Received: 21 June 2025 | Revised: 27 November 2025 | Accepted: 20 January 2026

DOI: 10.1002/ppp3.70189

Plants People Planet PPP Open Access

RESEARCH ARTICLE

Accelerating the prioritisation of plant species with underexplored medicinal potential: The *pm4mp* (Phylogenetic Methods for Medicinal Plants) R package

Giovanni Zecca | Elisa Toini | Massimo Labra | Fabrizio Grassi

Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy

Correspondence
Giovanni Zecca, Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy.
Email: giovanni.zecca@unimib.it; giovanni.zecca@gmail.com

Funding information
This work was supported by the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4—call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian MUR funded by the European Union—NextGenerationEU; Award Number: Project code CN_00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP H43C22000530001, project title 'National Biodiversity Future Center—NBFC'. This work was also supported by the PRIN (Project of Relevant National Interest) project Plants Bioprospecting Of Serine Proteases Inhibitors To Develop An Innovative Colon Cancer Prevention Strategy, 'PRESERVE', by the 'Ministero dell'Istruzione dell'Università e della Ricerca' (MIUR), PRIN 2020 - CUP H45E21000160001.

Societal Impact Statement

Medicinal plants used in ethnobotanical traditions to treat or prevent diseases have gained renewed interest for their largely untapped potential in drug discovery. In this study, we developed and tested novel methods to prioritise plant species based on their unexplored medicinal potential. By enabling researchers to target the most promising species, these approaches reduce the time and costs of bioprospecting while increasing the likelihood of identifying beneficial compounds. At the same time, they help minimise the environmental impact associated with research activities. Overall, our findings support more sustainable drug discovery practices and highlight the responsible use of biodiversity to advance human health.

Summary

- Plants are a key source of active compounds, with many drugs derived from them. Various methods have been used to explore the medicinal potential of unexploited taxa, but identifying the most active species remains challenging.
- Molecular phylogenetics holds promise for plant bioprospecting, but issues remain, especially when dealing with large-scale phylogeny. This paper presents a workflow that integrates new and existing methods to accelerate the identification and prioritisation of potential medicinal plants, focusing on the most promising regions of a phylogeny and assigning a value to each taxon based on its medicinal potential.
- We introduce *pm4mp*, an R package that implements the newly developed methods and is available for free on GitHub. *pm4mp* provides functionalities for identifying stable *hot nodes* across multiple analysis replicates, extracting *hot trees* for a disease of interest, prioritising target species by using new approaches and visualising the results graphically.
- We demonstrate the usefulness of *pm4mp* by analysing medicinal plant data on 10 diseases from a public database together with a phylogeny of 30,000+ land plants. Our findings show the effectiveness of the newly proposed methods,

which represent a substantial enhancement over the existing approaches for plant bioprospecting.

KEYWORDS
bioprospecting, Hidden Medicinal Plant Prediction, Hot Ancestry Score, hot nodes, hot species, hot trees, medicinal plants, *pm4mp*

