Int. J. Sci. R. Tech., 2026 3(8)
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

Original Article

Bridging The Data Divide: Standardization And Interoperability In Natural Product Databases & Artificial Intelligence In Drug Discovery

Hussain Asif Vindhani*, Satyam R. Dubey, Mohit Bulani

Department of Pharmaceutical Chemistry & Pharmacognosy, Principal K. M. Kundnani College of Pharmacy, Mumbai, India

ABSTRACT

Natural products (NPs) have traditionally been the primary source for modern pharmacotherapy. The recent integration of artificial intelligence (AI) and machine learning (ML) with molecular language processing represents a revolutionary opportunity in accelerating the discovery of drug leads from natural products, leveraging complex omics data and vast chemical spaces. But how does one resolve an inherent contradiction: natural product chemistry remains an acutely underserved field regarding available standardized, comprehensive, and machine-readable databases that are optimal for AI model training? Modern open-access natural product databases, such as The Natural Products Atlas, are designed to provide structure searchable, microbial metabolite data aligned with the concept of FAIR (Findability, Accessibility, Interoperability, and Reusability). Likewise, natural product databases like SuperNatural 3.0 aggregate comprehensive compound lists with bioactivity metadata. But in the broad domain of natural product chemistry, this openness is limited. Decades of accumulated knowledge in plant-based phytochemistry remains trapped in unstructured text data, scanned legacy journals, flat 2D image formats, or non-standardized institutional repositories with non-uniform molecular representation (lack of defined canonical SMILES strings, InChIs, or SELFIES) or stereochemical completeness. Simultaneously, new generations of AI-informed databases now exist, focusing on generative machine-learning datasets expanding to include up to 67 million NP-like molecules, showcasing the power of AI in radically increasing the searchable chemical space. Yet, the generative database also highlights the urgent need to develop standardized data pipelines to avoid hallucinations of synthesis algorithms and chemical structure misannotations. This review provides a holistic survey of the current state of natural product database standards, details key standardization bottlenecks (stereochemical loss due to data misalignment, nomenclature discrepancy, and metadata gaps), and presents detailed case studies of AI failure and breakthroughs. An integrated cheminformatics roadmap for NP data harmonization is suggested here to provide the prerequisite infrastructure for reliable and scalable AI-driven natural product drug discovery.

Keywords: Natural Products, Artificial Intelligence, Machine Learning, Cheminformatics, Database Standardization, Generative AI, Stereochemistry, Drug Discovery.

INTRODUCTION

Plant, microbial, marine invertebrate, and fungal-derived isolated natural products (NPs) have served as the cornerstone of global pharmaceutical therapy for centuries. From traditional botanical medicine to isolated secondary metabolites like morphine, paclitaxel, artemisinin, and penicillin, the chemical architectures of natural products offer unparalleled structural diversity, evolutionally optimized target selectivity, and a therapeutic breadth unlike synthetic small-molecule libraries developed by combinatorial chemistry methods occupying a considerably less complex and planar chemical space [1, 2]. Typically, secondary metabolites exhibit higher fractions of sp3-hybridized carbon atoms, feature complex fused ring systems, exhibit high degrees of chirality (proliferated through complex hydroxylation, methylation, and glycosyl conjugations), and constrain their natural behaviors through macrocyclic rings [3].

Artificial Intelligence (AI) and Deep Learning (DL) integration in recent years has precipitated a paradigm shift in the field of pharmacognosy and

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

pharmaceutical chemistry. Graph neural networks (GNNs), transformer-based molecular language models, and autoencoders are increasingly being utilized throughout the entire natural product drug discovery pipeline [1]. Algorithms are employed to perform high throughput metabolic deconvolution of liquid chromatography-mass spectrometry (LC-MS) data, predict biological target interactions, forecast pharmacokinetic and toxicity properties (ADMET), and perform de novo molecule generation [2, 3].

