Advanced Pharmaceutical and Biomedical Letters
Available online at https://advpbl.org.ng
Adv Pharm Biomed Lett, 2026:3, 78-91

Open Access Full Text Article

Review Article

Dusts and De-dusting of Natural Products Datasets: A Critical Review of the Challenges

Shadrach C. Eze[1,2], Lilian I. Nweke[3,4], Frances Ojobor[4], Wande M. Oluyemi[1,2], Charles O. Nnadi[4,5*]

[1]Department of Pharmaceutical and Medicinal Chemistry, College of Pharmacy, Afe Babalola University, Ado-Ekiti, Nigeria; [2]Research Laboratory for Rational Design of Drugs and Biomaterials, Pharmacy Building, Afe Babalola University, Ado-Ekiti, Nigeria; [3]Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmacy, State University of Medical and Applied Sciences, Igbo Eno 410001, Enugu State, Nigeria; [4]Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka 410001, Enugu State, Nigeria; [5]Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmacy, Maduka University Ekwegbe-Nsukka, 411110, Enugu State, Nigeria.
Correspondence: charles.nnadi@unn.edu.ng; Tel: +234 806 4947 734

Received 07 September, 2026 Revised 14 October, 2026 Accepted 19 October, 2026 Published 26 October, 2026

Copyright © 2026 the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ***ISSN (online: 3093-0715; print: 3093-0723)***

Eze SC, Nweke LI, Ojobor F, Oluyemi WM, Nnadi CO, 2026. Dusts and De-dusting of Natural Products Datasets: A Critical Review of the Challenges. Advanced Pharmaceutical and Biomedical Letters, 3(1): xxx. https://doi.org/10.59493/advpbl/2026.1.9

Abstract

Machine learning and artificial intelligence have become an established part of natural-products (NP) drug discovery. From bioactivity, hit-identification and lead optimization to toxicity and pharmacokinetic profiling, predictive models have sustained increased momentum. While many review works have focused on the models underlying predictions in NP drug discovery, a detailed understanding of how the quality of the data impacts the predictive process is still lacking. In this narrative review, we undertake the task of unravelling the shortcomings of the current datasets we have, the bottlenecks presented and ways to alleviate them. NP datasets are plagued by unique challenges—including chemical complexity, sparse bioactivity data, experimental variability, and significant toxicity data gaps—which collectively act as "dust", obscuring reliable predictions. We show how embedding data quality checks during curation and standardization during modelling can improve predictive performance. Additionally, data enrichment by augmentation and hybridization with multi-omics will be a good addition to NP drug discovery. We conclude that giving equal priority to data quality as algorithmic complexity ensures more reliable prediction and translation of AI-driven NP research into tangible therapeutic outcomes.

Keywords: Artificial intelligence, bioactivity, datasets, natural products, prediction, quality

Introduction

Machine learning (ML) and artificial intelligence (AI) have led to a paradigm shift in natural product (NP) drug discovery. It has transformed a historically serendipitous process into a more targeted, efficient, and computationally driven approach [1]. This transformation is characterized by the ability of AI to navigate the vast and complex chemical space of NPs and address long-standing bottlenecks, from compound identification to clinical translation [1–3].

Virtual screening represents one of the most established applications of ML in NP discovery. It is used to rapidly predict the bioactivity of NP libraries against specific targets. ML and deep learning (DL) architectures such as Random Forests (RF), Support Vector Machines (SVM), Graph Neural Networks (GNNs), and Convolutional Neural Networks (CNNs) are the most used models [4–6]. These models are also employed for early safety profiling, such as predicting cardiotoxicity. Another significant application is network pharmacology, which aligns with the poly-pharmacology principle: NPs interact with an array of biological targets to elicit their action, which is crucial for understanding the holistic effects of complex herbal formulations [7]. De novo design, on the other hand, leverages transformer architectures, a generative model, to propose novel, synthetically accessible scaffolds inspired by NP [8]. Platforms like DeepLigBuilder+ can generate 3D conformers of novel ligands, explicitly considering the stereochemical constraints that define NP structural richness [9].

The success of machine learning algorithms has largely depended on the availability of data. Models are fundamentally data-hungry,

and this presents a significant bottleneck to all predictive tasks [8, 10]. So, data quality is a critical factor that determines how much we can trust our models in outcome prediction. The principle of "garbage in, garbage out" (GIGO), used historically in computer science, is particularly pertinent in NP research. This is due to significant chemical complexity and data fragmentation that create unique challenges in NP, thereby compromising model performance and predictive reliability [11]. This issue can be conceptualized as "data dust" —a metaphor for the noise, biases, inconsistencies, and critical gaps that obscure the true biological signal within NP datasets [12]. In Table 1**,** the different databases used for natural product research, data types, and their potential dust are presented.

The first layer of "dust" stems from chemical representation errors. Natural products possess complex structures with multiple stereocenters, chiral centers, and tautomeric forms, yet many public databases contain incorrect SMILES (Simplified Molecular-Input Line-Entry System) strings, missing stereochemical information, or ambiguous protonation states [10, 13]. Such errors propagate through training datasets, leading models to learn spurious correlations rather than genuine structure-activity relationships [11]. Moving away from the structures, bioactivity data heterogeneity and publication bias are another significant challenge. NP bioactivity data are generated across diverse experimental assays (e.g., half maximal inhibitory concentration (IC_{50}), inhibitory constant (Ki), % inhibition, minimum inhibitory concentration (MIC)), and under varying conditions, making direct comparison and standardization very challenging [14–16]. Furthermore, the scientific literature is characterized by a pronounced positive result, which is preferentially published, while negative or inactive data remain largely unreported. While this is already a problem in the larger scientific community, it compounds an already fragile situation in modelling of NPs. This creates severely imbalanced datasets where active compounds dominate, yielding models that perform poorly in distinguishing truly inactive molecules, which is a fundamental requirement for virtual screening [16]. Recent analyses have revealed that AI-driven network pharmacology studies consistently and predictably converge on the same "hub" genes and pathways (e.g., PI3K-Akt, MAPK), suggesting that training data biases, rather than genuine biological discovery, drive these predictions [15].

Another obvious issue resides in the toxicity data gap. Unlike synthetic drug databases (e.g., Chemical Database of the European Molecular Biology Laboratory (ChEMBL), PubChem (the open chemical database at the National Institutes of Health)), NP-focused resources contain limited toxicological profiles [17]. Toxicity is inherently multi-faceted and includes hepatotoxicity, nephrotoxicity, cardiotoxicity, genotoxicity, and herb-drug interactions, with each requiring distinct experimental paradigms that are rarely systematically performed for NPs [18]. Moreover, toxicity data are scattered across different, often inaccessible sources, further compounding the challenge of building robust predictive models [19].

Many works in the past have approached AI and ML in NPs from the perspective of algorithms, outcomes, applications and potentials, yet data-quality challenges that underpin predictive model performance remain critically underexplored [20, 21]. This review addresses this gap by adopting a data-centric perspective, shifting the focus from algorithmic innovation to the quality, standardization, and integration of the data that fuels these models. In this context, we examine the entire lifecycle of NP data, from generation and curation to modelling and interpretation, with equal emphasis on bioactivity prediction and toxicity profiling, which are often neglected.

The scope of this review is comprehensive. First, we analyze the unique "data dust" challenges that often plague NP datasets, including structural misrepresentations, bioactivity heterogeneity, publication bias, and the critical scarcity of high-quality toxicity data. Then, we critically evaluate existing strategies for data quality enhancement, assessing their effectiveness and limitations in the NP context. We move on to explore emerging approaches for data integration, such as multi-omics, network toxicology, and transfer learning, with the potential to compensate for data sparsity and thus improve predictive reliability. Finally, we discuss the practical challenges of model validation, interpretability, and regulatory acceptance, offering recommendations where necessary on translating AI predictions into actionable drug discovery outcomes.

The framework involves four iterative steps: Curation ensures chemical structures, bioactivity data, and metadata are cleaned and validated for reliability. Standardization unifies ontologies, nomenclature, and formats to enable interoperability and comparison. Enrichment augments datasets via active learning, literature mining, and synthetic generation, addressing gaps in toxicity and inactivity data. Integration harmonizes chemical, biological, toxicological, genomic, and clinical information into multi- dimensional datasets, capturing the holistic complexity of NPs.

Data quality in natural product databases

Several factors contribute to challenges in the data quality of NP databases. For instance, Fig. 1 summarizes the "data dust" often encountered in NP research and modelling.

