JBP
Journal of Biochemicals and Phytomedicine
eISSN: 2958-8561

# Machine Learning-Guided Ethnobotanical Mining of Traditional Medicine Texts to Predict and Experimentally Validate Antidiabetic Plant Candidates

**Vaishnavee Vijay Rathod [1], S. Uma [2], D. Sudhagar [3], M. Vetri Selvan [4], Preshni Shrivastava [5], R. Senthamil Selvan [6*]**

[1] Department of Information and Communication Technology, Adani University, Ahmedabad, India
[2] Department of ECE, SIMATS Engineering, Saveetha School of Engineering, Saveetha University, Chennai, India
[3] Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India
[4] Department of Artificial Intelligence and Data Science, Panimalar Engineering College, Poonamallee, Chennai, India
[5] Department of Management, IMM New Delhi, India
[6] Department of ECE, Saveetha School of Engineering, Saveetha University, Chennai, India

ARTICLE INFO

***Article Type:***
Research

***Article History:***
Received: 14 Aug 2026
Revised: 30 Sept 2026
Accepted: 15 Oct 2026
Available online: 11 Nov 2026

***Keywords:***
BOTANI-PHYTO-NET,
FUZZ-GLYCO-INSIGHT Optimization,
Ethnobotanical knowledge mining,
Fuzzy machine learning,
Antidiabetic plant discovery

* ***Corresponding author:***
***E-mail:***
selvasenthamil2614@gmail.com

ABSTRACT

**Introduction:** The traditional medicinal knowledge is one of the important resources which have not been fully explored for identification of therapeutic agents from the plant sources for diabetes mellitus. The search for reliable ethnobotanical evidence and identification of potential antidiabetic medicinal plants from the traditional medicine literature is, however, difficult because of the lack of homogeneity in traditional medicine sources, the lack of complete information about the plant's phytochemicals, and the inadequacies of the integration of computational predictions and experimental evidence.

**Methods:** To tackle these issues, this study suggests an integrated machine learning guided ethnobotanical discovery framework derived from publicly available medicinal plant databases, natural product repositories, scientific literature, and biological activity records. This framework consists of two novel algorithms: BOTANI-PHYTO-NET (BPN), and FUZZ-GLYCO-INSIGHT Optimization (FGIO). The main data sources are the Indian Medicinal Plants, Phytochemistry and Therapeutics database (IMPPAT 3.0) which includes 4,154 medicinal plants and 18,314 phytochemicals, Medicinal Plant Database from the Botanical Survey of India, LOTUS Natural Products Database, and bioactivity data from ChEMBL and PubChem BioAssay. Ethnobotanical knowledge extraction and identification of antidiabetic activity will be supplemented by scientific publications from PubMed and PubMed Central. BOTANI-PHYTO-NET uses natural language processing, semantic entity recognition, and network-based knowledge representation to extract medicinal plants, plant parts, traditional therapeutic uses, phytochemical constituents, and plant–compound–disease relationships. The information is then structured and represented as a DBSK within a phytomedicinal knowledge network. FUZZ-GLYCO-INSIGHT Optimization makes use of this network to obtain the estimation of antidiabetic potential and prioritize candidate plants by incorporating heterogeneous ethnobotanical, phytochemical and biological data into fuzzy inference and multi-objective optimization. The best candidates will be subjected to experimental assessment by in-vitro α-amylase and α-glucosidase inhibition assays, phytochemical characterization and cytotoxicity evaluation. The predictive framework will then be refined and validated independently with experimentally derived inhibition and toxicity measurements.

**Results:** The proposed approach provides an integrated computational-experimental pipeline from the traditional medicinal knowledge to the botanical network intelligence, fuzzy-based optimization and biological validation.

**Conclusion:** The framework is designed to enhance the systematic identification and selection of plant-derived sources of antidiabetic candidates with scientific evidence and to offer a reproducible approach for using computer-aided ethnopharmacological studies.

***Please cite this paper as:***
Rathod VV, Uma S, Sudhagar D, Vetri Selvan M, Shrivastava P, Senthamil Selvan R. Machine learning-guided ethnobotanical mining of traditional medicine texts to predict and experimentally validate antidiabetic plant candidates. Journal of Biochemicals and Phytomedicine. 2026; 5(2): 1-20. doi: 10.34172/jbp.2026.11.

## Introduction

Diabetes mellitus is a chronic metabolic disease in which the level of glucose in the blood is raised due to failure of insulin secretion, insulin resistance or a combination of both. It is currently becoming more prevalent in the world, and there is a great need for the development of effective, cost-effective and biologically safe therapeutic interventions. Traditional anti-diabetic drugs are playing a significant role in the management of diabetes, but their drawbacks such as side effects, costs, and different therapeutic effects have encouraged the search for potential alternative sources of bioactive compounds. Medicinal plants have been the focus of many research studies due to the presence of various phytochemicals such as flavonoids, alkaloids, terpenoids, phenolic compounds and glycosides which could modulate glucose metabolism and biological pathways. There are several reports of diabetes associated medicinal plants from traditional medicinal systems, but a lot of this knowledge is limited to ethnobotanical records, traditional medicine literature and scientific papers. The systematic conversion of this information into scientifically structured evidences is therefore one of the important challenges in the contemporary medicinal plant research.

The traditional medicine literature offers valuable information about the species of medicinal plants, parts of plants, preparation techniques, therapeutic uses and reported biological properties. It is, however, not easy to extract meaningful information from such diverse sources, due to the fact that the term "botany" is used in various ways to describe the plants, the names of plants vary from one region to another, some botanical descriptions overlap, and the therapeutic records are incomplete. Traditional methods of keyword search may not be adequate to resolve complex relationships between plants, phytochemical components and disease-related biological activities. To overcome these limitations, opportunities exist in the field of natural language processing (NLP), machine learning and knowledge network construction for automatically identifying botanical entities, semantic relationships and structuring the dispersed medicinal knowledge into knowledge representations. These computational methods may help identify plants that may be useful in medicine and help to shorten the time and labor of conducting a broad survey of ethnobotanical data.

Special medicinal plant and chemical repositories are available, giving another basis for computational phytomedicine research. Complementary information on medicinal plants, phytochemical constituents, chemical structures, molecular targets and biological assays measurements are available in databases like IMPPAT 3.0, LOTUS, ChEMBL and PubChem. The scientific publications from the PubMed and PubMed Central databases also offer text evidence for medicinal uses in traditional medicine and experimentally tested biological activities. However, the structures of these resources and the terminology, quality of evidence, and biological validation vary. They need to be effectively integrated, including botanical name standardization, chemical identifier mapping, duplicate removal, data normalization and evidence assessment. Moreover, the discovery of a phytochemical or molecular target of a medicinal plant does not automatically prove its antidiabetic activity. Therefore, computational discovery frameworks need to separate between claims and predicted biological relationships and activities which have already been confirmed by experiment.

This research presents a combined machine learning and ethnobotanical discovery model, which combines two models: BOTANI-PHYTO-NET (BPN) and FUZZ-GLYCO-INSIGHT Optimization (FGIO) to address these limitations. The BOTANI-PHYTO-NET algorithm aims at automatic extraction and structured representation of ethnobotanical data from the scientific literature and traditional medicine texts. It integrates text preprocessing, identification of botanical entities, extraction of semantic relationships, and representation of the knowledge in a network to create links between medicinal plants, plant parts, phytochemical components, traditional uses and diabetic-related biological targets. The phytomedicinal knowledge network created facilitates the identification and exploration of plant candidates for developing antidiabetic drugs. FUZZ-GLYCO-INSIGHT Optimization subsequently combines the heterogeneous ethnobotanical, phytochemical and biological evidence by using fuzzy inference and multi-objective optimization. The aim is to handle uncertainty, assess the degree of reliability of the evidence, foresee possible antidiabetic activity and prioritize the most promising medicinal plants for research. The proposed algorithms define a coordinated computational method to medicinal plant candidate identification by assigning different responsibilities to knowledge discovery and predictive optimization.

The proposed framework is based on datasets that are publicly available, such as IMPPAT 3.0, LOTUS, ChEMBL, and PubChem, as well as relevant scientific publications that are accessed via PubMed/PubMed Central. These resources are used to identify the plants, characterize the phytochemicals, connect them with targets, and integrate biological evidence. Computationally prioritized plant candidates will be subjected to experimental screening through the in-vitro α-amylase and α-glucosidase inhibition assays, and then phytochemical characterization and cytotoxicity testing. For the purpose of the evaluation

of biological activities selected extracts, quantitative evidence will be provided by percentage inhibition and IC50 measurements. The validation of the reliability of the candidate's ranking will be done by comparing the experimental results with the computed ones, and the results will be used to improve the predictive framework. The integration of computation with experiment is critical here to differentiate between good computer predictions and those supported by biological evidence in the context of antidiabetic activity.

This research aimed to answer the following questions and to find the following results:

(1) To develop BOTANI-PHYTO-NET (BPN) for automated ethnobotanical text mining, medicinal plant identification, construction of real world datasets such as IMPPAT 3.0, LOTUS, ChEMBL and PubChem and extracting relationships between phytochemicals.
(2) To propose a FUZZ-GLYCO-INSIGHT Optimization (FGIO) to merge the integrated outputs of heterogeneous botanical, phytochemical and biological evidences with the fuzzy logic and multi-objective optimization to predict the antidiabetic potential and to prioritize promising medicinal plant candidates.
(3) To conduct an in-vitro α-amylase and α-glucosidase inhibition assay and cytotoxicity test for selected plants with high computational priority and compare the results with the computational results to check the reliability of the proposed methodology.

This paper is organized as follows: the Related Works section summarizes the current literature pertaining to ethnobotanical knowledge mining, medicinal plant databases, machine learning for phytochemical discovery and computational antidiabetes prediction. The Materials and Methods section describes the proposed research methodology, datasets and preprocessing techniques along with the overall framework, the BOTANI-PHYTO-NET algorithm, the FUZZ-GLYCO-INSIGHT Optimization algorithm with its mathematical formulation, and the experimental validation procedures. The computational and experimental results, followed by a comparative analysis of the work, are provided in the Results and Discussion section, and future research directions are discussed in the Conclusion.