But a paradox exists at the core of AI-driven natural product research: While AI models necessitate large, highly structured, and clean training data to deliver accurate physical and biological predictions, the natural product chemistry domain remains one of the most data-fragmented, unstandardized, and non-harmonized spaces in chemical research [4, 5]. By their very algorithm design, algorithms are innately limited by the quality of their training dataset (analogous to the computational truth of 'Garbage In, Garbage Out' or GIGO). Though accumulating knowledge in discovery-driven phytochemistry boasts several decades, the vast vast majority remains trapped within unstructured text documents, scanned legacy journals, flat 2D image formats, or non-interoperable institutional data repositories [4]. When computational chemists attempt to aggregate these multi-source datasets for training a predictive model, the datasets themselves create significant discrepancies in terms of structural formatting, missing stereochemical information, inconsistent taxonomical nomenclatures, and entirely absent standardized metadata regarding compound extraction, isolation, and biological activities [5].

2. The Current Landscape of Natural Product Databases

To better understand database fragmentation in this field, it is crucial to critically discuss the primary open access and commercial natural product databases currently utilized in computational research. Table 1 summarizes important features, structural representations, and FAIR-alignment status of key natural product databases.

Database Name	Coverage Scope	Number of Compounds	Structural Formatting	AI Standardization Challenges
The Natural Products Atlas [4]	Microbial NPs	~43,000	SMILES, InChI, Molfile	Restricted to microbial biospecies; excludes plant/marine metabolome.
SuperNatural 3.0 [5]	Plants, Fungi, Marine	>417,000	3D Mol2,SMILES, InChIKey	Inconsistent stereochemical completeness; algorithmic aggregation artifacts.
COCONUT (v2.0)	Open NP Aggregator	>400,000	SMILES, InChI, SELFIES	Constant cleanup efforts removing duplicate structures from legacy data feeds.
DNP (Dictionary of NPs)	Comprehensive General	>300,000	Proprietary / SDF	Paywall; non-machine readable

				for open access AI pipelines.
Generative NP-Like 67M [3]	In silico NP-like	67,000,000	SMILES, SELFIES	In silico generation; void of biological/physical feasibility validation.

Table 1: Comparative Survey of Leading Natural Product Databases & AI Readiness.

2.1 FAIR Principles in Pharmacognosy

Findable, Accessible, Interoperable, Reusable - these are the data principles mandated for modern informatics [4]. The Natural Products Atlas [4] stands as a prime example of a FAIR-aligned architecture in microbial pharmacognosy. Through strict peer-reviewed isolation citation requirements, data validation of skeletal structures and assignments, and persistent unique chemical identifiers (NPAtlas IDs), data fidelity is high. Yet microbial metabolites account for only a fraction of the total therapeutic space in nature. Phytochemicals derived from plants (e.g., complex alkaloids, polyphenols, triterpenoid saponins) remain fragmented and dispersed within heterogeneous databases (e.g., TCMSP, IMPPAT, KNApSAcK), many of which lack standardized structural validation protocols or continuous programmatic API accessibility [1, 5].

2.2 Large-Scale Aggregators: SuperNatural 3.0 and COCONUT

To close the niche gap, efforts like SuperNatural 3.0 [5] and the Collection of Open Natural Products (COCONUT) have aggregated hundreds of thousands of structures. SuperNatural 3.0 cross-references the chemical structure of compounds with predicted toxicity profiles, biological pathways, and vendor availability. But large aggregators present a dire paradox of their own: when legacy data is ingested from an uncurated primary literature source, introduced errors in the original published article (incorrect double-bond geometries, missing stereocenter uncertainties, and glycosidic linkage misassignments) are systematically copied and reinforced through the entire computational ecosystem [5].

3. The Standardization Crisis in AI-Driven Pharmacogonosy

The use of machine learning algorithms in drug discovery involves the translation of 3D chemical structures into machine-processible mathematical representations (binary molecular fingerprints, molecular graphs, or 1D string notations). When chemical data is unstandardized in this process, AI models “learn” erroneous patterns that greatly degrade predictive performance.

3.1 Stereochemical Loss and Chiral Ambiguity

Secondary metabolites are inherently chiral compounds. The specific spatial orientation of the molecule across the chiral centers defines their thermodynamic binding affinity at protein receptors. The enantiomers of carvedilol and limonin, for instance, exhibit drastically different therapeutic properties. Yet legacy data conversion pipelines often flatten 3D structures into non-stereospecific 2D representations. Exported as standard Simplified Molecular-Input Line-Entry System (SMILES) strings without explicit chiral flags (R/S or @/@@ identifiers), all stereoisomers are collapsed into one topological chemical entity [4, 5].