Poor chemical representations

NP databases are characterized by the presence of incorrect or ambiguous chemical structures, usually linked to errors in Simplified Molecular Input Line Entry System (SMILES) or International Chemical Identifier (InChI) notations. These notations are usually derived from literature mining of a very heterogeneous source [23]. The challenge of processing raw SMILES strings from heterogeneous sources is well known to include syntax errors, non-canonical forms, or inconsistencies that prevent proper chemical parsing. The LOTUS (natural prOducTs occUrrence database) initiative has taken steps to standardize its workflows by validating the chemical structures after translation into canonical forms from raw SMILES [24]. This process, however, is still messy and leads to loss of data and stereochemical detail. Even when SMILES strings are of required quality, duplication becomes another big challenge. Identical compounds usually appear in databases under different names, from different sources, or in different tautomeric forms, leading to inflated dataset sizes and biased model training [25]. This is particularly concerning in NP research due to the variety of forms and states that natural products can adopt. Alkaloids, flavonoids, and tannins especially, can occur in base forms and in different states of complexation with sugar moieties, thereby complicating clear compound identification and use in modelling [26]. De-duplication requires sophisticated algorithms that recognize and enforce structural equivalence despite varying representations, yet many databases lack rigorous deduplication protocols [25].

Table 1. Snapshot of popular NP databases, characteristics and possible data dusts

Database name	Data types	Possible Data Dusts	URL
Natural Products Resources			
COCONUT	Largest open-access collection of natural products; rich in metadata;	Limited pharmaceutical coverage.	(a)
DNP	Very rich in chemical and biological info.	Subscription-based; Access is limited.	(b)
AntiSMASH-DB	Hotspot for biosynthetic gene clusters, with genomics linked to metabolite prediction.	Limited chemical coverage; focused only on antibiotics	(c)
Natural Products Atlas	Well-curated microbial natural products, with strong visualization tools.	Strictly based on data from microbial sources.	(d)
Super Natural II	Incorporates predicted bioactivities of NPs	Poor in metadata. Predictions usually lack experimental validation.	(e)
TCMSP	Focused on Traditional Chinese Medicine; integrates pharmacokinetics and targets.	Limited chemical coverage	(f)
TCMID	Extensive TCM compound and formula database; links herbs to compounds.	Data quality varies; limited experimental validation.	(g)
ETCM	Curated TCM database with modern pharmacology annotations.	Narrow scope; mainly TCM-focused.	(h)
HIT	Links herbal compounds to protein targets; useful for network pharmacology.	Limited compound diversity; focused on herbal medicine.	(i)
KNApSAcK family	Broad coverage of metabolites across species; integrates organism info.	Data heterogeneity; limited bioactivity annotations.	(j)
IMPPAT	Indian medicinal plants database; integrates phytochemistry and therapeutic uses.	Regional focus; less global compound coverage.	(k)
AfroDb	Focused on African medicinal plants; unique regional coverage.	Smaller dataset compared to others.	(l)
SANDB	Specialized in South American natural products.	Regional scope; limited global coverage.	(m)
Toxicology Resources			
CTD	Database that links chemicals, genes, and diseases, with a strong focus on toxicology.	Not specialized in natural products	(n)
Tox21	Comprehensive and publicly available dataset on toxicity	Severe class imbalance; actives comprising less than 5% of dataset	(o)
ToxCast	Large toxicity-based database focused on environmental pollutants	Very low on natural product coverage; mainly synthetic compounds	(p)
Bioactivity Resources			
Drug Bank	Rich in pharmacology, targets, and clinical data.	Very low on natural product coverage.	(q)
ChEMBL	A large bioactivity database that integrates assays, targets, and compounds.	Bioactivity data can be very noisy and dispersed across assays, which requires careful and expert curation.	(r)
Binding DB	Large database includes target and inhibition kinetics of small molecules	Heterogeneity of activity and assay types	(s)
Biological/Structural resources			
UniProt	Gold standard for protein sequence and annotation	Needs integration with chemical databases	(t)
PDB	Structural data for biomolecules	Little or no direct compound bioactivity data.	(u)

Table adapted from Xu et al.[22]. ADME(T): Absorption, Distribution, Metabolism, Elimination and Toxicity
(a)https://coconut.naturalproducts.net/ (b) http://dnp.chemnetbase.com/ (c) https://antismash-db
(d) https://www.npatlas.org/ (e) https://bioinf-applied.charite.de/supernatural_3/
(f) https://www.tcmsp.com/tcmsp.php (g) http://www.megabionet.org/tcmid/
(h) http://www.tcmip.cn/ETCM2/ (i) http://www.badd-cao.net:2345/ (j) https://www.knapsackfamily.com
(k) https://cb.imsc.res.in/imppat/ (l) https://zinc.docking.org/ (m) https://sanddb.ru.ac.za/
(n) https://ctdbase.org/ (o) https://tripod.nih.gov/tox21/challenge/ (p) https://www.epa.gov/comptox-tools/exploring-toxcast-data#Download (q) https://www.drugbank.ca (r) http://www.ebi.ac.uk/chembl/
(s) https://www.bindingdb.org/rwd/bind/index.jsp (t) https://www.uniprot.org/
(u) https://www.wwpdb.org/
DNP (Dictionary of Natural Products); antiSMASH-DB (antibiotics & Secondary Metabolite Analysis SHell DataBase); TCMSP (Traditional Chinese Medicine Systems Pharmacology database and analysis platform); TCMID (Traditional Chinese Medicine Integrated Database); ETCM (Encyclopedia of Traditional Chinese Medicine); HIT (Herbal Ingredients' Targets database); IMPPAT (Indian Medicinal Plants, Phytochemistry and Therapeutics database); AfroDb (African Natural Product Database); SANDB (South African Natural Compounds Database); CTD (Comparative Toxicogenomic Database); UniProt (Universal Protein Resource); PDB (Protein Data Bank).

Natural product (NP) databases often neglect stereochemistry, despite its crucial role in bioactivity [13, 27]. NPs possess multiple stereocenters and defined 3D conformations, yet many databases store only 2D structures or omit stereochemical detail in SMILES. This simplification undermines predictive modelling, as typically only one or two stereoisomers drive bioactivity. Explicit enumeration of isomers aids models in learning fine chemical details and provides toxicity insights [13]. Ambiguous stereochemical data leads to incorrect correlations [13, 28]. Annotation challenges extend SMILES standardization, where missing stereo details merge minor isomers into canonical forms, critically reducing predictive accuracy.

Tautomers and protonation states

Natural products (NPs) often exist in multiple tautomeric forms influenced by proton concentration, directly shaping molecular properties, binding, and ADME-T profiles [29]. Yet, most databases store only one representation, missing pH-dependent variability. Tautomers, constitutional isomers in equilibrium, are critical: physalins interconvert in solution, while curcumin's diketone and keto-enol forms show solvent-dependent antioxidant mechanisms [30]. Curcumin's dianion drives >98% reactivity despite 2% abundance [31]. Protonation states also matter: α-tomatine's antifungal activity requires its unprotonated form, while aporphine alkaloids exploit pH trapping. Complex cases like sanggenon C/D highlight computational challenges with multiple pKa values [32]. The Natural Product Activity and Species Source Database, version 3.0 (NPASS 3.0) advances by including quantitative ADME-T data [33], but standard cheminformatics often neglects tautomerism and protonation, impairing predictions for pH-sensitive NPs such as alkaloids, phenols, and flavonoids [26].

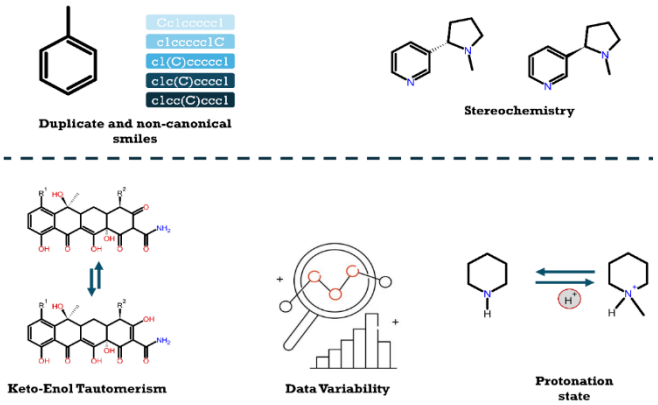


Fig. 1. Data "dust" in NP research and predictive modelling

Salts and solvates, mixtures and extracts

Salts and solvates add significant complexity to NP data dust. NPs often exist as free bases, salts, hydrates, solvates, and cocrystals, each with distinct solubility, stability, and bioavailability [34]. Berberine illustrates this: its hydrochloride salt solubility is ~1.8 mg/mL, while the sulphate salt reaches 30 mg/mL, a 16-fold difference explaining inconsistent IC_{50} values when salt forms are unspecified [35]. Databases worsen the issue, with salts comprising 45.9% of solid-form registries versus 19.3% free forms, yet most fail to distinguish them. Models trained on berberine chloride differ fundamentally from those trained on sulphate, though both are labelled "berberine." Mixtures and extracts pose greater challenges, as composition varies with genetics, environment, harvest, and processing. Magnolol and honokiol, positional isomers from *Magnolia officinalis*, block hERG channels individually (IC_{50} 5.2 and 4.7 μM) but combine additively, increasing cardiotoxic risk [36]. Their synergistic α-glucosidase inhibition further shows extract-level activity cannot be inferred from single compounds. Most databases record only isolated activities, disconnecting stored entities from therapeutic extracts. Dereplication and spectroscopy-based characterization can identify key metabolites driving these effects [37].