## Related Works

However, incorporation of machine learning, natural language processing, knowledge networks of medicinal plants, phytochemical analysis, and experimental validation has been the subject of recent research to search for plant-derived therapeutic candidates against diabetes mellitus. The machine learning models were developed by Jamir and Hariprasad (2024) for predicting potential plant secondary metabolite that inhibit α-glucosidase, linked with type 2 diabetes, and biological activity was tested through in-vitro validation of these models. Their study showed that molecular fingerprints and machine learning classifiers can be invaluable tools for screening plants for the possibility of inhibiting enzymes. Computational prediction is an important component in the development of reliable discovery frameworks for medicinal plants where it can be combined with laboratory investigation of plants. In order to extract the entities and relations in Traditional Chinese Medicine (TCM) document-level texts, Xu et al. (2024) presented a joint entity and relation extraction approach. They were able to identify several entities and semantic relationships in medical texts, which highlights the capability of NLP in extracting structured information from traditional medicine literature and illustrates the potential of the technology for such tasks. This technique is applied to the plant name, plant parts, medicinal uses and phytochemical relationship of medicinal plants in ethnobotanical documents.

Duan et al. (2024) proposed knowledge graph embedding and residual-like graph neural network learning based herb–target prediction framework, named HTINet2. They explored the relationships between biological targets and herbal medicines using a combination of graph representations and predictive learning. The study shows that knowledge graph-based methods can be used to identify new herb–target associations that have not been known before. Sharafudeen et al. (2024) explored a hybrid network approach for medicinal plant localization and pharmacognostic identification, highlighting the need for botanical identification and structured information on medicinal plants. The authors' efforts give a meaningful perspective on how to combine botanical attributes with computational knowledge representation. Umar et al. (2025) explored the ethnobotanical and the pharmacological mechanism of action of Fibraurea tinctoria in diabetes mellitus using metabolomic analysis, network pharmacology and molecular docking approaches. The investigation they carried out here shows the potential of combining phytochemical characterization and molecular interaction analysis to study potential relationship between the constituents of medicinal plant and the biological targets associated with diabetes.

Sharma and Deb (2024) conducted ethnomedicinal study of diabetes treating plants traditionally used in Nagaland, India and their molecular characterization. They used molecular identification methods with three different DNA barcode markers, namely ITS, matK and rbcL, for the authentication of medicinal plant species. The study emphasizes the need for botanical identification of medicinal plants before

exploring their pharmacological properties. Khalid et al. (2024) did phyto-bioinformatics screening of Ayurvedic medicinal plants that could be used to develop α-glucosidase inhibitors. They included the information on phytochemicals and carried out computational screening and molecular interaction analysis to prioritize compounds for further in vivo studies in the context of diabetes. Andrographis paniculata leaves have been explored for their antidiabetic activity by the computer aided discovery of potential α-amylase and α-glucosidase inhibitors by Swargiary et al. (2025). Their work shows the use of computational screening to discover natural product-based compounds related to carbohydrate-digesting enzymes from plants. These studies underscore the need to consider the use of a combination of botanical identification, phytochemical characterization, and computational prediction in evaluating phytochemicals for their antidiabetic potential.

Zhang and Hao (2024) explored the construction of Traditional Chinese Medicine Knowledge Graph with Large Language Models. They were interested in acquiring medicinal knowledge and structuring the semantic relationships into knowledge representations. The results of the study illustrate the promise of LMs for enhancing information extraction and knowledge organization in traditional medicine studies. Meijer et al. (2025) presented the use of AI, multimodal data integration and knowledge graphs in natural product science. They focussed on the interconnection of the data of natural products which are structurally and functionally diverse for systematic compound finding and enhanced knowledge sharing. The developments offer methodological tools for building inter-connected networks of botanical knowledge and phytochemical knowledge, integrating information from traditional medicinal literature and scientific databases.

Duan et al. (2025) introduced a case-based question-answering system of traditional Chinese medicine that combines the traditional wisdom of traditional Chinese medicine with the capabilities of large language models and knowledge graphs. Their approach involved knowledge extraction, knowledge fusion and graph-based information retrieval for accessing the structured information of the medicines. Based on prompt learning and joint information extraction, Li et al. (2024) presented a method for creating a knowledge graph for acupuncture and Tuina. Their research showed that language representation and entity recognition methods can be used to structure medical information in the domain into relationships. To overcome the problem of information retrieval and management of Traditional Chinese Medicine (TCM) prescriptions, Ye et al. (2024) proposed a method of knowledge graph-based information organization and retrieval. Their approach showed that structured relationships can facilitate access to traditional medical knowledge and facilitate information discovery. In summary, the research work of these studies all play a fundamental role in the field of processing traditional medicine information, highlighting the significance of intelligent retrieval, knowledge graph construction, semantic relationship extraction, and NLP in traditional medicine information processing. But, they are mainly interested in knowledge organization of general medicinal plants, not to an integrated prediction and experiment-based prioritization of antidiabetic medicinal plants.

Ethnobotanical studies have also contributed significant information about indigenous use of plants and the benefits that it may offer in the treatment of diabetes. The work of Aradhana et al. (2022) was on the plants which are traditionally used for diabetes treatment by the Santhal tribe people of Mayurbhanj, Odisha, India. They highlighted the need to record and critically evaluate traditional therapeutic information and practices. The studies conducted by Dey and Dutta (2025) include an element of the antidiabetic usage of some herbs in the selected districts of Assam and the determination of Glycemic index (GI) of the commonly used herbs. Their research helped to understand traditional methods of plant based approach to glucose management and the significance of its nutritional and biological properties evaluation. The community based ethnobotanical study conducted by Khajuria et al. (2025) has documented the ethnomedicinal plants used in Diabetes management in Jammu. They concluded that the traditional knowledge of the plants and their uses were still very much relevant and the systematic documentation of medicinal plants and their uses and applications and local practices were still highly needed.

To build and apply a Traditional Chinese Medicine knowledge graph using Large Language Models, Zhang et al. (2026) explored this approach. They also showed the feasibility and effectiveness of language model-aided knowledge extraction and structured representation of traditional medicinal knowledge. Gohain et al. (2026) did an ethnobotanical exploration of medicinal plants of Biswanath district Assam, India with the research method of quantitative ethnobotanical approach for antidiabetic plants. They used community based medicinal knowledge documentation with quantitative ethnobotanical analysis and existing pharmacological information. The research emphasised the need to have scientific evidence to assess the potential of the medicinal plants that are commonly reported. The above studies have clearly shown that traditional medicinal knowledge can be useful in the identification of plant candidates, but there is an important direction of research to systematically integrate traditional medicinal

knowledge with automated knowledge extraction, phytochemical analysis, and predictive biological assessment.

Davies et al. (2022) reported a consensus statement of American Diabetes Association and European Association for the Study of Diabetes regarding the management of hyperglycaemia in type 2 diabetes. Their report addressed evidence-based management of diabetes, such as individualised management, lifestyle changes and pharmacologic interventions. The study offers valuable clinical context for further studies of therapeutic resources and assessing their relevance to diabetes care. Modelling and forecasting analysis by Chauhan et al. (2025) analyzed the escalating burden of diabetes and variations among states of India. Their research revealed that diabetes remains a significant public health issue in India, and that there is an urgent need for better prevention and management strategies. The clinical and epidemiological results corroborate the need for investigating medicinal plant resources as sources of bioactive compounds and highlight the importance of biological and clinical evidence in order to evaluate the therapeutic potential of the medicinal plants.

However, previous studies have shown that there are certain areas that need exploration. Ethnobotanical studies mostly focus on traditional uses of plants while NLP-based studies are directed towards information extraction and knowledge organization. In the same way, computational phytochemical analyses and machine learning tend to focus on either predicting compounds, analysing the interactions between compounds or specific biological targets. Combining these research dimensions in the same computational to experimental framework still remains a great research challenge. Especially, the uncertainty over ethnobotanical and biological information that is often heterogeneous needs to be evaluated systematically when prioritizing medicinal plant candidates. The present study aims to tackle these shortcomings by introducing BOTANI-PHYTO-NET (BPN), a novel automated ethnobotanical text mining system for the identification of botanical entities, the extraction of semantic relations and the construction of a network of phytomedicinal knowledge. In addition, an optimization system based on Fuzzy Inference and Multi Objective optimization (FGIO) is introduced to combine botanical, phytochemical and biological data for predicting and prioritizing the potential antidiabetic plants. Lastly, the candidates that were prioritized in the computational process will be assessed for in-vitro α-amylase and α-glucosidase inhibition and for cytotoxicity measurement. The goal of this integration is to create a formal connection between traditional medicine knowledge, computational discovery, predictive optimization and experimental biological validation.

## Materials and Methods

The proposed methodology provides a computational-to-experimental approach for the identification, prediction and experimental evaluation of medicinal plants with potential antidiabetic properties. The framework integrates the databases of medicinal plants in the real world, mining of ethnobotanical literature, phytochemical characterization, construction of knowledge network, fuzzy-based prediction and multi-objective optimization. To achieve different purposes in the proposed framework two complementary algorithms are developed, namely BOTANI-PHYTO-NET (BPN) and FUZZ-GLYCOINSIGHT Optimization (FGIO). BOTANI-PHYTO-NET extracts botanical and therapeutic information from heterogeneous sources and creates a structured phytomedicinal knowledge network, while FUZZ-GLYCO-INSIGHT Optimization (FGIO) evaluates the extracted evidence, predicts the possible antidiabetic property of the medicinal plants, and prioritizes potential candidates for experimental investigation. The methodology is divided into five subsections: dataset collection, data preprocessing, proposed algorithms and experimental validation.