3.2 Tautomerism, Protonation States, and Salts

Phytochemicals such flavonoids and alkaloids exhibit dynamic tautomerization and pH-dependent protonation processes. A database recording a molecule in its neutral (unprotonated) state while another records a protonated salt will cause machine learning models to treat them as completely separate chemical entities, artificially inflating the size of the dataset and distorting feature extraction algorithms [2].

3.3 String Representation Vulnerabilities: SMILES vs. SELFIES

Molecular language models (i.e., GPT-based transformer or Variational Autoencoders) encode chemical structures as sequential string language. The standard SMILES notation operates on a strict format of ring closure and valence constraints. When generative models learn to generate novel NP-like structure using SMILES, over 30-50% of the output strings are syntactically invalid (unclosed rings or pentavalent carbon atoms) [3]. To resolve this limitation, the cheminformatics community created a solution known as SELFIES (Self-Referencing Embedded Strings). SELFIES provides 100% mathematical guarantees by translating strings into localized structural rules. The SELFIES translation from legacy SMILES is now a primary prerequisite for scaling generatively NP discovery pipelines [3].

4. Real-World Case Studies: AI Failures & Solutions

To clearly demonstrate the necessity of standardization in a practical context, we will present two seminal case studies showing how cheminformatics corrections directly influence AI performance in natural product discovery.

Case Study 1: The Stereochemistry Bottleneck & NPstereo Language Model

Context: Machine learning models trained on flat, non-curated natural product databases often fail to predict accurate ligand binding affinities, as traditional published literature typically lacks explicit stereochemical flags (R/S configuration).

Problem: When predicting a multi-stereocenter diterpene or alkaloid, the flat SMILES representation obligates the neural network to infer 3D conformations, leading to severe false positives in virtual screening.

Solution & Impact: Researchers created an open source tool called NPstereo, a transformer-based language model trained exclusively on the open-access COCONUT database. NPstereo functions as an automated post-hoc standardization algorithm. Based on local chemical contexts, NPstereo predicts and flags missing chiral annotations at an >80% accuracy rate across key metabolite classes (alkaloids, terpenoids, and steroids). This case study demonstrates that AI must first be applied to clean and standardize historical data before it can be used to discover novel drug leads with consistency.

Case Study 2: Syntactic Hallucinations in 67 Million Generative NP-Like Compounds

Context: Generative molecular language models were utilized by Tay et al. (2023) [3] to construct an immense virtual database of 67 million natural product-like molecules, expanding the searchable chemical space beyond what is known to exist in Nature.

Problem: In initial training with standard canonical SMILES as input, the model produced a high percentage of syntactically invalid output, 'hallucinating' chemically impossible architectures (e.g., hypervalent oxygen atoms, broken aromaticity, and invalid stereocenters).

Solution & Impact: By translating the underlying structure pipeline to robust, cheminformatics-validated representations (including SELFIES and strict valence-checked InChIKeys), the research group removed syntactically erroneous compounds. This enabled deep learning models to efficiently screen the 67 million virtual compounds and identify novel bio-inspired lead compounds targeting multi-drug resistant bacterial pathogens.

Standardized 5-Part Botanical & Phytochemical Analysis Framework

To create AI databases that capture the complete biological and chemical detail of medicinal botanicals, future database architectures must standardize data ingestion across five necessary pharmacognostical components. Below is the standardized framework required for digital monograph integration:

5.1 Botanical & Ethnopharmacological Profile

Taxonomical integrity is paramount. Databases must combine validated Latin binomials from World Flora Online (WFO) with family designation, geographical site (latitude/longitude coordinates), and structured ethnobotanical keywords (Traditional Ayurvedic, TCM, Unani uses, etc.).

5.2 Phytochemical Composition & Chemical Classes

Molecules must be classified into hierarchically secondary metabolite (e.g., monoterpenoid indole alkaloids, prenylated flavonoids, proanthocyanidins) categories with standardized Chemical Entities of Biological Interest (ChEBI) ontology terms.

5.3 Extraction, Isolation & Processing Metadata

AI models anticipating compound yield or bioactivity potential must encode extraction metadata: solvent polarity (i.e., supercritical CO2, ethanol, water), extraction technique (i.e., Soxhlet, ultrasound-assisted extraction), and chromatographic separation methods (HPLC-PDA and LC-MS/MS).