Positivity bias

Natural product bioactivity databases suffer from severe imbalance, with active compounds vastly outnumbering inactive ones [38]. Literature confirmation bias drives this, as positive results are preferentially published while negative findings remain underreported. Consequently, models trained mainly on actives struggle to identify genuinely inactive molecules, undermining virtual screening. Public databases like BindingDB illustrate this imbalance [39]. Potency analyses show activity classes dominated by actives, with few inactive or weakly active entries. Even active compounds often have heterogeneous measurements, with only subsets qualifying as high-confidence activity data (HCD) [28]. The lack of systematically collected inactive data critically limits robust modelling. In network pharmacology, this bias manifests as the "convergence problem," where AI-driven predictions of multi-target interactions risk distortion, despite network pharmacology's central role in NP research [40]. However, these predictions consistently converge on the same set of "hub" genes and pathways—most notably PI3K-Akt (Phosphoinositide 3-kinase / Protein kinase B signaling pathway), MAPK (Mitogen-Activated Protein Kinase), and TNF (Tumor Necrosis Factor) signaling [14, 15]. This convergence is not an indication of biological validity but rather a reflection of training data bias. The underlying databases and literature from which these network predictions are derived are themselves biased toward well-studied pathways, creating a self-reinforcing cycle. Models are trained on data where these pathways are overrepresented, and consequently, they predict the same pathways for novel compounds.

Scarcity of toxicity data

High-quality toxicity data for natural products is scarce or largely unavailable [41]. This gap is not merely an inconvenience; rather, it is a fundamental impediment to translating AI-driven NP discoveries into safe and effective therapeutics. Historically, NP research has prioritized the discovery of bioactive compounds with therapeutic potential, with safety assessment often treated as an afterthought [42]. Even plants with notable toxic bioactive compounds are carefully avoided in the systematic search and study [43]. The result is an under-reporting of toxicity data in the primary literature and, consequently, in curated databases. Researchers are more motivated to publish exciting discoveries of a new active compound than to document the lack of safety of known compounds [16]. This efficacy-centric orientation is now recognized as a major liability. We have already noted that toxicity is not a stand-alone property, but encompasses a diverse array of adverse effects [41]. Each of these endpoints requires distinct experimental

paradigms, biological models, and interpretive frameworks, making toxicity data collection, standardization, and integration challenging. Also, since toxicity can manifest through multiple mechanisms, tracking it becomes very difficult. The NPASS 3.0 database represents a notable advance, incorporating 34,975 toxicity records across seven categorical toxicity labels (carcinogenicity, organ-specific toxicity, irritation, mutagenicity, nephrotoxicity, genotoxicity, neurotoxicity) from external resources such as Tox-Val and TOXRIC [33]. However, many toxicity endpoints reported in databases are predicted endpoints with very little underlying experimental background. Thus, comprehensive toxicity coverage remains the exception, rather than the norm in NP databases [19, 41, 44].

Data leakage and non-independent evaluation

Data leakage is a critical threat in natural product modelling, arising when external data contaminates training, inflating performance [45]. In binding affinity prediction, nearly half of CASF complexes overlap with PDB-Bind training data, enabling models to memorize ligand-protein similarities [46]. Filtering algorithms revealed benchmark drops, exposing leakage-driven superiority. ToxBench showed random splitting systematically overestimates AUROC compared to scaffold splitting [46]. Leakage also occurs in preprocessing: applying feature selection or normalization before splitting lets test data influence training. A study across thirty datasets found leakage inflated performance in nearly half, with feature selection most distorting [47]. Rigorous strategies—scaffold- or cluster-based splitting and fitting preprocessing only on training data—are essential, though leakage persists [45–47].

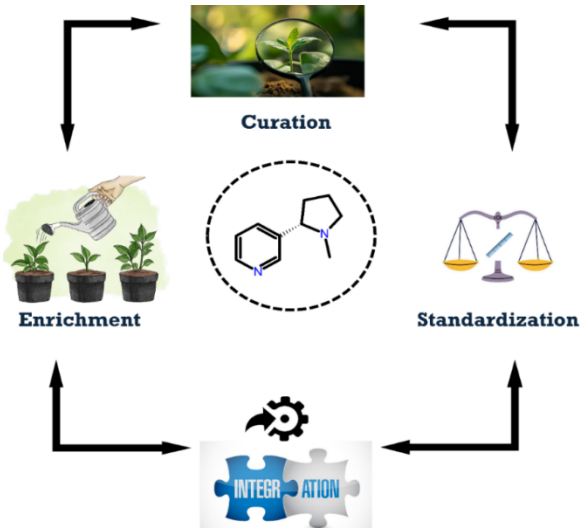


Fig. 2: Overall pipeline for quality data infrastructure in NP research and predictive modelling

Producing quality datasets for modelling of NPs

The pipeline for ensuring quality data for predictive modelling is as shown in **Fig. 2**.

Data curation and cleaning

Incorrect chemical representations often arise from literature mining errors, optical structure recognition (OCSR) inaccuracies, or simple data entry mistakes [10, 48]. Thus, automatic validation embedded in the mining pipeline must be implemented (Fig. 2). Open-source cheminformatics toolkits—particularly RDKit, Open Babel, and the Chemistry Development Kit (CDK) —provide essential functionality for detecting and correcting structural errors, including syntax validation, aromaticity detection, valence checking, and stereochemical assignment [10, 48–50]. The Cheminformatics Microservice, a recently developed open-source solution, unifies access to these toolkits through a single REST (REpresentational State Transfer) interface, eliminating the need for researchers to install and maintain multiple software packages [51]. This platform provides ready-to-use routines for chemical structure validation, descriptor calculation, 3D conformer generation, and format conversion. The integration of RDKit, Open Babel, and CDK within a single service enables comprehensive validation workflows that catch errors that any single toolkit might miss.

However, even with these tools, the challenge of structure validation is far from trivial. Different toolkits can produce inconsistent results, particularly for complex NP structures. An RDKit mol object might slightly differ from one generated from Open Babel [52]. Researchers using these tools must understand their specific requirements and limitations, particularly regarding stereochemical information. The need for multi-tool validation workflows is clear, as using a single toolkit may miss errors that another would detect.

Standardization of bioactivity data

We have noted that heterogeneity of bioactivity measurements represents a major obstacle to model training and cross-study comparison. These different metrics are not directly comparable without conversion, unit harmonization, and standardization [10]. The Chemical Database of the European Molecular Biology Laboratory (ChEMBL) database exemplifies best practices in bioactivity data standardisation [53]. ChEMBL applies rigorous curation protocols to convert activity measurements to unified scales, particularly pIC_{50} (the negative logarithm of IC_{50}, in molar units) and pKi (negative logarithm of the inhibition constant), which are directly comparable across assays. A typical standardisation protocol includes unit conversion (to nM), value transformation to pIC_{50}/pKi scales, activity flagging (or classification), and confidence annotations [54]. IC_{50} and Ki values are usually converted to pIC_{50} and pKi scores, standardized by unit, and merged with compound structures for cheminformatics analysis [55]. This has become a standard approach in modelling, even in synthetic compound databases. For natural product datasets, however, the lack of standardized assay conditions poses particular challenges. Many assays are reported as effects of crude extracts or fractions of plant parts, complicating clear assay endpoint annotation [56]. Assay interference from colloidal aggregation, chelation, chemical reactivity, and light attenuation can produce false assay readouts that masquerade as genuine bioactivity[57].

Even after standardizing potency values, NP datasets often contain duplicates and conflicting bioactivity measurements for the same compound-target pair. These inconsistencies stem from multiple publications, assay variability, differing protocols, and data entry errors [58]. Best practice is to retain all records with metadata and confidence annotations, then apply consensus strategies (median/mean) for high-confidence measurements, remove outliers, and maintain provenance [55, 58]. A hop extract study

showed harmonization's importance: ABTS, DPPH, and TPC yielded different rankings, while unit harmonization required excluding datasets reported in alternative calibration systems (ferulic vs. gallic acid equivalents) [56]. Thus, standardisation must consider both values and experimental context.

Consistency in data naming, storage and transfer

Interoperability is as important as the quality of the dataset itself. NP datasets must adopt consistent naming conventions, ontologies, and data formats to ensure similar outcomes across tools, platforms and use cases [59]. This step is essential for enabling meaningful comparisons and aggregations across different studies. The FAIR principles—Findable, Accessible, Interoperable, and Reusable—provide an essential framework for ensuring that research data are discoverable and useful for future studies [60]. These principles are particularly critical in NP research, where data are generated across diverse disciplines and often remain siloed in incompatible formats. The MassBank database exemplifies FAIR implementation at a high level. As an open-access mass spectral data repository, MassBank implements the FAIR principles through extensive cross-referencing (keywords and unique identifiers), accessibility by making the datasets freely available with clear usage licenses, interoperability through standard data formats for software integration, and reusability [61] and reusability [61]. The COCONUT (COlleCtion of Open Natural ProdUcTs) natural products database represents a community-driven platforms that further illustrate FAIR implementation in NP research, employing open standards and ontology-based annotations to simplify data reuse [62]. These platforms, combined with the ongoing efforts of the ELIXIR European life sciences infrastructure to recognize and integrate community resources, are laying the groundwork for a more interoperable NP data ecosystem [63].