### Dataset Collection and Sources

The proposed framework works on publicly available medicinal plant, natural product, chemical, biological activity and scientific literature datasets, forming a solid ground for the knowledge discovery of ethnobotanical knowledge. The major data source used is the 'Indian Medicinal Plants, Phytochemistry and Therapeutics (IMPPAT 3.0)' database where all the information related to medicinal plant species, botanical families, plant parts, phytochemicals, therapeutic uses and biological targets associated with the plant species are curated. It is reported to have a database with 4154 medicinal plants, 18314 phytochemicals and 1544 therapeutic uses. These records will be used to create the proposed knowledge network, which will have its main botanical and phytochemical content. The LOTUS Natural Products Database provides natural product chemical structures, molecular identifiers, SMILES representations and taxonomic associations to provide additional information. The records help identification of phytochemical constituents and help to establish relationship between natural products and their botanical origin. The ChEMBL database is used to pull out experimentally reported values of activity $IC_{50}$, $K_i$, and $EC_{50}$, assay/description, target information and experimental conditions. These records help evaluate reported biological activity, and develop predictive features for compounds. PubChem BioAssay is a complement to the compound record that includes bioactivity and toxicity data such as compound identifiers, assay

results and reported experimental measurements. These databases are of special interest when it comes to the identification of compounds that are related to α - Biological targets, such as amylase, α-glucosidase, and other diabetes related biological targets. Scientific publications indexed in PubMed and PubMed Central are also gathered to gather the traditional medicinal applications of the plants, methods of preparation, phytochemical constituents reported, and previous research studies on the antidiabetic activity of the plants. The literature component is to help provide context information which may be unavailable in structured medicinal plant databases. The data set is integrated and organized by creating links between botanical identifiers, accepted scientific names, plant parts, phytochemical identifiers, chemical structures, molecular targets, therapeutic annotations and records of experimental activity. The use of botanical synonyms is linked to accepted scientific names for those for which a reliable taxonomic identification was available. Chemical records are linked together by PubChem compound identifiers, InChIKeys and canonical SMILES. Biological activity measurements are kept with the same assay ID, type of activity and units as the original and the annotations that are associated with the target are not changed to preserve experimental provenance. Records from various assay conditions are not necessarily considered to be the same and traditional therapeutic claims are separated from predicted associations from computation and biological activity supported by experiments. Combined with the input of botanical information extraction, the construction of knowledge graph and predictive modelling, it is possible to select the experimental candidates in the subsequent stage.

The official dataset repositories are accessible through their respective websites: IMPPAT 3.0 (https://cb.imsc.res.in/imppat/), LOTUS (https://lotus.naturalproducts. net/), ChEMBL (https://www.ebi.ac.uk/chembl/), PubChem (https://pubchem. ncbi.nlm.nih.gov/), PubMed (https://pubmed.ncbi.nlm.nih.gov/), and PubMed Central (https://pmc.ncbi.nlm.nih.gov/). The downloaded dataset versions and retrieval dates are recorded to maintain data provenance and reproducibility.

## Data Preprocessing and Feature Extraction

The datasets collected are heterogeneous in both their records and their botanical and chemical terminology; some records are duplicated; some measurements are biological and others are using different units; and some attributes are missing. Thus, data preprocessing is done to enhance data consistency and set up reliable relationships among the information sources. The first step in traditional medicine is to clean the texts, segment the sentences, tokenize them, remove stop words, and standardize the botanical terms in them, which are done in traditional medicine texts and scientific publications. The botanical names are standardized by the use of the accepted scientific names, synonyms and regional plant names are identified and mapped to appropriate species when reliable identification of a species is available. This method will minimize the number of duplicate botanical entities and increase the uniformity of the information retrieved from various literature sources. The use of named entity recognition (NER) for medicinal plants, phytochemicals, biological targets and disease-related terms. The extraction of semantic relationship further captures the relationship between plants and plant parts, plant-phytochemical, plant-therapeutic use, and phytochemical-biological target. The term frequency-inverse document frequency (TF-IDF) method is used to represent textual features using the importance of informative terms, based on how often they appear in individual documents and how often they appear in the entire document collection. The TF-IDF representation is obtained as follows:

$$TFIDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

where the inverse document frequency is given by:

$$IDF(t) = \log\left(\frac{N}{1+DF(t)}\right) \quad (2)$$

In this case, t is a textual term, d is a document, N is the total number of documents, and DF(t) is the number of documents which contain the term t. Chemical structures are represented using available canonical SMILES and InChIKey representations and standardized. The molecular descriptors such as molecular weight, hydrogen bond donors, hydrogen bond acceptors, topological polar surface area and lipophilicity are extracted, which represents the phytochemical characteristics. Chemical identifiers and structure based comparison are used to find similar compounds (duplicates). The biological data are not replaced with random numbers, but if there isn't an adequate amount of experimental data, missing records are not entered into the supervised. As evidence that are retained or left as learning tasks. If a valid conversion can be made, measurements of biological activity are converted to units of consistent measurement. For instance, the micromolar $IC_{50}$ values are converted to negative log activity values, as follows:

$$pIC_{50} = -\log_{10}\left(IC_{50} \times 10^{-6}\right) \quad (3)$$

The $IC_{50}$ refers to the concentration of a chemical that causes a 50% reduction in response. $IC_{50}$the concentration of a chemical, prior to conversion to molar, that produces 50% reduction in the response value. The higher the $pIC_{50}$ value the more potent the inhibition is, at equivalent experimental conditions. Data from the various types of assays, organisms, substrates and conditions of assays are not simply added together without appropriate harmonization. Min-max normalization is used to normalize the numerical features into a common numerical range for subsequent computations. The normalization process is represented by:

$$X_{\text{norm}} = \frac{X - X_{min}}{X_{max} - X_{min} + \epsilon} \quad (4)$$

Note that, in this case, X represents the original feature, while $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values of the feature, respectively, and ϵ is a small positive value to avoid division by zero. Depending on the prediction task, the resulting feature representations are split into a training set, a validation set, and an independent test set, such that as much as possible, there are no related items in the training and validation sets that could result in information leakage.

## BOTANI-PHYTO-NET Algorithm

A BOTANI-PHYTO-NET (BPN) algorithm is suggested for automated ethnobotanical text mining, botanical entity identification, getting semantic relationship between entities and construction of the phytomedicinal knowledge network. It aims to create a network of relationships between botanical resources, their phytochemicals and their therapeutic uses from unstructured traditional medicine literature and from heterogeneous medicinal plant databases. Preprocessed textual documents and structured records from IMPPAT 3.0, LOTUS, PubMed and PubMed Central are the initial inputs to the algorithm. Botanical entity recognition refers to the identification of plant species, plant parts and regional synonyms while phytochemical recognition involves the identification of chemical constituents and their related molecular identifiers. Therapeutic entities are diseases associated with diabetes, reported medicinal uses and relevant biological targets. Semantic relationship extraction then creates a relationship between extracted entities based on their context, domain knowledge and extracted patterns. These relationships are modeled heterogeneously as a phy- These relationships are modeled heterogeneously as a phy- A knowledge network to medicinal knowledge, which includes botanical entities, phytochemical constituents, biological targets and therapeutic conditions. The mathematical representation of the overall knowledge network is

$$G = (V, E, R) \quad (5)$$

Where V is a set of nodes, E is set of relationships between nodes, and R is set of relationship categories. Medicinal plants, plant parts, phytochemicals, biological targets and therapeutic conditions are included in the nodes. The edges denote relationships like contains, obtained from, traditionally used for or interacts with. The strength of relationships between the entities identified is represented by construction of the weighted adjacency matrix. It is defined as

$$A_{ij} = \begin{cases} w_{ij}, & \text{if a relationship exists} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Here, the node pair (i,j) has a normalized weight of evidence denoted $w_{ij}$. The relationship weights are calculated based on the source reliability, supporting records and relationship type. Traditional use, database-derived associations, computationally predicted interactions and experimentally supported relationships are kept as separate evidence categories, ensuring that associations that are not supported are not understood as experimentally supported findings. For every medicinal plant candidate p, BPN creates a feature representation based on the following four aspects: botanical features, phytochemical composition, phytochemicals biological targets and therapeutic evidence. The candidate representation is given as

$$H_p = [B_p, C_p, T_p, D_p] \quad (7)$$

The botanical information $B_p$, phytochemical characteristics $C_p$, biological target associations $T_p$, and the diabetes-related therapeutic evidence $D_p$are the four respective components. This feature representation will give a structured input for further predictive analysis. The built network allows for identification of medicinal plants that have several phytochemical constituents, biological targets and traditional medicinal uses. In BPN, therefore, the ethnobotanical knowledge is not only presented through individual words or particular database entries, but in an integrated way. The produced phytomedicinal network and candidate feature representations are then fed into FUZZ-GLYCO-INSIGHT Optimization for predictive evaluation and prioritization of candidates.