5.4 Pharmacological Activity & Molecular Target Mechanisms

Bioactivity records must be more than qualitative description (i.e., 'anti-inflammatory') and include quantitative, structured endpoint parameters such as IC50, EC50, Ki, protein target UniProt ID, and biological signal transduction pathways affected (e.g., NF-kB inhibition, MAPK phosphorylation).

5.5 Standardization, Biomarkers & Quality Control

Monographs must include quantified reference biomarkers (e.g., curcuminoids in Curcuma longa; sennosides in Cassia angustifolia), potential adulterants, heavy metal contamination limits and standardized analytical HPTLC/HPLC fingerprint profiles.

DISCUSSION & FUTURE ROADMAP FOR AI HARMONIZATION

To unlock the full potential of artificial intelligence in natural product drug discovery, an urgent paradigm shift in how phytochemical data is curated, published, and shared across the academic and industry research community is required. No amount of algorithmic intelligence can compensate for fundamentally flawed input data. We suggest a three-tiered roadmap for global natural product database harmonization:

Journal & Publisher Mandatory Reporting Standards: Academic research journals accepting novel isolation or pharmacognostical research papers must mandatorily require submission of fully validated chemical structure data files (FAIR compliant SDF or SELFIES) with confirmed 3D stereochemistry, raw mass spectra data (MS/MS), and NMR FID files in addition to classical manuscript text [1, 4].

Automated ML Data Cleaning Pipelines: Implementation of open-source cheminformatics validation pipelines (i.e., RDKit, Open Babel, CDK) that automate salt removal, tautomerization, protonation state assignment at physiological (pH 7.4), and flagging of ambiguous chiral centers upon data entry [5].

Integration of Multi-Omics and Ethnopharmacology: Linking standardized chemical structures with genomic biosynthetic gene cluster (BGC) data from antiSMASH and metabolomic networks from GNPS (Global Natural Products Social Molecular Networking) to form multi-layered knowledge graphs for AI reasoning [1, 2].

CONCLUSION

Applying artificial intelligence to natural product chemistry represents one of the most exciting edges of contemporary pharmaceutical discovery. But as demonstrated throughout this review, the field bottleneck has shifted from algorithms to data quality and standardization. By overcoming stereochemical ambiguity, embracing robust string notations like SELFIES, consolidating databases under FAIR principles, and prescribing comprehensive phytochemical metadata, the global scientific community can bridge the data divide. The standardization of 'Nature's Pharmacy' will enable next-generation AI workflows to unlock novel therapeutics for complex human disease with unprecedented accuracy and speed.

REFERENCES

1. Geetha, J., Ranganathan, N., Gulothungan, G., & Chopra, H. (2026). Integrating Artificial Intelligence and Machine Learning in Natural Product Discovery: From Omics Data to Drug Design. Natural Resources & Health, https://doi.org/10.53365/nrfhh/217344
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/ph19220301

3. Tay, D. W. P., Yeo, N. Z. X., Adaikkappan, K., Lim, Y. H., & Ang, S. J. (2023). 67 million natural product-like compound database generated via molecular language processing. Scientific Data, 10, 220. https://doi.org/10.1038/s41597-023-02207-x
4. Santen, J. A. V., Jacob, G., Singh, A., Aniebok, V., Balunas, M. J., Bunsko, D., et al. (2019). The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Central Science, 5(11), 1824–1833. https://doi.org/10.1021/acscentsci.9b00806
5. Gallo, K., Kemmler, E., Goede, A., Becker, F., Dunkel, M., Preissner, R., & Banerjee, P. (2022). SuperNatural 3.0—a database of natural products and natural product-based derivatives. Nucleic Acids Research, 50(D1), D654–D660. https://doi.org/10.1093/nar/gkac1088

HOW TO CITE: Hussain Asif Vindhani*, Satyam R. Dubey, Mohit Bulani, Bridging The Data Divide: Standardization And Interoperability In Natural Product Databases & Artificial Intelligence In Drug Discovery, Int. J. Sci. R. Tech., 2026, 3 (8), 648-653. https://doi.org/10.5281/zenodo.21980457