Natural products are sourced from thousands of plant species, each with complex taxonomic histories—accepted scientific names, synonyms, vernacular names, and frequently changing classifications. Without rigorous name standardisation, linking chemical structures to biological sources becomes impossible [64, 65]. World Flora R package is a tool that enables researchers to match and standardize plant names against this authoritative database, addressing the challenge of synonyms and taxonomic changes [66]. A recent study published in Scientific Reports highlights a critical nuance: storing accepted scientific names alone can lead to misinterpretation of botanical data [67]. Because taxonomic databases are updated over time, a name that resolves to a particular accepted name in one database version may resolve differently in a later version (name intransitivity) [68]. Approximately 1% of names in the World Checklist of Vascular Plants (WCVP) and 6% of names in WFO (World Flora Online) fall into this, with discrepancies increasing as new taxonomy versions are released [64]. The authors recommend that botanical datasets store verbatim names as found and resolve them to accepted names "as close to the point of end-use as possible, using the most recent taxonomic database". This practice preserves data provenance while enabling ongoing name harmonization.

Data enrichment

Addressing gaps in toxicity data, inactive compounds, and NPs from understudied organisms is also challenging. Data enrichment addresses these gaps through strategic augmentation, computational prediction, and intelligent experimental design [69, 70]. Active learning leverages AI to identify the most informative experiments to run next, maximizing data quality per experimental effort. This approach is particularly valuable in NP research, where experimental validation is resource-intensive [71]. In a recent study, active learning was applied to identify a synthetic mimic of a macrocyclic natural product antifungal [72]. Starting with just 100 compounds, through five rounds of 20-compound selection and model refinement, the authors successfully identified florylpicoxamid, a broad-spectrum crop antifungal. Exploitation versus exploration, as a tool of enrichment, enhanced this discovery [73]. Critically, the procedure broadened the domain of applicability of successive models while also enhancing predictive accuracy. For NP toxicity prediction, active learning could be applied to prioritize the most informative compounds for toxicity testing, dramatically reducing the number of experiments needed to build robust safety models while ensuring coverage of diverse chemical space [19, 74, 75].

Where experimental data are scarce, generative models can augment training sets by generating realistic synthetic data. The goal is to expand dataset size and diversity while maintaining fidelity to real experimental measurements [8, 13]. A study on herbal medicine homologous classification demonstrates this approach for near-infrared (NIR) spectral data [76]. Researchers proposed DDIM-NIR, a novel U-Net-based one-dimensional denoising diffusion implicit model, specifically tailored for generating high-fidelity NIR spectra of medicinal herbs. The method was applied to a comprehensive dataset of 48 varieties of medicinal and food homologous (MFH) herbs across eight functional groups. The diffusion model approach was selected over GAN-based alternatives because it avoids common issues like convergence instability, mode collapse, and difficulty modelling long-range spectral dependencies. Results showed that data augmentation with the diffusion model significantly improved downstream classification performance, addressing the critical challenge of data scarcity that limits machine learning model robustness in NP research. This approach could be adapted for generating realistic bioactivity or toxicity measurements for NPs where experimental data are limited. However, the quality of synthetic data depends heavily on the quality and representativeness of the training data, and extensive validation is required to ensure that augmented data do not introduce systematic biases [77–79].

Data mining for unstructured data

There remains a vast amount of bioactivity and toxicity data hidden in text, tables, and figures—data that remain inaccessible to conventional computational analysis [23, 80]. Despite the efforts made, there remain significant gaps. Many groups have advanced Natural Language Processing (NLP) and multimodal AI systems to extract this information systematically [48, 49]. The Bio-Miner is a multimodal agentic system specifically designed for protein-ligand bioactivity extraction from literature. Through a structured decomposition of tasks, Bio-Miner enables robust handling of complex chemical representations, especially Markush structures, and supports extraction from text, tables, and figures. Evaluation on the Bio-Vista benchmark showed Bio-Miner achieved an F1 score of 0.33 for extracting complete bioactivity triplets, while a

one-shot baseline achieved only 0.00037, highlighting the inherent difficulty of the task and the necessity of principled task decomposition [80]. Using Bio-Miner, a bioactivity pre-training database was constructed containing 82,262 data points extracted from 11,683 papers within two days. Models pre-trained on this database demonstrated improved performance (3.9% and 3.6% improvement in root mean square error (RMSE)) on independent test sets compared to models trained only on existing curated data [80]. Bio-Miner has also been integrated into a human-in-the-loop (HITL) workflow for data augmentation of NLRP3 data available in ChEMBL, and automated labelling of 242 structures within the Pose Busters database at an average rate of 2 minutes/structure [80]. Similarly, NaturaPredicta™ applies NLP techniques to the PubMed database to predict bioactivity and identify bioactive compounds in botanical ingredients using functional scores from existing approved ingredients. This is in contrast to molecular structure-based data extraction techniques, thereby highlighting the potential of mining unstructured text for functional information beyond simple chemical descriptors [81]. Rajan and coworkers recently deployed Machine Learning Assisted Recognition of Chemical Structures from Unstructured Sources (MARCUS) for natural product literature curation with COCONUT database schema-aware mapping [23]. MARCUS integrates text annotation, optical chemical structure recognition Optical Chemical Structure Recognition (OCSR), and direct submission to the COCONUT database. Using a fine-tuned GPT-4 model for chemical entity recognition and an ensemble of OCSR tools (DECIMER - Deep Learning for Chemical Structure Recognition, Mol-NexTR, Mol-Scribe) with human-in-the-loop validation, it supports stereochemical validation, taxonomic metadata linking, and manual refinement through Ketcher editing [23].

Data integration

Multi-Omics and systems toxicology

Chemical structure and single bioactivity measurements are one-dimensional data, which always fail to capture the complex biological context in which NPs exert their effects [10, 48]. Multi-omics integration and systems-level approaches will address this limitation. Metabolomics provides a direct readout of the small-molecule constituents of a biological system, offering a crucial link between predicted compounds and real-world metabolic profiles [11, 21]. Multi-omics strategies offer a systemic framework to delineate intricate NP-target interactomes. This is an improvement over conventional approaches, which merely interrogate discrete molecular layers and fail to capture cross-omics perturbations spanning the genome, epigenome, transcriptome, and proteome [75]. AI makes it possible to integrate these high-dimensional datasets and identify disease-associated molecular signatures, prioritizing candidate targets that single-omics analyses would miss. De-replication is the first beneficiary of this approach, which rapidly identifies known compounds to prioritize novel ones [12, 28]. Previously inaccessible sources come within reach when computational chemical methods are integrated with metabolomics. This allows the identification of novel bioactive scaffolds and discovery of cryptic metabolites, which otherwise would not have been possible.

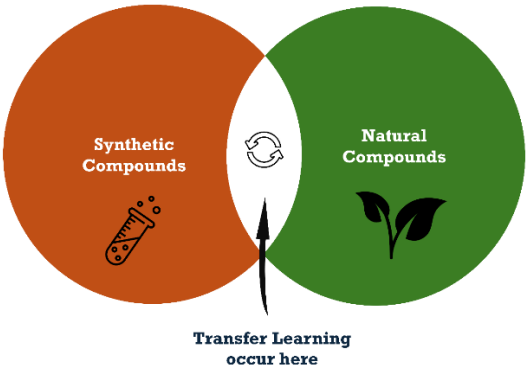


Fig. 3. Transfer learning using foundational models grounded in synthetic data, and fine-tuned on limited NP data

Metabolomics data, when integrated with AI-driven predictions, can help researchers validate predicted compounds (before an experimental set-up), identify novel active metabolites and contextualize activity by linking bioactivity to actual metabolic profiles [74, 75]. Metabolomics can be extended into network toxicology to allow us to elucidate molecular mechanisms of toxicities [7]. Perturbations caused by chemical exposures can easily be linked to key metabolites by integrating chemical–target information, disease-associated genes, and pathway enrichment analysis [74]. Network toxicology constructs an interaction network model of "toxicity-gene-target-drug" to depict the toxic properties of a compound, enabling the understanding of the mechanisms and material basis of toxicity at the systemic level.