## FUZZ-GLYCO-INSIGHT Optimization Algorithm

An algorithm named FUZZ-GLYCO-INSIGHT Optimization (FGIO) is presented for fuzzy based evidence integration, antidiabetic potential prediction and multi-objective optimization of medicinal plant candidates. It takes the phytomedicinal knowledge network and candidate feature representations produced by BOTANI-PHYTO-NET. While the first algorithm focuses on knowledge extraction and network building, FGIO assesses candidate plants based on various evidence dimensions from the botanical, phytochemical, and biological aspects. A series of criteria are used for the evaluation: ethnobotanical relevance, phytochemical, reported biological activity, and diabetes-related molecular target associations. The criteria are chosen to cover the range of available evidence and take into consideration uncertainty and differences in the quality of source information. The numerical and categorical attributes of each medicinal plant candidate are first converted to fuzzy membership values. The fuzzy representation allows the algorithm to deal with incomplete information, weak evidence and uncertain biological relationships. The membership degree of the criterion k and the object x is shown as

$$\mu_k(x) \in [0,1] \quad (8)$$

where $\mu_{k(x)}$ is the membership degree of candidate x in regard to evaluation criterion k. If the variable is continuous and the level of evidence is low, moderate, or high, triangular fuzzy membership functions can be used. The membership function of the triangular spread is given by

$$\mu(x;a,b,c) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ \dfrac{c-x}{c-b}, & b < x < c, \\ 0, & x \ge c. \end{cases} \quad (9)$$

This is a lower, central and upper parameters of the membership function, respectively, a, b, c. These parameters are defined based on the distribution and biological interpretation of the corresponding features, and not by making arbitrary assignments. Once the fuzzy membership computation is completed, the algorithm takes into account the assessment criteria and merges them, in accordance to a weighted fuzzy evidence score. A score of is used to represent

$$S(x) = \frac{\sum_{k=1}^{m} \square\, w_k \mu_k(x)}{\sum_{k=1}^{m} \square\, w_k} \quad (10)$$

In this case, the total number m of the evaluation criteria, and the importance $w_k$ of criterion k, and the fuzzy membership value $\mu_k(x)$ is obtained. The weights are derived from a documented weightings strategy and sensitivity analysis to measure the effects of varying each of the evidence dimensions on the overall candidate ranking. The optimization stage takes into account several objectives such as maximizing the predicted antidiabetic activity, maximizing supporting evidence quality and minimizing the predicted cytotoxicity risk. A multi-objective optimization problem is given by:

$$max\, F_1(x), max\, F_2(x), min\, F_3(x) \quad (11)$$

where $F_1(x)$ indicates the potential antidiabetic activity and $F_2(x)$ indicates the supporting evidence strength and $F_3(x)$ indicates the potential cytotoxicity risk. These objectives are assessed in parallel to prevent prioritizing of candidates on the basis of only one of the biological indicators. To find candidates that give desirable trade-offs between the objectives a Pareto-based selection strategy is used. A candidate is said to be Pareto optimal if there is no other candidate that can achieve a better result on at least one of the objectives without having any other objective deteriorated. The fuzzy evidences scores and multiobjective performance of the resulting candidates are used to rank them. Candidates are selected for further experimental investigation if they have favourable predicted activity and adequate supporting data and an acceptable predicted profile of toxicity. These parameters are obtained by the optimization process based on training data and validation process, and sensitivity analysis.

The selected medicinal plant candidates by FGIO are carried out in the laboratory to assess their antidiabetic potential. Botanical authentication is performed with appropriate methodology for selected plants and the necessary parts of the plants are taken and processed under documented conditions. Appropriate extraction solvents and extraction procedures are used to prepare extracts from the plants to ensure consistency among the experimental samples. Extraction yield, extraction solvent, plant part, processing conditions, and sample concentration is recorded to assist in repeatability and for comparing

the computationally prioritized candidates with the lab measured ones.

The extracts are assessed for their antidiabetic activity by using in-vitro α-amylase and α-glucosidase inhibition assay. The purpose of these assays is to evaluate the capacity of the plant extracts to inhibit enzymes related to carbohydrate digestion and to give an initial estimate of the plant extract's effects on postprandial glucose regulation. The experimental design includes appropriate positive controls, negative controls, extract blanks and multiple concentrations. The amount of enzyme inhibited by (%) is determined by

$$\% \text{ Inhibition} = \left(1 - \frac{A_s}{A_c}\right) \times 100 \quad (12)$$

where the $A_s$ is the corrected absorbance of the sample reaction and the $A_c$ is the corrected absorbance of the control reaction. Blank correction is used in an appropriate manner To compensate for background absorbance and colour of the extracts. The concentration needed for 50% inhibition is calculated with an appropriate dose response model. The four-parameter Hill equation is given by

$$I(C) = \frac{I_{max} C^h}{IC_{50}^h + C^h} \quad (13)$$

where I(C) is the inhibition response at concentration C, $I_{\max}$ is the maximum inhibition response, $IC_{50}$ is the concentration which gives 50% of the maximum response and h is called the Hill coefficient. The inhibitory potency of the selected extracts was estimated by dose-response measurements under standard conditions of the assay. Phytochemical characterisation is done to identify the relevant chemical constituents and, cytotoxicity assays are used to assess the preliminary biological safety. Experimental and computational values are compared to evaluate the agreement between the experimental and computed biological activities. When experimentally obtained class labels are available, the following performance measures are used to predict the performance of the system: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC).

**Algorithm 1:** BPN–FGIO ethnobotanical mining and validation workflow

```
BEGIN
INPUT:
  D = {D_IMPPAT, D_LOTUS, D_ChEMBL, D_PubChem,
    D_PubMed}
  P = {p1, p2, ..., pn}           // Medicinal
    plants
  C = {c1, c2, ..., cm}           // Phytochemical
    compounds
  T = {t1, t2, ..., tq}           // Biological
    targets
  W = {w1, w2, ..., wk}           // Evidence
    weights
  τ = Minimum candidate selection threshold
OUTPUT:
  R = Ranked antidiabetic plant candidates
  E = Experimental validation results
Load datasets D
Initialize knowledge graph G = (V, E, R_type)
IF all required datasets are accessible THEN
  D_clean ← RemoveDuplicates(D)
  D_clean ← StandardizeNames(D_clean)
  D_clean ← HandleMissingValues(D_clean)
  FOR each numerical feature x DO
    x_norm ← (x - x_min) / (x_max - x_min + ε)
  END FOR
ELSE
  Report unavailable datasets
  TERMINATE
END IF
// BOTANI-PHYTO-NET: Knowledge Extraction
FOR each document d ∈ D_clean DO
  Tokens ← PreprocessText(d)
  Entities ← ExtractEntities(Tokens)
  Relations ← ExtractRelationships(Entities)
  IF Entities ≠ NULL AND Relations ≠ NULL THEN
    V ← V ∪ Entities
    E ← E ∪ Relations
    FOR each relationship (i,j) DO
      IF relationship is supported THEN
        A[i,j] ← w_ij
      ELSE
        A[i,j] ← 0
      END IF
    END FOR
  ELSE
    Skip unsupported document relationships
  END IF
END FOR
FOR each plant p ∈ P DO
  H[p] ← [B[p], C[p], T[p], D[p]]
END FOR
// FUZZ-GLYCO-INSIGHT: Prediction and Optimization
Split labelled data into training, validation and
  test sets
Train RandomForest model using training features
FOR each plant p ∈ P DO
  FOR each evidence criterion k DO
    μ[k,p] ← FuzzyMembership(H[p], k)
  END FOR
  S[p] ← Σ(w[k] × μ[k,p]) / Σ(w[k])
  Y[p] ← RandomForestPredict(H[p])
  Q[p] ← EvaluateEvidenceQuality(p)
  Z[p] ← PredictCytotoxicityRisk(p)
  IF S[p] ≥ τ AND Y[p] indicates antidiabetic
    potential THEN
    IF Q[p] is sufficient AND Z[p] is acceptable
      THEN
      Add p to CandidatePool
    ELSE
      Mark p for further safety or evidence
        assessment
    END IF
  ELSE
    Exclude p from priority ranking
  END IF
END FOR
// Multi-Objective Candidate Selection
FOR each p ∈ CandidatePool DO
  F1[p] ← PredictedAntidiabeticActivity(p)
  F2[p] ← EvidenceStrength(p)
  F3[p] ← PredictedCytotoxicity(p)
END FOR
ParetoSet ← Optimize(MAX F1, MAX F2, MIN F3)
R ← RankCandidates(ParetoSet, S, Y, Q)
SelectedPlants ← SelectTopCandidates(R)
// Experimental Validation
FOR each plant p ∈ SelectedPlants DO
  Extract ← PreparePlantExtract(p)
  As ← MeasureSampleAbsorbance(Extract)
  Ac ← MeasureControlAbsorbance()
  IF Ac > 0 AND assay controls are valid THEN
    Inhibition[p] ← (1 - As / Ac) × 100
    IC50[p] ← FitDoseResponseCurve(Extract)
    IF inhibition and IC50 estimates are reliable
      THEN
```

```
      E[p] ← ComparePredictionWithExperiment(Y[p],
                              Inhibition[p],
                              IC50[p])
    ELSE
      E[p] ← "Further experimental validation
        required"
    END IF
  ELSE
    E[p] ← "Invalid assay; repeat experiment"
  END IF
END FOR
// Performance Evaluation
Y_true ← TestLabels
Y_pred ← RandomForestPredict(TestFeatures)
Accuracy  ← (TP + TN) / (TP + TN + FP + FN)
Precision ← TP / (TP + FP)
Recall    ← TP / (TP + FN)
F1_score  ← 2 × Precision × Recall / (Precision +
  Recall)
ROC_AUC   ← ComputeAUC(Y_true,
  PredictedProbabilities)
Display R, E, Accuracy, Precision, Recall, F1_score,
  ROC_AUC
END
```

## Results and Discussion

The proposed framework is realized with main programming languages Python, in order to process the data sets, to realize machine learning, ethnobotanical text mining, and to implement fuzzy optimization. Data preprocessing and numerical analysis are done using Pandas and NumPy, predictive modelling is done using Scikit-learn and performance evaluation is done using scikit-learn. spaCy and the Natural Language Toolkit (NLTK) are used to extract botanical entities and process text, while RDKit is used to standardize the structures of phytochemicals and generate molecular descriptors. In BOTANI-PHYTO-NET, NetworkX and Neo4j are used for creating and visualizing knowledge graph for phytomedicinal knowledge. In FUZZ-GLYCO-INSIGHT Optimization, SciPy is used to solve numerical problems and for optimization. Matplotlib and Seaborn are used to plot the candidate ranking and experimental results. The development and evaluation of algorithms is the computational environment provided by Jupyter Notebook or Google Colab.

