.
69. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)]
70. Grisoni, F.; Merk, D.; Friedrich, L.; Schneider, G. Design of natural-product-inspired multitarget ligands by machine learning. *ChemMedChem* **2019**, *14*, 1129–1134. [[CrossRef](#)]
71. Sharma, K. Cholinesterase inhibitors as Alzheimer’s therapeutics (review). *Mol. Med. Rep.* **2019**, *20*, 1479–1487. [[CrossRef](#)] [[PubMed](#)]
72. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J.A.; Tagliabue, S.G.; Todeschini, R.; Schneider, G. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **2018**, *1*, 44. [[CrossRef](#)]
73. Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4067–4072. [[CrossRef](#)] [[PubMed](#)]
74. Schneider, P.; Schneider, G. A computational method for unveiling the target promiscuity of pharmacologically active compounds. *Angew. Chem. Int. Ed.* **2017**, *56*, 11520–11524. [[CrossRef](#)] [[PubMed](#)]
75. Pereira, F. Machine learning methods to predict the terrestrial and marine origin of natural products. *Mol. Inform.* **2021**, *40*, e2060034. [[CrossRef](#)]
76. Ahmadian, M.; Suh, J.M.; Hah, N.; Liddle, C.; Atkins, A.R.; Downes, M.; Evans, R.M. PPAR γ signaling and metabolism: The good, the bad and the future. *Nat. Med.* **2013**, *19*, 557–566. [[CrossRef](#)]
77. Lehrke, M.; Lazar, M.A. The many faces of PPAR γ . *Cell* **2005**, *123*, 993–999. [[CrossRef](#)] [[PubMed](#)]
78. Rupp, M.; Schroeter, T.; Steri, R.; Zettl, H.; Proschak, E.; Hansen, K.; Rau, O.; Schwarz, O.; Müller-Kuhrt, L.; Schubert-Zsilavecz, M.; et al. From machine learning to natural product derivatives that selectively activate transcription factor PPAR γ . *ChemMedChem* **2010**, *5*, 191–194. [[CrossRef](#)] [[PubMed](#)]
79. Zeidan, M.; Rayan, M.; Zeidan, N.; Falah, M.; Rayan, A. Indexing natural products for their potential anti-diabetic activity: Filtering and mapping discriminative physicochemical properties. *Molecules* **2017**, *22*, 1563. [[CrossRef](#)]
80. Yoo, S.; Yang, H.C.; Lee, S.; Shin, J.; Min, S.; Lee, E.; Song, M.; Lee, D. A deep learning-based approach for identifying the medicinal uses of plant-derived natural compounds. *Front. Pharmacol.* **2020**, *11*, 584875. [[CrossRef](#)]
81. Lolmède, K.; Duffaut, C.; Zakaroff-Girard, A.; Bouloumié, A. Immune cells in adipose tissue: Key players in metabolic disorders. *Diabetes Metab.* **2011**, *37*, 283–290. [[CrossRef](#)]
82. Hotamisligil, G.S. Inflammation and metabolic disorders. *Nature* **2006**, *444*, 860–867. [[CrossRef](#)]
83. Aswad, M.; Rayan, M.; Abu-Lafi, S.; Falah, M.; Raiyn, J.; Abdallah, Z.; Rayan, A. Nature is the best source of anti-inflammatory drugs: Indexing natural products for their anti-inflammatory bioactivity. *Inflamm. Res.* **2018**, *67*, 67–75. [[CrossRef](#)] [[PubMed](#)]
84. Galvez-Llompert, M.; del Carmen Recio Iglesias, M.; Gálvez, J.; García-Domenech, R. Novel potential agents for ulcerative colitis by molecular topology: Suppression of IL-6 production in Caco-2 and RAW 264.7 cell lines. *Mol. Divers.* **2013**, *17*, 573–593. [[CrossRef](#)]
85. Chagas-Paula, D.; Oliveira, T.; Zhang, T.; Edrada-Ebel, R.; Da Costa, F. Prediction of anti-inflammatory plants and discovery of their biomarkers by machine learning algorithms and metabolomic studies. *Planta Med.* **2015**, *81*, 450–458. [[CrossRef](#)] [[PubMed](#)]
86. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18. [[CrossRef](#)] [[PubMed](#)]
87. He, X.; Zhao, K.; Chu, X. AutoML: A survey of the state-of-the-art. *Knowl.-Based Syst.* **2021**, *212*, 106622. [[CrossRef](#)]
88. Detsi, A.; Kontogiorgis, C.; Hadjipavlou-Litina, D. Coumarin derivatives: An updated patent review (2015–2016). *Expert Opin. Ther. Pat.* **2017**, *27*, 1201–1226. [[CrossRef](#)]
89. Hu, Y.; Shen, Y.; Wu, X.; Tu, X.; Wang, G.-X. Synthesis and biological evaluation of coumarin derivatives containing imidazole skeleton as potential antibacterial agents. *Eur. J. Med. Chem.* **2018**, *143*, 958–969. [[CrossRef](#)]
90. Park, S.; Ko, Y.H.; Lee, B.; Shin, B.; Beck, B.R. Abstract 35: Molecular optimization of phase III trial failed anticancer drugs using target affinity and toxicity-centered multiple properties reinforcement learning. In Proceedings of the Poster Presentations—Proffered Abstracts, Ljubljana, Slovenia, 15 June 2020; p. 35.
91. Zhou, Z.; Kearnes, S.; Li, L.; Zare, R.N.; Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **2019**, *9*, 10752. [[CrossRef](#)]
92. Zhavoronkov, A.; Ivanenkov, Y.A.; Aliper, A.; Veselov, M.S.; Aladinskiy, V.A.; Aladinskaya, A.V.; Terentiev, V.A.; Polykovskiy, D.A.; Kuznetsov, M.D.; Asadulaev, A.; et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040. [[CrossRef](#)] [[PubMed](#)]

-
93. Shin, B.; Park, S.; Bak, J.; Ho, J.C. Controlled molecule generator for optimizing multiple chemical properties. In Proceedings of the Conference on Health, Inference, and Learning, Virtual Event, 8 April 2021; pp. 146–153.
 94. Li, Y.; Pei, J.; Lai, L. Structure-based *de novo* drug design using 3D deep generative models. *Chem. Sci.* **2021**, *12*, 13664–13675. [[CrossRef](#)] [[PubMed](#)]