Transfer learning

Transfer learning complements data augmentation in addressing sparse NP datasets (**Fig. 3**). Models are pre-trained on large, high-quality sources like ChEMBL or ZINC, then fine-tuned on smaller NP datasets [82]. This mirrors other fields where foundational models adapt to downstream tasks [83]. The KAN-PROSPECT framework applied cross-modal transfer learning, pre-trained on ~3,800 drugs and fine-tuned on 400 NPs [78]. It outperformed baselines in dual-label prediction (efficacy/adverse effects), excelling in few-shot and zero-shot scenarios. The TransGenGRU model extended this by combining transfer learning with generative modelling, producing novel NP derivatives with specific inhibitory activities [82]. While effective, challenges remain due to the large gap between NP and synthetic chemical space. [77]. In this context, negative transfers can ensure a situation where the generated molecules tilt towards synthetic scaffolds due to data imbalance. NPs differ significantly in structural complexity, stereo-centers, and physicochemical properties from typical synthetic drugs. Transfer learning therefore struggles to create unique mappings between the two chemical spaces, leading to poorer outcomes than expected [79]. Domain adaptation techniques are designed to handle these differences. This was employed in the Twisted Domain Adversarial Network (TwistDAN) framework to expand the original chemical space with up to 97% novel scaffolds [84]. The TwistDAN framework uses a semi-supervised adversarial learning approach via gradient reversal layers. It combines supervised classification on a labelled domain (D0) with adversarial domain confusion across all samples, forcing the feature extractor to learn representations that are simultaneously discriminative for the target task while being

indistinguishable across source and target domains. This prevents overfitting to domain-specific patterns and enables generalisation to the NP domain. Other techniques include robust assessment of source-target similarity, chemical space overlap, embedding alignment, scaffold coverage, and domain shift in transfer learning tasks.

Hybrid models

Research efforts have increasingly integrated physics-based calculations (molecular docking and binding free energy calculations) with ML and AI, in a hybrid model, to ensure more robust and interpretable predictions (Fig. 4). This is necessary to enforce a mechanistic ground truth, from which interpretability emerges in the prediction [85]. Molecular docking and MM-GBSA (Molecular Mechanics with Generalized Born and Surface Area solvation) and even more rigorous free energy perturbation (FEP) calculations provide physics-based estimates of ligand-target binding affinity [86, 87]. While computationally intensive, these methods generate features that reflect actual physical interactions, thereby complementing purely statistical pattern recognition. The workflow is akin to an inverted cone, where faster and less accurate ML and AI models perform the initial screening of millions of compounds. Then rigorous re-ranking and filtering are done using molecular docking, free energy calculations and solvation energy estimations [88].

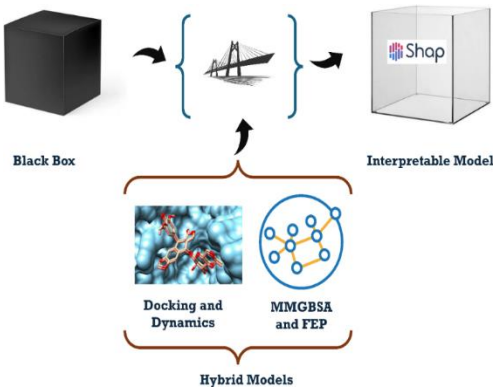


Fig. 4. From Black-box to interpretable model using SHAP analysis and hybrid models

Interpretable AI

The "black box" nature of many AI models limits their acceptance in pharmaceutical research and regulatory contexts (Fig. 4) [89]. Interpretable AI addresses this by providing explanations for predictions, building trust and facilitating experimental validation. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) are the most widely used interpretability frameworks [90]. SHAP and LIME identify feature importance at global and local levels, enabling chemically meaningful insights into compound behavior. For example, SHAP and LIME analyses identified "EffectiveRotorCount3D" and "YStericQuadrupole3D" as the most influential descriptors for drug activity—features corresponding to molecular flexibility and steric effects that are chemically interpretable [91]. The integration of SHAP and LIME enables chemically meaningful insights into compound behavior, supporting informed molecular design and increasing model transparency. Combining ensemble modelling with explainability improves both prediction robustness and interpretability, enhancing confidence in AI-driven decision-making in the drug discovery process. This dual-layer approach is scalable for broader datasets and offers actionable insights for rational therapeutic development and regulatory alignment [89, 90]. SHAP and LIME, however, have limitations. SHAP has significant processing costs and difficulties with linked data because it erroneously assumes feature independence [92]. Meanwhile, LIME exhibits high local instability and poor global accuracy, due to the utilization of random neighborhood disturbances [93]. Ultimately, both techniques can produce false feature significance scores when features are substantially collinear. In addition, feature attribution does not carry the weight of mechanistic explanation.

Data provenance

Data provenance—the documented history of data origins, transformations, and ownership—is vital yet often overlooked in natural product (NP) research. While the "dusting" framework addresses errors, gaps, and enrichment, it must also ensure traceability. Without provenance, curated datasets remain unreliable, as authenticity, reproducibility, and fitness cannot be verified [94]. Provenance requires source attribution—including plant species, geographic origin, voucher specimens, and experimental conditions [95]—plus transformation history recording conversions, filtering, and algorithmic steps [49]. Version control tracks updates and corrections. Technically, frameworks like W3C PROV and blockchain ensure tamper-proof records [96]; culturally, journals and funders must mandate documentation. Databases such as ChEMBL and MassBank exemplify best practices by linking compounds to literature and protocols [12]. Thus, provenance must be integrated as a cross-cutting requirement, ensuring every data point is traceable from biological origin to predictive modelling, enabling transparency, reproducibility, and trust for scientific and regulatory acceptance.

Challenges in predictive modelling of natural products, and way forward

The stand-off between predictive accuracy and biological interpretability remains a deep tension in natural product drug discovery. Deep learning models, particularly Graph Neural Networks and Transformers, achieve superior performance but operate as "black boxes" whose decision-making processes are opaque to human understanding [12, 97]. This lack of understanding means we cannot reproduce or extrapolate to data outside of the train/test sets. While interpretability frameworks like SHAP and LIME provide post-hoc explanations, they offer approximations rather than true mechanistic understanding [12].

AI models can rapidly screen compounds, yet experimental validation remains costly and slow, creating a bottleneck [71]. Active learning helps by prioritizing informative compounds [11], but toxicity assessment still requires multiple assays and animal models [19]. Bridging computational and experimental labs demands institutional support. Data sharing faces barriers—intellectual property concerns and lack of curation incentives [10, 48]. A cultural shift is needed, recognizing curation as a scientific contribution. FAIR principles guide stewardship, requiring funding [60].

Federated learning offers solutions for proprietary data [98]. Sustainability must also be addressed: AI can predict harvesting levels, identify alternatives, and support conservation, yet ecological considerations remain under-integrated, requiring collaboration with ecologists and conservation biologists [21].

Limitations of the study

This review highlights diverse forms of data dust in NP predictive modelling but acknowledges limitations. Its scope is broad—covering bioactivity, toxicity, curation, standardisation, enrichment, integration, challenges, and future directions—yet sacrifices depth, with each subtopic warranting a dedicated review. Readers seeking technical detail may find it insufficient, and rapid AI advances mean tools and practices described are only representative. The proposed "dusting" framework (Curation, Standardisation, Enrichment, Integration) synthesizes prior works without empirical validation, leaving utility and scalability untested. Perspectives and biases may shape conclusions, yet the synthesis remains valuable. These limitations should be seen as opportunities for further investigation, not flaws.

Conclusion

In this review, we have provided data "dusting" as an analogy for rigorous frameworks —curation, standardisation, enrichment, and integration— to transform raw, noisy data into reliable inputs for AI-driven prediction. High-quality, standardized datasets are the true foundation upon which the future of AI-driven phytomedicine will be built. Our algorithmic sophistication cannot compensate for erroneous, biased, or incomplete datasets. By systematically "dusting the dataset," we can improve confidence in AI predictions and help accelerate discovery of life-saving medicines from the reservoir of nature's abundance. This, therefore, is a call on researchers, funding agencies, publishers, and institutions to balance focus on novel architectures with novel data curation strategies. This will deliver the promise of prediction in natural product drug discovery.

Declarations

Ethical approval
Not applicable

Informed consent
Not applicable

Consent for publication
Consent for publication was obtained from all authors

Availability of data and materials
All relevant data are included within the manuscript

Conflict of interest
The authors declare no conflict of interest

Funding
The authors did not receive any funding for the conduct of this study

Authors' contributions – CrediT
CON: Conceptualization, Methodology, Writing – review and editing; WMO: Writing – review and editing; SCE: Methodology, Data curation, Writing – original draft; LIN: Writing – review and editing; FO: Writing – review and editing.

Acknowledgements
None

Artificial Intelligence (AI) disclosure
The authors used generative AI for language editing and grammar improvement. The authors reviewed and take full responsibility for the final content.