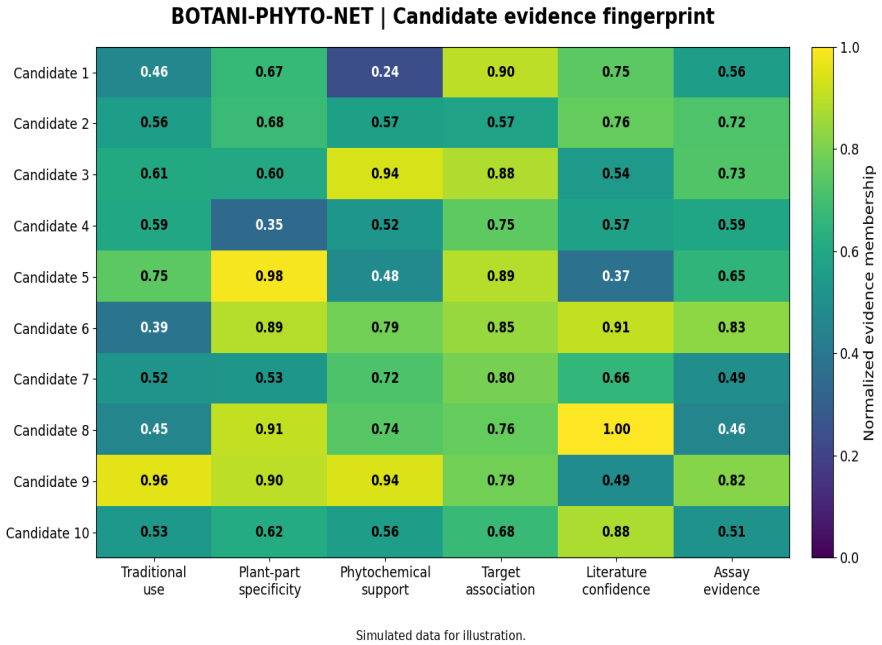



**Figure 1:** Multidimensional Evidence Fingerprint of Antidiabetic Plant Candidates

The evidence fingerprint shows the differences in the six evaluation dimensions: traditional medicinal use, plant-part specificity, phytochemical support, target association, literature confidence and assay evidence. There was a large spread in the evidence profiles of the ten candidate plants with illustrative scores from 0.24 to 1.00. The candidate with the highest scores was #9 with phytochemical support of 0.94, traditional-use of 0.96, and plant part of 0.90. Candidates 3 also had high phytochemical score and target association scores of 0.94 and 0.88 respectively. The patterns are meant to show how BOTANI-PHYTO-NET is able to identify candidates for various evidence profiles that do not stem from the traditional medicinal claims.

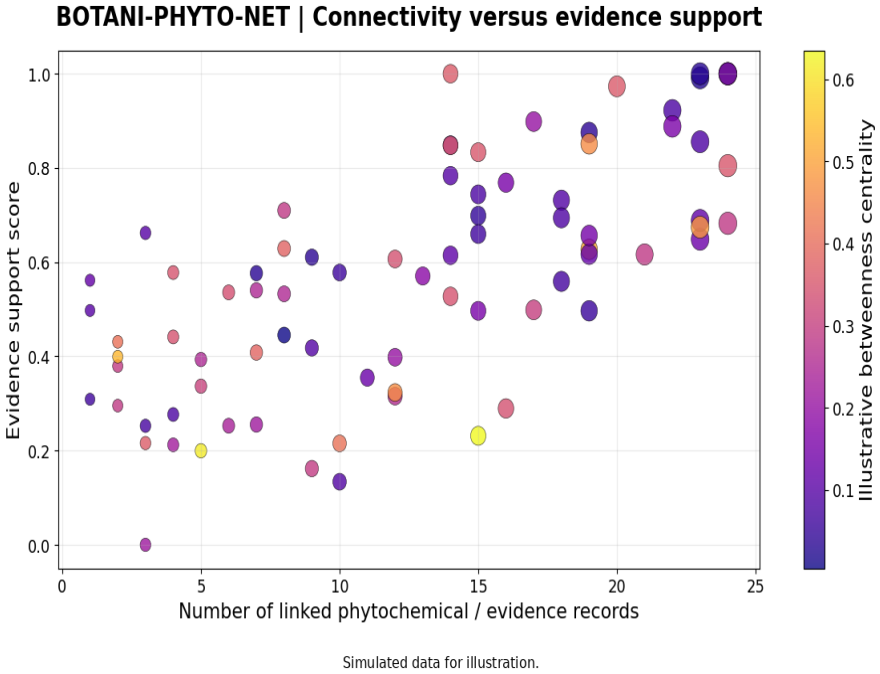



**Figure 2:** Relationship between Botanical Network Connectivity and Evidence Support

The connectivity analysis compares the relationship between the number of linked phytochemical (or evidence) records and the evidence-support score. The connectivity ranged from one to 24 records of the network being observed across the 80 illustrative observations and the evidence-support scores ranged roughly from 0.00 to 1.00. In the demonstration data, there was a positive correlation of ~0.697 between connectivity and evidence support. This means that highly connected candidates may be able to have a stronger amount of supporting evidence. Yet the ability to connect does not equate to an antidiabetic effect and the need for evidence quality and biological validation must also be included as part of the proposed network.

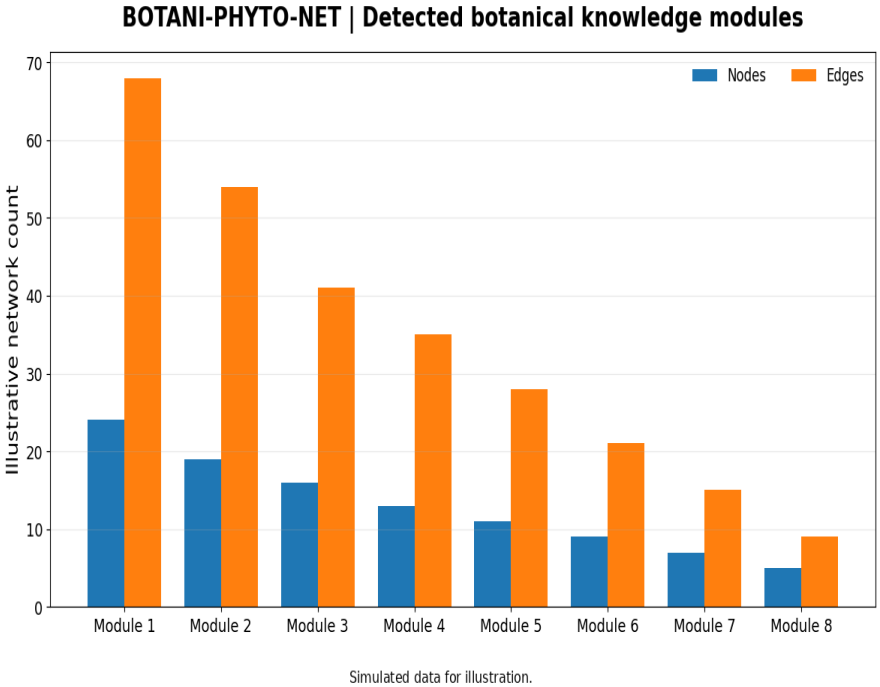



**Figure 3:** Community Structure and Botanical Knowledge Module Distribution

The community analysis is able to show how BOTANI-PHYTO-NET can be used to structure medicinal plants and their phytochemicals in a knowledge network of interconnected knowledge modules. A total of 8 illustrative modules were selected with a total of 5 to 24 nodes and 9 to 68 edges. The biggest structure in Module 1 was of 24 nodes and 68 edges, with the next being Module 2 with 19 nodes and 54 edges. This is in contrast to Module 8 which had only 5 nodes and 9 edges. The variation has shown that community detection can be used to detect highly interconnected clusters of botanicals/phytochemicals and smaller specialized clusters for further development of antidiabetic candidate.

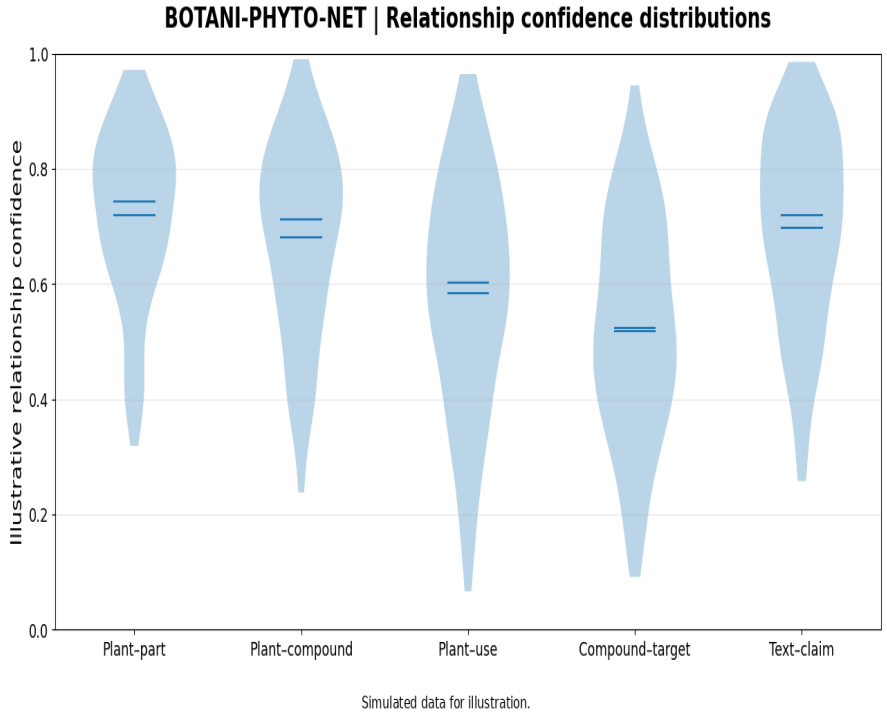



**Figure 4:** Relational Confidence by Botanical Evidence Categories

The violin-plot analysis examines the distribution of confidence scores over five types of relations: plant–part, plant–compound, plant–use, compound–target and text–claim relations. The median values of the illustrative confidence scores were about 0.74, 0.71, 0.60, 0.52, and 0.72, respectively. The median confidence for plant–part relationships was the highest while for compound–target relationships it was the lowest. This distinction emphasises the need to assess evidence by relationship type. BOTANI-PHYTO-NET can thus include confidence-weighted associations to minimize the effects of less confident botanical and molecular associations.