References

1. Ferreira FJN, Carneiro AS, 2025. AI-Driven Drug Discovery: A Comprehensive Review. ACS Omega, 10: 23889–23903. https://doi.org/10.1021/acsomega.5c00549.
2. Paerhati Y, Aikebaier A, Dilimulati D, Baishan A, Yusufujiang N, Qiu X, Wusiman Y, Zhou W, 2026. Rethinking nature's pharmacy: AI era and natural product drug discovery. Pharmaceuticals, 19(2): 301. https://doi.org/10.3390/ph19020301
3. Bilotta M, Rocca R, Alcaro S, 2025. Next-generation drug discovery: The AI revolution in pharmaceutical research. Artificial Intelligence in the Life Sciences, 8: 100149. https://doi.org/10.1016/j.ailsci.2025.100149
4. Simoben CV, Babiaka SB, Moumbock AFA, Namba-Nzanguim CT, Eni DB, Medina-Franco JL, Günther S, Ntie-Kang F, Sippl W, 2023. Challenges in natural product-based drug discovery assisted with in silico-based methods. RSC Advances, 13: 31578–31594. https://doi.org/10.1039/d3ra06831e
5. Ayoka TO, Nnadi CO, 2025. A robust SMILES-based prediction of aqueous solubility of diverse antiplasmodial compounds using machine learning algorithms. Letters in Applied NanoBioScience, 14(3): 187. https://doi.org/10.33263/LIANBS143.187.
6. Hou C, Gao Y, Lin X, Wu J, Li N, Lv H, Cheng-Chung Chu W, 2025. A review of recent artificial intelligence for traditional medicine. Journal of Traditional and Complementary Medicine, 15: 215–228. https://doi.org/https://doi.org/10.1016/j.jtcme.2025.02.009.
7. Sunday JE, Okoro EC, Nwose AA, Ugwuanyi CS, Mgbe CC, Onwuasoanya PC, Abana SI, Ugwuodo BE, Nnadi CO, Obonga WO, 2026. Network pharmacology identification of Aframomum-derived anti-arthritic agents. Discover Chemistry, 3(1):350. https://doi.org/10.1007/s44371-026-00824-0.
8. Bhadra S, Pundir C, Mukherjee M, Kar A, Banerjee S, Charoensup R, Duangyod T, Mukherjee PK, 2025.

Reimagining ethnopharmacology with generative AI: Towards inclusive, ethical, and data-driven traditional medicine. Pharmacological Research, 221: 108002. https://doi.org/https://doi.org/10.1016/j.phrs.2025.108002.

9. Li Y, Pei J, Lai L, 2021. Structure-based de novo drug design using 3D deep generative models. Chemical Science, 12: 13664–13675. https://doi.org/10.1039/d1sc04444c.
10. Esaki T, Ikeda K, 2026. Data curation in cheminformatics: importance and implementation. Journal of Cheminformatics, 18: 43. https://doi.org/10.1186/s13321-026-01174-w.
11. Gangwal A, Ansari A, Siddique MUM, Sawale J, Padmane S, Wadekar R, 2026. Integrating artificial intelligence into drug discovery from medicinal plants: Current applications and infrastructural challenges. Chemical Biology & Drug Design, 108. John Wiley & Sons, Ltd: e70372. https://doi.org/10.1111/cbdd.70372.
12. Bender A, Cortés-Ciriano I, 2021. 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 Discovery Today, 26: 511–524. https://doi.org/https://doi.org/10.1016/j.drudis.2020.12.009.
13. Pereira TO, Abbasi M, Oliveira RI, Guedes RA, Salvador JAR, Arrais JP, 2023. Artificial intelligence for prediction of biological activities and generation of molecular hits using stereochemical information. Journal of Computer-Aided Molecular Design, 37: 791–806. https://doi.org/10.1007/s10822-023-00539-9.
14. Hasanzadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA, 2025. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. NPJ Digital Medicine, 8: 154. https://doi.org/10.1038/s41746-025-01503-7.
15. Diao X, Zhang H, Wang S, Wang Z, Zhang Q, 2026. Rethinking network analysis in ethnopharmacology: a multi-omics and AI roadmap to overcome conceptual and methodological biases. Frontiers in Pharmacology, 17: 1748478. https://doi.org/10.3389/fphar.2026.1748478
16. Sieg J, Flachsenberg F, Rarey M, 2019. In need of bias control: Evaluating chemical data for machine learning in structure-based virtual screening. Journal of Chemical Information and Modelling, 59: 947–961. https://doi.org/10.1021/acs.jcim.8b00712.
17. Didigwu OK, Nnadi CO, 2024. Drug-likeness, pharmacokinetics, and toxicity prediction of phytotoxic terpenoids. Proceedings, 102(1), 47. https://doi.org/10.3390/proceedings2024102047.
18. Didigwu OK, Nnadi CO, 2024. Theoretical drug-likeness, pharmacokinetics, and toxicities of phytotoxic terpenoids from the Toxic Plants-Phytotoxins. Tropical Journal of Natural Product Research, 8(10): 8867-8873. https://doi.org/10.26538/tjnpr/v8i10.31.
19. Nnadi CO, Banda SH, Uzonwanne MU, Didigwu OK, 2024. Cancer stem cells as potential targets of phytotoxic alkaloids: Drug-likeness prediction and molecular docking studies. Proceedings, 100(1): 21. https://doi.org/10.3390/proceedings2024100021.
20. Hou C, Gao Y, Lin X, Wu J, Li N, Lv H, Chu WC, 2025. A review of recent artificial intelligence for traditional medicine. Journal of Traditional and Complementary Medicine, 15: 215–228. https://doi.org/https://doi.org/10.1016/j.jtcme.2025.02.009.
21. Othman ZK, Ahmed MM, Kasimieh O, Musa SS, Branda F, Cue EG, Ocampo JMA, Prisno DEL, Vimolmangkang S, 2025. Artificial intelligence for natural product drug discovery and development: current landscape, applications, and future directions. Intelligence-Based Medicine, 12: 100316. https://doi.org/10.1016/j.ibmed.2025.100316.
22. Xu T, Xu Y, Zhang J, Zhou Y, Feng H, Zhang A, Zhang Y, 2026. Advances in AI for predicting pharmacological properties of natural medicines. Life Sciences, 387: 124180. https://doi.org/10.1016/j.lfs.2025.124180.
23. Rajan K, Weißenborn V, Lederer L, Zielesny A, Steinbeck C, 2025. MARCUS: molecular annotation and recognition for curating unravelled structures. Digital Discovery, 4: 3137–3148. https://doi.org/10.1039/d5dd00313j.
24. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, Graham JG, Stephan R, Page R, Vondrášek J, Steinbeck C, 2022. The LOTUS initiative for open knowledge management in natural products research. Edited by David A Donoso, Anna Akhmanova, and Charles Tapley Hoyt. ELife, 11: e70780. https://doi.org/10.7554/eLife.70780.
25. Kaur R, Singh HJ, Chana I, 2026. Machine learning-based data deduplication: Techniques, challenges, and future directions. Concurrency and Computation: Practice and Experience, 38: e70574. https://doi.org/10.1002/cpe.70574.
26. Letchuman S, Madhuranga HDT, Madhurangi BLNK, Premarathna AD, Saravanan M, 2025. Alkaloids unveiled: A comprehensive analysis of novel therapeutic properties, mechanisms, and plant-based innovations. Intelligent Pharmacy, 3: 268–276. https://doi.org/10.1016/j.ipha.2024.09.007.
27. Chen Y, Kirchmair J, 2020. Cheminformatics in Natural Product-based Drug Discovery. Molecular Informatics, 39: 2000171. https://doi.org/10.1002/minf.202000171.
28. Bender A, Cortes-Ciriano I, 2021. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: A discussion of chemical and biological data. Drug Discovery Today, 26: 1040–1052. https://doi.org/10.1016/j.drudis.2020.11.037.
29. Araujo de Oliveira AP, Wegermann CA, Ferreira AMDC, 2024. Keto-enol tautomerism in the development of new drugs. Frontiers in Chemical Biology, 3: 1400642. https://doi.org/10.3389/fchbi.2024.1400642.
30. Wu JP, Li LY, Li JR, Yu M, Zhao J, Xu QM, Gu YC, Zhang T, Zou ZM, 2022. Silencing tautomerization to isolate unstable physalins from Physalis minima. Journal of Natural Products, 85(6):1522-39. https://doi.org/10.1021/acs.jnatprod.2c00101.