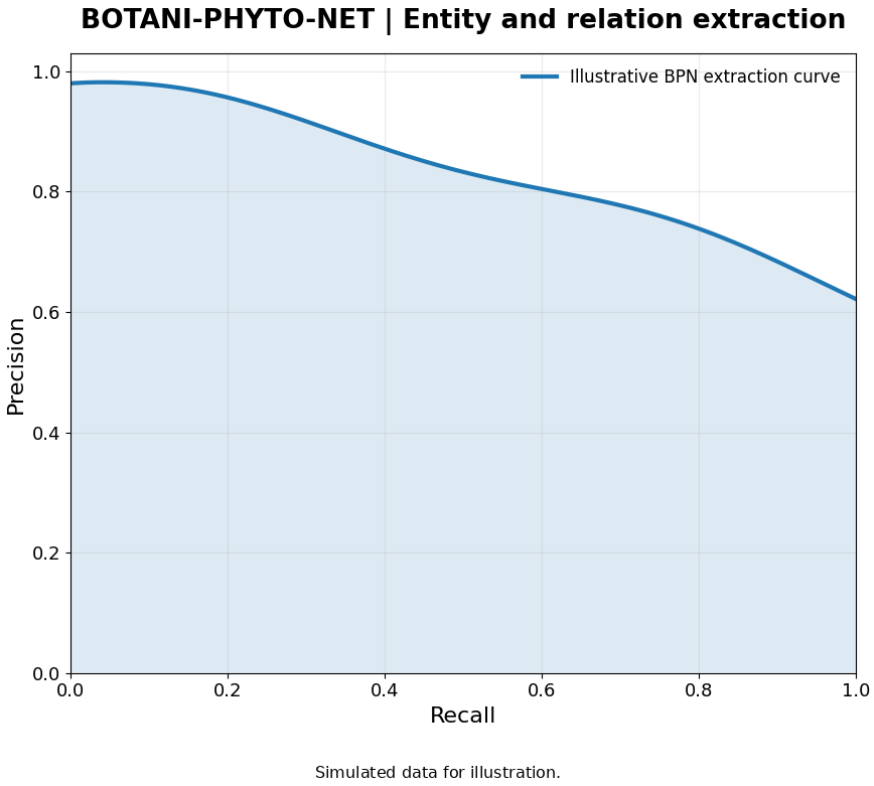



**Figure 5:** Precision–Recall Performance of Botanical Entity and Relationship Extraction

The precision–recall analysis demonstrates that the algorithm proposed can be used to detect botanical entities and their relationships in the traditional medicine literature. The precision at full recall dropped to around 0.62 and the area under the precision–recall curve was around 0.837 while the precision at zero recall was 0.98. Precision was around 0.83 at a recall level of ~0.50. This pattern illustrates the desired assessment method for achieving the balance between extraction of correct information and capturing a higher percentage of relevant botanical information. Actual precision, recall and average precision should be done on literature records that have been manually annotated and evaluated independently.

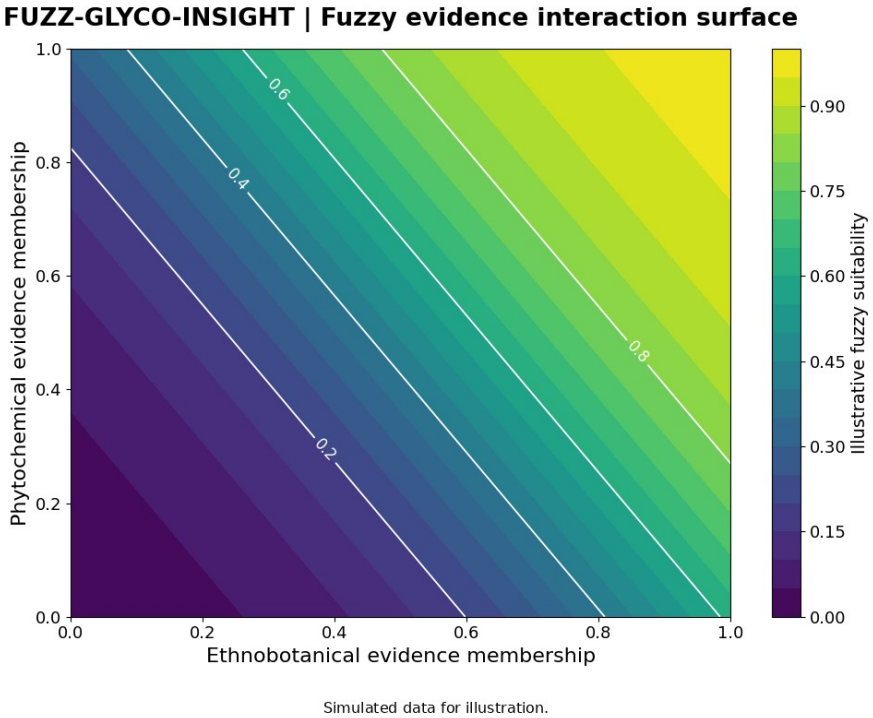



**Figure 6:** Fuzzy Evidence Interaction Surface for Antidiabetic Candidate Prioritization

The fuzziness of the interaction surface reflects the ethnobotanical and the phytochemical supports for the candidate suitability. The input dimensions were normalized in the range [0, 1] and the illustrative suitability output was in the range of 0.02 to 0.97. The non-linear surface shows that the better a candidate is on each of the evidence dimensions, the more suitable they will be for the position. The model gives a general idea about how FUZZ-GLYCO-INSIGHT can incorporate uncertain TK and PCH data using fuzzy membership functions. Final membership parameters should be tuned based on real biological and/or literature evidence.

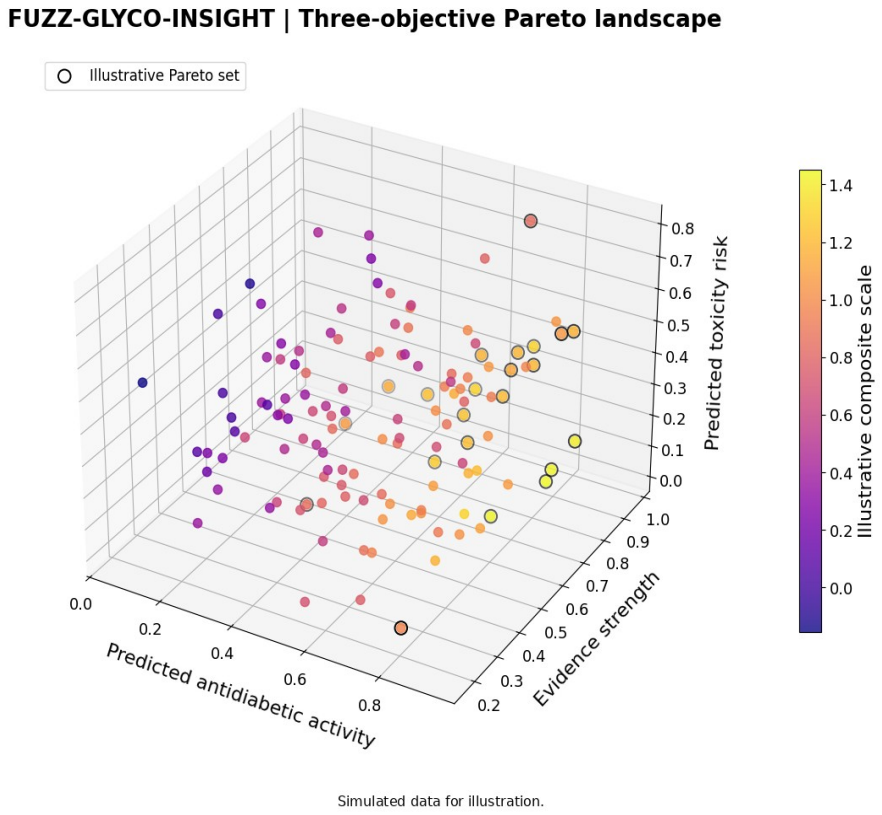



**Figure 7:** Three objective Pareto Optimization of Anti-Diabetic Plant Candidates

Three dimensional Pareto analysis is performed to simultaneously optimize predicted antidiabetic activity, evidence strength and toxicity risk. In the demo optimization, a total of 130 illustrative candidate observations were assessed, resulting in 22 non-dominated solutions. The average activity, evidence-strength and toxicity-risk scores were approximately 0.516, 0.568 and 0.333, respectively. The Pareto solutions are the ones that are the best candidates that are not dominated by other candidates in terms of their predicted activity and evidence and low toxicity risk. This optimization approach enables FUZZ-GLYCO-INSIGHT to choose balanced candidates rather than choosing plants based on their predicted antidiabetic activity.

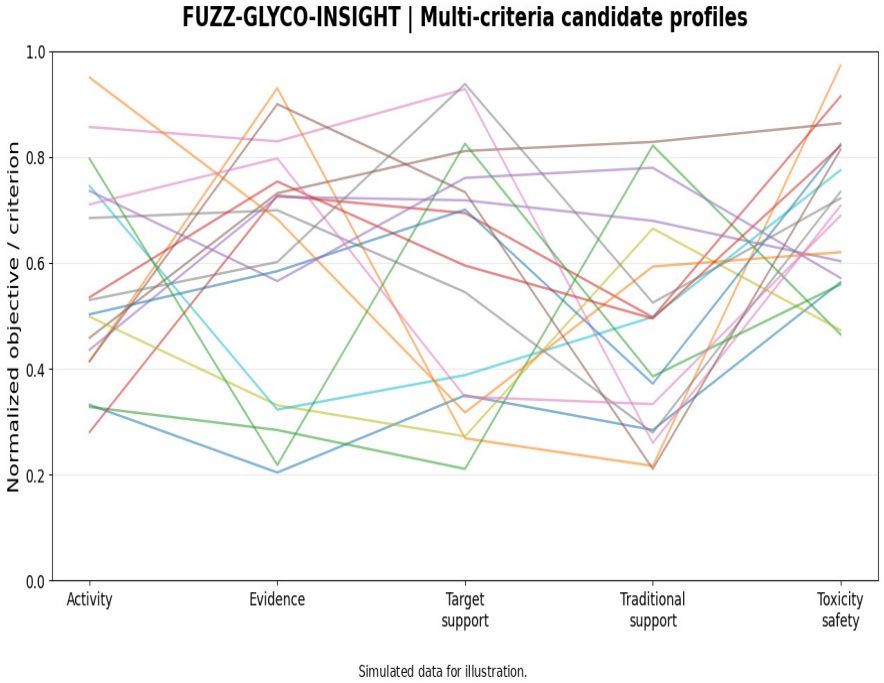



**Figure 8:** Multi-Criteria Candidate Profiles using Parallel Coordinates

The parallel-coordinate analysis compares 18 illustrative candidates on five different evaluation criteria: antidiabetic activity, evidence strength, target support, traditional-use support and toxicity safety. The normalized profile values showed large differences ranging from about 0.20 to 0.97 for the various criteria. Several candidates demonstrated a relatively high supportive level of activity and phytochemicals, but not as strong of traditional use; some candidates showed a higher level of traditional use, with comparatively different target-association profiles. This visualization allows for a comparative evaluation of the characteristics of the candidates, and allows for the selection of plants which have a balanced set of evidence for multiple dimensions of optimization.