31. Truong DH, Dinh TT, Trinh TM, Pham TH, Pham MQ, Gawlik-Dziki U, Dao DQ, 2025. HOO radical scavenging activity of curcumin I and III in physiological conditions: a theoretical investigation on the influence of acid–base equilibrium and tautomerism. RSC Advances, 15(8):5649-64. https://doi.org/https://doi.org/10.1039/d4ra07769e.
32. Schwarzinger J, Adelsberger S, Ortmayr K, Stellnberger SL, Tahir A, Hädrich G, Pichler V, Rollinger JM, Grienke U, Dailey LA, 2024. Biopharmaceutical profiling of anti-infective sanggenons from Morus alba root bark for inhalation administration. International Journal of Pharmaceutics: X, 8:100272. https://doi.org/10.1016/j.ijpx.2024.100272.
33. Lin H, Hou D, Jiang X, Yin R, Yu S, Lu S, Wang S, Ju D, Zhao H, Chen Y, Zeng X, 2026. NPASS database update 2026: comprehensive quantitative composition, bioactivity, and ADME-Tox data of natural products for biomedical research. Nucleic Acids Research, 54(D1): D1519-D1527. https://doi.org/10.1093/nar/gkaf1196.
34. Wang M, Wang Z, Zhang J, Zhan J, Wu C, Yu W, Fan W, Tang J, Zhang Q, Zhang J, 2022. Sustainable bioactive salts fully composed of natural products for enhanced pharmaceutical applicability. ACS Sustainable Chemistry & Engineering, 10(31):10369-10382. https://doi.org/10.1021/acssuschemeng.2c03138.
35. Battu SK, Repka MA, Maddineni S, Chittiboyina AG, Avery MA, Majumdar S, 2010. Physicochemical characterisation of berberine chloride: a perspective in the development of a solution dosage form for oral delivery. AAPS PharmSciTech, 11(3):1466-1475. https://doi.org/10.1208/s12249-010-9520-y.
36. Zhao W, Hao T, Lu H, Jiang C, Xu J, Zhang Z, Stalin A, Zhang X, Pan L, Xu J, 2026. hERG Channel Blockade and Additive Interactions of Magnolol and Honokiol from Magnolia Species. Toxicology Letters, 2:113152. https://doi.org/10.1016/j.toxlet.2026.113152.
37. Kellogg JJ, Jordan RT, Ranaweera MM, Custer K, Anez SG, Bendlin J, Chacon FT, Chen X, 2026. Cultivar to chemotype: characterising complex botanicals with mass spectrometry metabolomics. Natural Product Reports, 43(6):1181-200. https://doi.org/https://doi.org/10.1039/d5np00040h.
38. Albattah W, Khan RU, 2025. Impact of imbalanced features on large datasets. Frontiers in Big Data, 8: 1455442. https://doi.org/10.3389/fdata.2025.1455442.
39. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J, 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44: D1045–D1053. https://doi.org/10.1093/nar/gkv1072.
40. Panossian A, 2025. Trends and pitfalls in the progress of network pharmacology research on natural products. Pharmaceuticals, 18(4): 538. https://doi.org/10.3390/ph18040538.
41. Nnadi CO, Ozioko LM, Eneje GC, Onah CM, Obonga WO, 2020. In vivo antitrypanosomal effect and in silico prediction of chronic toxicity of N-methylholaphyllamine in rats. Tropical Journal of Pharmaceutical Research, 19(11): 2369-2375. http://dx.doi.org/10.4314/tjpr.v19i11.18.
42. Kumar V, Lalotra GS, Sasikala P, Rajput DS, Kaluri R, Lakshmanna K, Shorfuzzaman M, Alsufyani A, Uddin M, 2022. Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques. Healthcare, 10(7): 1293. https://doi.org/10.3390/healthcare10071293.
43. Zeng T, Li J, Wu R, 2024. Natural product databases for drug discovery: Features and applications. Pharmaceutical Science Advances, 2: 100050. https://doi.org/10.1016/j.pscia.2024.100050.
44. Beressa TB, Annu A, Mtewa AG, 2020. Toxicity Protocols for Natural Products in the Drug Development Process. In Poisonous Plants and Phytochemicals in Drug Discovery, 189–212. https://doi.org/10.1002/9781119650034.ch9.
45. Graber D, Stockinger P, Meyer F, Mishra S, Horn C, Buller R, 2025. Resolving data bias improves generalisation in binding affinity prediction. Nature Machine Intelligence, 7(10):1713-1725. https://doi.org/10.1038/s42256-025-01124-5.
46. Kartic YS, Yi S, Park T, 2026. ToxBench: A leakage-audited multi-task benchmark for predictive toxicology with calibration, uncertainty, and applicability domain analysis. BMC Bioinformatics, https://doi.org/10.1186/s12859-026-06621-x.
47. Becker AE, Recamonde-Mendoza M, 2026. Mind the Gap: Investigating the Impact of Data Leakage on Machine Learning Predictive Models. In Intelligent Systems, ed. Rosiane de Freitas and Diego Furtado, 166–180. Cham: Springer Nature Switzerland.
48. Schiebroek CCG, Landrum GA, Riniker S, 2026. Balancing data quantity and quality: Evaluating curation strategies for bioactivity prediction in lead optimisation. Journal of Chemical Information and Modelling, 66(13): 7446–7452. https://doi.org/10.1021/acs.jcim.6c01018.
49. Oguike OE, Ugwuishiwu CH, Asogwa CN, Nnadi CO, Obonga WO, Attama AA, 2022. Systematic review on the application of machine learning to quantitative structure–activity relationship modelling against Plasmodium falciparum. Molecular Diversity, 26(6): 3447-3462. https://doi.org/10.1007/s11030-022-10380-1.
50. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR, 2011. Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3: 33. https://doi.org/10.1186/1758-2946-3-33.
51. Rajan K, Chandrasekhar V, Sharma N, Kanakam SRS, Baensch F, Steinbeck C, 2025. Cheminformatics Microservice V3: a web portal for chemical structure manipulation and analysis. Journal of Cheminformatics, 17: 142. https://doi.org/10.1186/s13321-025-01094-1.
52. Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, Bellis LJ, De Veij M, Leach AR, 2020. An open-source chemical structure curation pipeline using RDKit. Journal of Cheminformatics, 12: 51. https://doi.org/10.1186/s13321-020-00456-1.

53. Zdrazil B, 2025. Fifteen years of ChEMBL and its role in cheminformatics and drug discovery. Journal of Cheminformatics, 17: 32. https://doi.org/10.1186/s13321-025-00963-z.
54. Theisen R, Wang T, Ravikumar B, Rahman R, Cichońska A, 2024. Leveraging multiple data types for improved compound-kinase bioactivity prediction. Nature Communications, 15: 7596. https://doi.org/10.1038/s41467-024-52055-5.
55. Igwezeke FO, Nnadi CO, 2026. TransQSAR-pf: A Bio-Informed QSAR framework using Plasmodium falciparum stress signatures for enhanced antiplasmodial activity prediction. Engineering Proceedings, 124(1): 37. https://doi.org/10.3390/engproc2026124037.
56. Ayoka TO, Nnadi CO, 2026. Application of machine learning models in predicting malaria prevalence in Nigeria: An analysis of the 2015–2020 demographic and health surveys. Journal of Parasitic Diseases, 50(2):431-443. https://doi.org/10.1007/s12639-025-01880-6
57. Nnadi CO, Althaus JB, Nwodo NJ, Schmidt TJ, 2018. A 3D-QSAR study on the antitrypanosomal and cytotoxic activities of steroid alkaloids by comparative molecular field analysis. Molecules, 23(5): 1113. https://doi.org/10.3390/molecules23051113
58. Aditya T, Srinivas TA, Sravanthi Y, Kumar Y, Srihith IV, 2023. Data Standardisation: Key to Effective Data Integration. Advanced Innovations in Computer Programming Languages, 6(1): 1-4. https://doi.org/10.5281/zenodo.10060920.
59. Stumpfe D, Bajorath J, 2011. Assessing the confidence level of public domain compound activity data and the impact of alternative potency measurements on SAR analysis. Journal of Chemical Information and Modelling, 51: 3131–3137. https://doi.org/10.1021/ci2004434
60. Wilkinson MD., Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3: 160018. https://doi.org/10.1038/sdata.2016.18.
61. Neumann S, Meier R, Wenk M, Elapavalore A, Nishioka T, Schulze T, Stravs M, Tsugawa H, Matsuda F, Schymanski EL, 2026. MassBank: An open and FAIR mass spectral data resource. Nucleic Acids Research, 54(D1): D601–D606. https://doi.org/10.1093/nar/gkaf1193.
62. Chandrasekhar V, Rajan K, Kanakam SR, Sharma N, Weißenborn V, Schaub J, Steinbeck C, 2025. COCONUT 2.0: A comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Research, 53(D1): D634–D643. https://doi.org/10.1093/nar/gkae1063.
63. Harrow J, Drysdale R, Smith A, Repo S, Lanfear J, Blomberg N, 2021. ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics, 37(16): 2506–2511. https://doi.org/10.1093/bioinformatics/btab481.
64. Bennett BC, Balick MJ, 2014. Does the name really matter? The importance of botanical nomenclature and plant taxonomy in biomedical research. Journal of Ethnopharmacology, 152: 387–392. https://doi.org/https://doi.org/10.1016/j.jep.2013.11.042.
65. Rees T, 2014. Taxamatch, an algorithm for near ('fuzzy') matching of scientific names in taxonomic databases. PLOS ONE, 9(9): e107510. https://doi.org/10.1371/journal.pone.0107510.
66. Kindt R, 2020. WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. Applications in Plant Sciences, 8: e11388. https://doi.org/https://doi.org/10.1002/aps3.11388
67. Richard-Bollans A, Allkin B, Civita F, Patmore K, Govaerts R, 2026. Storing accepted scientific names alone can lead to misinterpretation of botanical data. Scientific Reports, 16: 22651. https://doi.org/10.1038/s41598-026-47142-0.
68. Boyle B, Hopkins N, Lu Z, Raygoza Garay JA, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, McKay SJ, Lowry S, 2013. The taxonomic name resolution service: an online tool for automated standardisation of plant names. BMC Bioinformatics, 14(1):16. https://doi.org/10.1186/1471-2105-14-16.
69. Ikerionwu C, Ugwuishiwu C, Okpala I, James I, Okoronkwo M, Nnadi CO, Orji U, Ebem D, Ike A, 2022. Application of machine and deep learning algorithms in optical microscopic detection of Plasmodium: A malaria diagnostic tool for the future. Photodiagnosis and Photodynamic Therapy, 40:103198. https://doi.org/10.1016/j.pdpdt.2022.103198.
70. Salifu D, Chepkemoi L, Ibrahim EA, Nkoba K, Tonnang HE, 2025. Data augmentation and machine learning algorithms for multi-class imbalanced morphometric data of stingless bees. Heliyon, 11(3): e42214. https://doi.org/10.1016/j.heliyon.2025.e42214.
71. Wang X, Hu Y, Zhou X, Li S, 2022. Network pharmacology and traditional medicine: Setting the new standards by combining in silico and experimental work. Frontiers in Pharmacology, 13: 1002537. https://doi.org/10.3389/fphar.2022.1002537.
72. Meyer KG, Bravo-Altamirano K, Herrick J, Loy BA, Yao C, Nugent B, Buchan Z, Daeuble JF, Heemstra R, Jones DM, Wilmot J, 2021. Discovery of florylpicoxamid, a mimic of a macrocyclic natural product. Bioorganic and Medicinal Chemistry, 50: 116455. https://doi.org/10.1016/j.bmc.2021.116455.
73. Apeh IS, Ayoka TO, Nnadi CO, Obonga WO, 2025. Modelling the quantitative structure–activity relationships of 1, 2, 4-triazolo [1, 5-a] pyrimidin-7-amine analogues in the inhibition of Plasmodium falciparum. Engineering Proceedings, 87(1): 52. https://doi.org/10.3390/engproc2025087052.
74. Ayoka TO, Uchegbu UJ, Alabuike CC, Nnadi CO, 2024. Efficient classification and regression models for the QSAR of chloroquine analogues against chloroquine-

sensitive and chloroquine-resistant Plasmodium falciparum. Letters in Applied NanoBioScience, 13(2): 90. https://doi.org/10.33263/LIANBS132.090.
75. Umoette KS, Nnadi CO, Obonga WO, 2023. A robust regression-based modelling to predict antiplasmodial activity of thiazolyl–pyrimidine hybrid derivatives against Plasmodium falciparum. Chemistry Proceedings, 14(1): 52. https://doi.org/10.3390/ecsoc-27-16167.
76. Yang Y, Zhang H, Wang F, Qin Y, Yang Z, Zhai D, Zhou L, Li P, 2026. A deep learning-based generative model for NIR spectral data augmentation and herbal medicine-food homologous classification. Infrared Physics and Technology, 153: 106341. https://doi.org/10.1016/j.infrared.2025.106341.
77. Queiroz LP, Rebello CM, Costa EA, Santana VV, Rodrigues BCL, Rodrigues AE, Ribeiro AM, Nogueira IBR, 2023. Transfer Learning Approach to Develop Natural Molecules with Specific Flavour Requirements. Industrial and Engineering Chemistry Research, 62: 9062–9076. https://doi.org/10.1021/acs.iecr.3c00722.
78. Du Z, Wang Z, Chen Y, Li B, Wan X, Ren T, Chen H, Liu L, Jin Q, Zhang Y, Zhang Y, 2026. KAN-PROSPECT: a Kolmogorov–Arnold Networks–integrated framework for predicting the effects and adverse reactions of natural products via transfer learning. Scientific Reports, 16: 27045. https://doi.org/10.1038/s41598-026-57761-2.
79. Zhang C, Zhai Y, Gong Z, Duan H, She Y, Yang Y, Su A, 2024. Transfer learning across different chemical domains: Virtual screening of organic materials with deep learning models pretrained on small molecule and chemical reaction data. Journal of Cheminformatics, 16: 89. https://doi.org/10.1186/s13321-024-00886-1.
80. Yan J, Zhu J, Yang Y, Liu Q, Zhang K, Zhang Z, Liu X, Zhang B, Gao K, Xiao J, Chen E, 2025. BioMiner: A multi-modal system for automated mining of protein-ligand bioactivity data from literature. BioRxiv, https://doi.org/10.1101/2025.04.22.648951.
81. Nwizu CR, Nnadi CO, Igboasoiyi AC, 2026. Structure-based virtual screening identifies ritobegron and resveratrol as dual inhibitors of NMDA and PTP-1B for Alzheimer's disease therapy. Tropical Journal of Phytochemistry and Pharmaceutical Sciences, 5(2):468-4674. https://doi.org/10.26538/tjpps/v5i2.4.
82. Lin W, Wang P, Zhang Y, Lu W, Yang M, Zhu Z, Yin X, Lv J, Zheng Z, Mei S, Huang Y, 2025. AI-driven transfer learning and generative model (TransGenGRU) enables the drug discovery of novel natural guaianolide sesquiterpene derivatives as potent NLRP3 inhibitors. Journal of Medicinal Chemistry, 68: 21534–21559. https://doi.org/10.1021/acs.jmedchem.5c01663.
83. Choi J, Nam G, Choi J, Jung Y, 2025. A perspective on foundation models in chemistry. JACS Au, 5(4): 1499–1518. https://doi.org/10.1021/jacsau.4c01160.
84. Aljanabi QA, Huang Z, Fan Z, Wu S, Liu S, Chen M, Chen H, Zhang Y, Xu J, Deng L, 2026. TwistDAN: Twisted domain adversarial network for synthetic accessibility assessment. Journal of Chemical Information and Modelling, 66: 2791–2806. https://doi.org/10.1021/acs.jcim.5c02935.
85. Cournia Z, Allen B, Sherman W, 2017. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. Journal of Chemical Information and Modelling, 57(12): 2911-2937. https://doi.org/10.1021/acs.jcim.7b00564.
86. De Ruiter A, Oostenbrink C, 2011. Free energy calculations of protein-ligand interactions. Current Opinion in Chemical Biology, 15(4): 547-552. https://doi.org/10.1016/j.cbpa.2011.05.021.
87. Wang E, Sun H, Wang J, Wang Z, Liu H, Zhang JZ, Hou T, 2019. End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. Chemical Reviews, 119(16): 9478-508. https://doi.org/10.1021/acs.chemrev.9b00055.
88. Akkus E, Tayfuroglu O, Yildiz M, Kocak A, 2022. Accurate binding free energy method from end-state MD simulations. Journal of Chemical Information and Modelling, 62: 4095–4106. https://doi.org/10.1021/acs.jcim.2c00601.
89. Ennab M, Mcheick H, 2024. Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11: 1444763. https://doi.org/10.3389/frobt.2024.1444763.
90. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, Menegaz G, 2025. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1): 2400304. https://doi.org/10.1002/aisy.202400304.
91. Ezema IF, Akwu VC, Didigwu OK, Ogbonna JE, Ugwu AC, Nnadi CO, 2023. Exploring different drug targets responsible for the inhibitory activity of N, N'-substituted diamine derivatives in Leishmania. Engineering Proceedings, 56(1): 178. https://doi.org/10.3390/ASEC2023-16264.
92. Takefuji, Y, 2025. Reevaluating feature importance in machine learning: concerns regarding SHAP interpretations in the context of the EU Artificial Intelligence Act. Water Research, 280: 123514. https://doi.org/10.1016/j.watres.2025.123514.
93. Huang X, Marques-Silva J, 2024. On the failings of Shapley values for explainability. International Journal of Approximate Reasoning, 171: 109112. https://doi.org/10.1016/j.ijar.2023.109112.
94. Tropsha A, 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics 29. John Wiley & Sons, Ltd: 476–488. https://doi.org/10.1002/minf.201000061.
95. Ertl P, Schuffenhauer A, 2009. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics, 1: 8. https://doi.org/10.1186/1758-2946-1-8.
96. Jalbani A, Weerawarna R, Al-Zubaidi K, 2025. Enhancing data provenance in AI with blockchain technology: A comprehensive quality model. CSI Transactions on ICT,

13: 213–224. https://doi.org/10.1007/s40012-025-00419-7.

97. Muthuraj R, Chandrasekaran J, 2026. Nature meets machine: The AI renaissance in natural product drug discovery. Natural Products and Bioprospecting, 16: 37. https://doi.org/10.1007/s13659-025-00589-6.

98. Aledhari M, Razzak R, Parizi RM, Saeed F, 2020. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Access, 8: 140699–140725. https://doi.org/10.1109/ACCESS.2020.3013541.