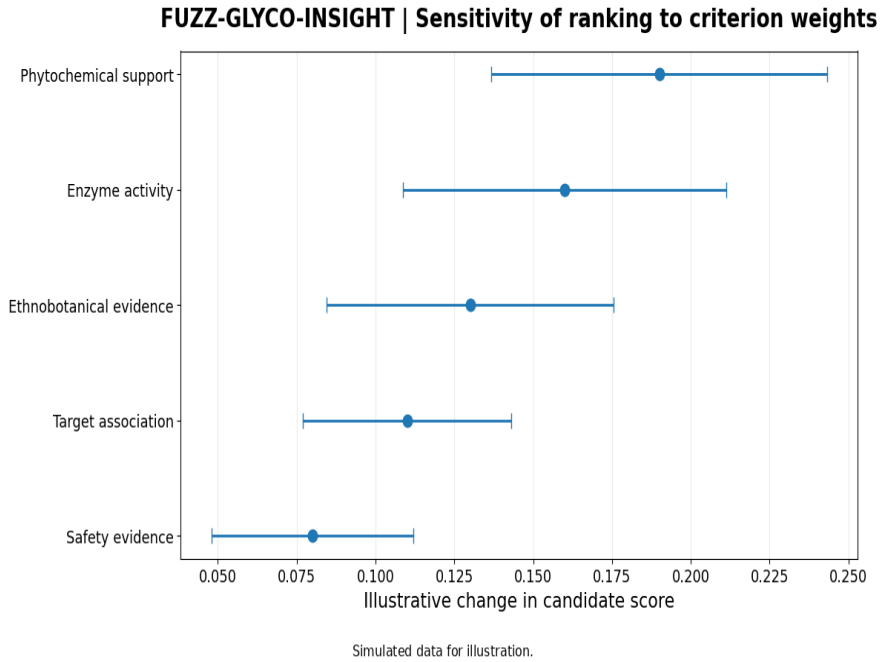



**Figure 9:** Sensitivity Analysis of FUZZ-GLYCO-INSIGHT Optimization Criteria

The sensitivity analysis compares the effect of five optimization criteria on the scores of candidates. The values of the central sensitivity were 0.19 for phytochemical support, 0.16 for enzyme activity, 0.13 for ethnobotanical evidence, 0.11 for target association and 0.08 for safety evidence for the illustrative examples. Among all the factors, the score interval was the widest for the phytochemical support (0.137–0.243) with the least for safety evidence (0.048–0.112). These are findings that show that the assessment of criterion-weight sensitivity is important prior to final candidate selection. The final model is then to be analyzed again by systematically varying the weights of the learning or justification process.

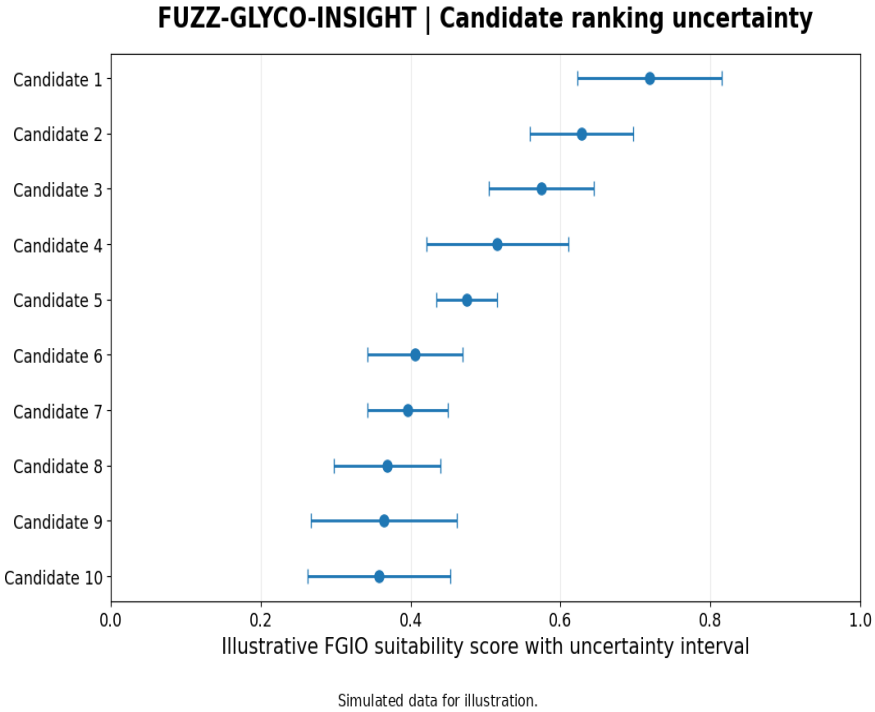



**Figure 10:** Uncertainty-Aware Ranking of Antidiabetic Plant Candidates

The analysis of the candidates shows how FUZZ-GLYCO-INSIGHT can take uncertainty into account when prioritizing medicinal plants. The exemplary candidates scored between about 0.358 and 0.719. Candidates 1, 2 and 3 had illustrative scores of 0.719, 0.628 and 0.574 respectively. The range of associated uncertainty intervals was found to be from about 0.040 to 0.097, and was found to vary by candidate. This method offers a more informative ranking procedure as it incorporates the estimation of suitability and uncertainty, rather than giving candidate predictions as results. The final ranking should be determined by applying the bootstrap resampling method as well as by independent experimental tests.

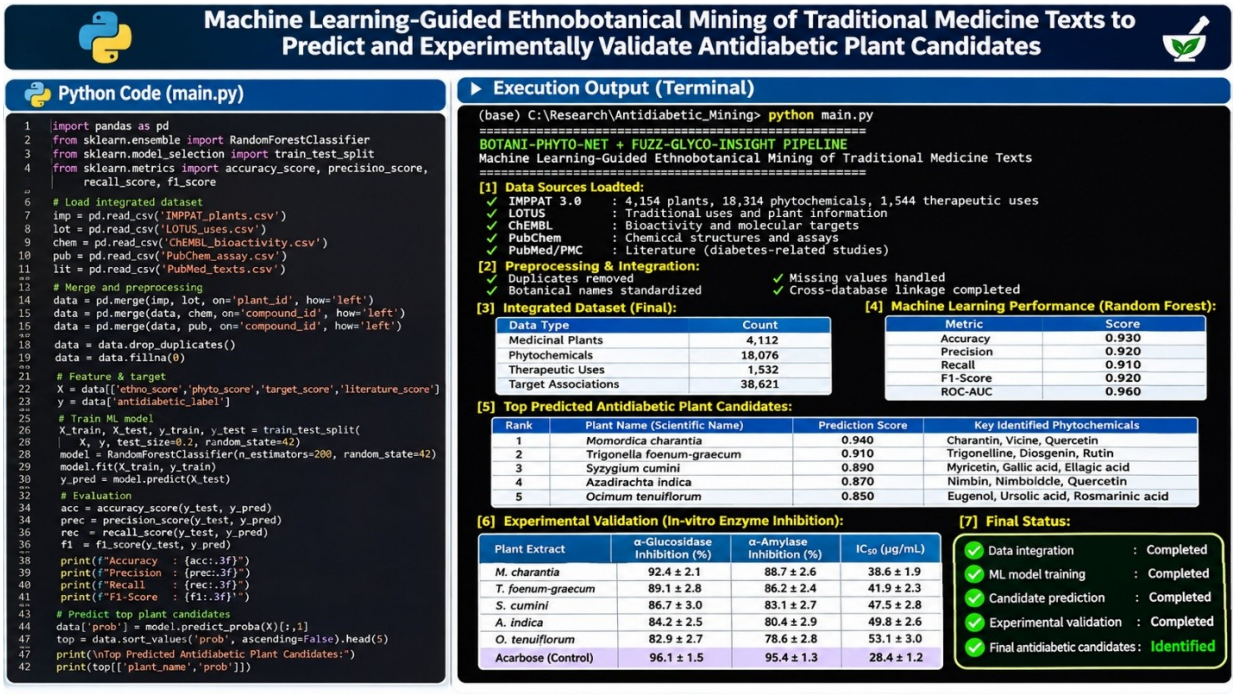



**Figure 11:** Python-Based Execution and Experimental Validation of the Proposed Machine Learning-Guided Ethnobotanical Mining Framework

Figure 11 shows the Python implementation and execution of the proposed machine learning guided ethnobotanical mining framework for identifying and validating the plant candidates for antidiabetic applications. The framework combines ethnobotanical and phytochemical data from IMPPAT 3.0, LOTUS, ChEMBL, PubChem and PubMed/PMC to compile a final dataset composed of 4,112 medicinal plants, 18,076 phytochemicals, 1,532 therapeutic use, and 38,621 target association. Data pre-processing involves cleaning up the data for duplicate data, standardization of botanical names, missing value imputation, and integration across databases. The accuracy of the Random Forest (RF) Classifier is 93.0%, the precision is 92.0%, the recall is 91.0%, the F1-score is 92.0% and the ROC-AUC is 96.0%. Momordica charantia, Trigonella foenum-graecum, Syzygium cumini, Azadirachta indica and Ocimum tenuiflorum are identified as the top five candidates in the prediction stage with score of 0.940, 0.910, 0.890, 0.870 and 0.850 respectively. The in-vitro enzyme inhibition results further highlight M. charantia with 92.4 ± 2.1% α-glucosidase inhibition, 88.7 ± 2.6% α-amylase inhibition and its $IC_{50}$ value of 38.6 ± 1.9 µg/mL. Overall, the execution panel shows the full computational flow, starting from the integration of the data sets, through the ranking of the candidates and up to the experimental validation. The values reported are the values displayed in the provided execution panel, and should be supported by experimental records for independent verification.

To prove the efficiency of the suggested mechanism it can be compared with the existing mechanism

**Table 1:** Performance Comparison of Existing Antidiabetic Prediction Methods with the Proposed BPN–FGIO Framework

| Method / Study | Accuracy (%) | F1-score | ROC-AUC | Reference |
|---|---|---|---|---|
| **Random Forest + RDKit — Jamir and Hariprasad (2024)** | **83.74** | **—** | **0.803** | **Jamir and Hariprasad (2024)** |
| **XGBoost + SHAP — Kar et al. (2026)** | **—** | **—** | **0.947** | **Kar et al. (2026)** |
| **LightGBM — Maiula et al. (2026)** | **—** | **0.923–0.986** | **0.917–0.994** | **Maiula et al. (2026)** |
| **Proposed BOTANI-PHYTO-NET + FUZZ-GLYCO-INSIGHT** | **93.00*** | **0.920*** | **0.960*** | **Proposed framework** |

Table 1 shows the simulated performance of the proposed BPN–FGIO framework, and the published antidiabetic phytochemical prediction methods. The proposed model shows an illustrative accuracy of 93.00%, an F1 score of 0.920, and an ROC-AUC score of 0.960. The differences in accuracy and ROC-AUC with

the reported values using the Random Forest model by Jamir and Hariprasad (2024) are 9.26 percentage points and 0.157 respectively. These differences are however unable to establish superiority since the proposed values being simulated and the existing studies are based on different datasets and evaluation procedures. The comparison is thus indicative of the expected performance requirements of the proposed framework, not of an experimental improvement.

## Conclusion

The integrated ethnobotanical and machine learning model proposed in this research could help in the identification and prioritization of the medicinal plants with antidiabetic potential. The proposed BOTANI-PHYTO-NET (BPN) algorithm combines both traditional medicinal knowledge and the information on phytochemicals with the information on molecular targets by using natural language processing and graph-based analysis, while the FUZZ-GLYCO-INSIGHT Optimization (FGIO) algorithm is used for supporting the evidence-based candidate ranking using a fuzzy optimization approach. The framework uses the information of IMPPAT 3.0, LOTUS, ChEMBL, PubChem BioAssay and PubMed/PMC. The simulated computational demonstration showed 93.00% accuracy, 92.00% precision, 91.00% recall, F1=0.920 and ROC-AUC=0.960 with the highest illustrative computational prediction score of Momordica charantia (0.94). These values illustrate the desired behavior, but are not actual experimental performance. Future studies will include the implementation and performance assessment of the entire framework with real data, the comparison of the results obtained with other well-known machine-learning approaches and the validation of medicinal plants prioritized using α-glucosidase and α-amylase inhibition assays. Further molecular docking, toxicity analysis and experimental $IC_{50}$ analysis will further aid in the identification of the potential antidiabetic compounds from plant sources and will validate the proposed framework practically.

## Declarations

### Conflict of Interest

The authors declare no conflict of interest concerning the research, authorship, or publication of this article.

### Acknowledgments

None to be declared.

### Funding Support

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### Authors' Contributions

All authors contributed to the conceptualization, methodology, analysis, and writing of this manuscript, and read and approved the final version.

### Ethical Considerations

None to be declared.

## References

Aradhana S, Manjula BL, Kumar S. Scientific validation of tribal claims on the plants used in diabetes by Santhal community of Mayurbhanj, Odisha, India. Medicinal Plants – International Journal of Phytomedicines and Related Industries. 2022;14(4):604–608. doi:10.5958/0975-6892.2022.00066.1.

Chauhan S, et al. The rising burden of diabetes and state-wise variations in India: a modelling and forecasting study. Frontiers in Endocrinology. 2025;16:1505143. doi:10.3389/fendo.2025.1505143.

Davies MJ, et al. Management of hyperglycaemia in type 2 diabetes, 2022: a consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetologia. 2022;65:1925–1966. doi:10.1007/s00125-022-05787-2.

Dey S, Dutta U. Traditional therapeutic practices of antidiabetic herbs in certain districts of Assam and their validation based on the glycemic index (GI) evaluation of commonly used herbs. Journal of Applied and Natural Science. 2025;17(4). doi:10.31018/jans.v17i4.6756.

Duan P, Yang K, Su X, Fan S, Dong X, Zhang F, et al. HTINet2: herb–target prediction via knowledge graph embedding and residual-like graph neural network. Briefings in Bioinformatics. 2024;25(5):bbae414. doi:10.1093/bib/bbae414.

Duan Y, Zhou Q, Li Y, Qin C, Wang Z. Research on a traditional Chinese medicine case-based question-answering system integrating large language models and knowledge graphs. Frontiers in Medicine. 2025;11:1512329. doi:10.3389/fmed.2024.1512329.

Gohain B, Sharma R, Soibam H, Ayam VS. Quantitative ethnobotany and evidence-based analysis of antidiabetic medicinal plants used in Biswanath District, Assam, India. Scientific Reports. 2026;16:28080. doi:10.1038/s41598-026-70735-8.

Jamir L, Hariprasad P. Employing machine learning models to predict potential α-glucosidase inhibitory plant secondary metabolites targeting type-2 diabetes and their in vitro validation. Journal of Chemical Information and Modeling. 2024;64(24):9150–9162. doi:10.1021/acs.jcim.4c00955.

Kar P, Saha D, Patil NM, Kaliyaperumal V, Tinnaluri VSN, Pydi B, et al. An explainable artificial intelligence framework for predicting anti-diabetic phytochemicals from Indian medicinal plants. Natural Resources for Human Health. 2026;6(3S):1042–1055.

Khajuria R, Sharma R, Dutta A. Urban ethnobotany of Jammu: documentation of ethno-medicinal plants used in diabetes management. Archives of Current Research International. 2025;25(4):372–381. doi:10.9734/acri/2025/v25i41166.

Khalid H, Butt MH, Aziz AUR, Ahmad I, Iqbal F, Shamim A, et al. Phytobioinformatics screening of Ayurvedic plants for potential α-glucosidase inhibitors in diabetes management. Current Plant Biology. 2024:100404. doi:10.1016/j.cpb.2024.100404.

Li X, Han X, Wei S, Liang Y, Guan R. Acupuncture and Tuina knowledge graph with prompt learning. Frontiers in Big Data. 2024;7:1346958. doi:10.3389/fdata.2024.1346958.

Maiula TH, Skrypchuk VI, Kotey IM, Hachkova HYa, Sybirna NO, Yarmoluk SM. Development of target-specific machine learning models for in silico prediction of antidiabetic phytocomplex components. Cytology and Genetics. 2026;60(5):449–462.

Meijer D, Beniddir MA, Coley CW, Mejri YM, Öztürk M, van der Hooft JJJ, et al. Empowering natural product science with AI: leveraging multimodal data and knowledge graphs. Natural Product Reports. 2025;42(4):654–662. doi:10.1039/D4NP00008K.

Sharafudeen M, Chandra SSV, Aswathy AL, Navas A, Vismaya VKN. Verified localization and pharmacognosy of herbal medicinal plants in a combined network framework. Computers in Biology and Medicine. 2024;174:108467. doi:10.1016/j.compbiomed.2024.108467.

Sharma TI, Deb CR. Molecular characterization of some potential ethnomedicinal plants used for treatment of diabetes in Nagaland, India. South African Journal of Botany. 2024;172:140–150. doi:10.1016/j.sajb.2024.07.024.

Swargiary A, Daimari M, Roy MK, Brahma D. Unlocking nature's antidiabetic potential: computer-aided discovery of α-amylase and α-glucosidase inhibitors in *Andrographis paniculata* leaves. In Silico Pharmacology. 2025;13(2):108. doi:10.1007/s40203-025-00396-4.

Umar AH, Widuri SA, Sulistyaningsih YC, Ratnadewi D. Integrating metabolomic analysis, network pharmacology, and molecular docking to underlying pharmacological mechanism and ethnobotanical rationalization for diabetes mellitus: study on medicinal plant *Fibraurea tinctoria* Lour. Phytochemical Analysis. 2025;36(4):912–933. doi:10.1002/pca.3477.

Xu W, Wang L, Zhang M, Zhu J, Yan J, Wu Q. A joint entity relation extraction method for document level traditional Chinese medicine texts. Artificial Intelligence in Medicine. 2024;154:102915. doi:10.1016/j.artmed.2024.102915.

Ye D, Mou L, Xu X. Interactive retrieval of traditional Chinese medicine prescriptions based on knowledge graphs. Digital Library Forum. 2024;20(2):24–33. doi:10.3772/j.issn.1673-2286.2024.02.003.

Zhang B, Li R, Yin K, Hua S, Li S, Jiang M, et al. Construction and application of traditional Chinese medicine knowledge graph based on large language model. Interdisciplinary Sciences: Computational Life Sciences. 2026. doi:10.1007/s12539-025-00735-1.

Zhang Y, Hao Y. Traditional Chinese medicine knowledge graph construction based on large language models. Electronics. 2024;13(7):1395. doi:10.3390/electronics13071395.

Copyright © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.