1 | INTRODUCTION

Throughout their history, plants have differentiated a wide variety of chemical compounds, leading to the evolution of various groups of specialised metabolites with endogenous biological functions, such as alkaloids, terpenoids and phenols (Rønsted et al., 2012; Wink, 2003). Plants are therefore a primary source of functionally active compounds, and many current pharmacological drugs for the treatment and cure of various ailments and diseases have been derived from botanicals (Gurib-Fakim, 2006; Tu, 2011; Khaiwa et al., 2021). It has been reported that approximately 80% of the global population relies on traditional herbal medicines, which remain the primary means of disease prevention for millions of individuals worldwide (Davis & Choisy, 2024). Recent estimates have valued the global medicinal plant market at $201 billion in 2023, with projected growth reaching $375.6 billion by 2032 (Zamani et al., 2025). Nonetheless, a substantial proportion of the world's plants remain to be assessed for their medicinal properties, with only 16% of plants deemed to be therapeutic having been subjected to biological activity testing (Davis & Choisy, 2024). Researchers have applied a variety of approaches to exploit the potential medicinal utility of untapped taxa. The random screening strategy has been extensively utilised in the selection of potential medicinal plants, resulting in significant discoveries, including the identification of taxol (Wani & Horwitz, 2014). While this approach has proven useful, it is also costly and time consuming. Consequently, several authors have adopted a selection strategy of candidate plants based on available ethnobotanical information (Albuquerque et al., 2012; Khafagi & Dewedar, 2000; Lulekal et al., 2008). Cross-cultural comparisons of medicinal flora have revealed a link between taxonomic affinity and ethnomedicinal use in various communities, providing a promising direction for bioscreening studies (Domingo-Fernández et al., 2023; Reinaldo et al., 2020; Saslis-Lagoudakis et al., 2011). However, identifying the most biologically active species from the wide range of existing plants for the purpose of further screening remains a challenging task (Zaman et al., 2021). In this context, molecular phylogenetics have shown to be an effective way to identify plants for a specific use, overcoming the limitations of more traditional taxonomic approaches (Rønsted et al., 2012; Saslis-Lagoudakis et al., 2012). Phylogenetic tools have been used to demonstrate clustering of traditionally used and bioactive species (Rønsted et al., 2008; Saslis-Lagoudakis et al., 2011; Zhu et al., 2011) and to predict the medicinal properties of as yet unexploited plants, effectively narrowing the number of possible target species (Ernst et al., 2016; Pellicer et al., 2018). These methods are based on the assumption that evolutionarily related taxa tend to possess comparable medicinal properties. However, specialised metabolites and phylogeny may exhibit a lack of congruence because of various phenomena such as convergent evolution, whereby similar traits arise independently in distantly related taxa in response to common environmental pressures. For instance, the capacity to synthesise cyanogenic glycosides has apparently evolved independently in numerous plant families, serving as a form of chemical defence (Rønsted et al., 2012). Therefore, in order to be used predictively, the association between phylogeny and medicinal properties must be tested. For this purpose, a range of metrics, frequently borrowed from phylogenetic comparative methods and community phylogenetics, have been employed to assess the strength of the phylogenetic signal, the existence of significant phylogenetic clustering or overdispersion and the extent of phylogenetic similarity or divergence between predefined groups of plants (Rønsted et al., 2012; Saslis-Lagoudakis et al., 2012; Ernst et al., 2016; Pellicer et al., 2018; Souza-Neto et al., 2016; Thompson & Hawkins, 2025). In the past, a widely used and proven effective approach has been based on the identification of 'hot nodes' (i.e. those nodes that show a significant overabundance of medicinal plants among the terminal taxa distal to them) (Ernst et al., 2016; Pellicer et al., 2018; Rønsted et al., 2012; Saslis-Lagoudakis et al., 2011, 2012; Yessoufou et al., 2015; Zaman et al., 2021). The most appealing aspect of this method is the ability to test for the presence of clades with significant clustering of medicinal plants while locating them within a phylogeny of interest. Nevertheless, it is possible to highlight some potential critical issues with this approach. As pointed out by Atienza-Barthelemy et al. (2024), unlike other metrics, hot nodes depend solely on the topology of the tree, completely ignoring branch length information. In the same paper, these authors found a partial inconsistency between the results obtained using hot nodes approach and those obtained using other common phylogenetic divergence metrics. Moreover, especially when dealing with large phylogenies, the number of identified hot nodes can be substantial, which prompts the question of whether all of them are equally relevant from a predictive point of view. Indeed, when the number of identified hot nodes is large, the number of potential target taxa is even larger, which again poses a problem of prioritisation. Although the deepest hot nodes may be interesting from an evolutionary perspective, they may not be very useful in practice for the identification of new medicinal species. Conversely, an exclusive focus on the shallowest hot nodes may result in the loss of information pertaining to phylogenetic context. Finally, since the identification of hot nodes is based on a randomisation procedure, it is expected that the different replications

will not have completely identical results. However, as far as we know, to what extent this may influence the final results and whether there is consistency between the hot nodes identified in the different analysis replicates has never been ascertained in previous work.

Until recently, knowledge about medicinal plants was only available in specialist journals, manuals and textbooks. The advent and increasing availability of scientific databases have led to the development of a new paradigm in the collection, integration and dissemination of information on medicinal plants (Fathifar et al., 2023). Contemporary databases have the capacity to encompass a vast array of information pertaining to the global flora, with much of this information deriving from scientific literature. If properly managed and regularly updated, scientific databases that explicitly link known medicinal uses to specific natural compounds represent an invaluable resource for plant bioprospecting and in silico screening (Buenz et al., 2018).

In parallel with the spread of databases, an increasing number of molecular phylogenies have become available in recent decades, many of which have been used to further test evolutionary and ecological hypotheses (Piel et al., 2009). Similarly, phylogeny-guided selection of target plants for bioprospecting has also benefited from this availability. Although choosing a specific reference phylogeny may depend on the objectives of the individual study, it is possible to outline some basic features that may be useful for bioprospecting. Ideally, the phylogeny chosen should be robust, based on multiple molecular markers, and provide a broad phylogenetic context in relation to the purpose of the study.

In this study, we present three new methods useful to accelerate the identification and the prioritisation of new medicinal plants. These methods can be used to identify consistent hot nodes between multiple replicates of the analysis, to extend the concept of a hot node to the idea of a 'hot tree' and to apply new functionalities capable of prioritising the target species, assigning each a value based on its medicinal potential. Auxiliary functions are also introduced to facilitate the downloading and preprocessing of input data and to display the results in graphical form.

In the following, leveraging the information available in an existing database and phylogeny, we describe our approach through its application to a group of diseases selected as an example to illustrate the novelties introduced. The described workflow (Figure 1) involves the use of both existing and newly developed software. We chose the R programming language to provide users with a single, consistent framework for analysis. R is a free and platform-independent software environment that is widely used for data analysis, statistical computing and has several phylogeny-specific packages, which makes it suitable for our work. All custom scripts developed for this project were implemented in the new R package *pm4mp*, freely available at https://github.com/gzecca/pm4mp.git.

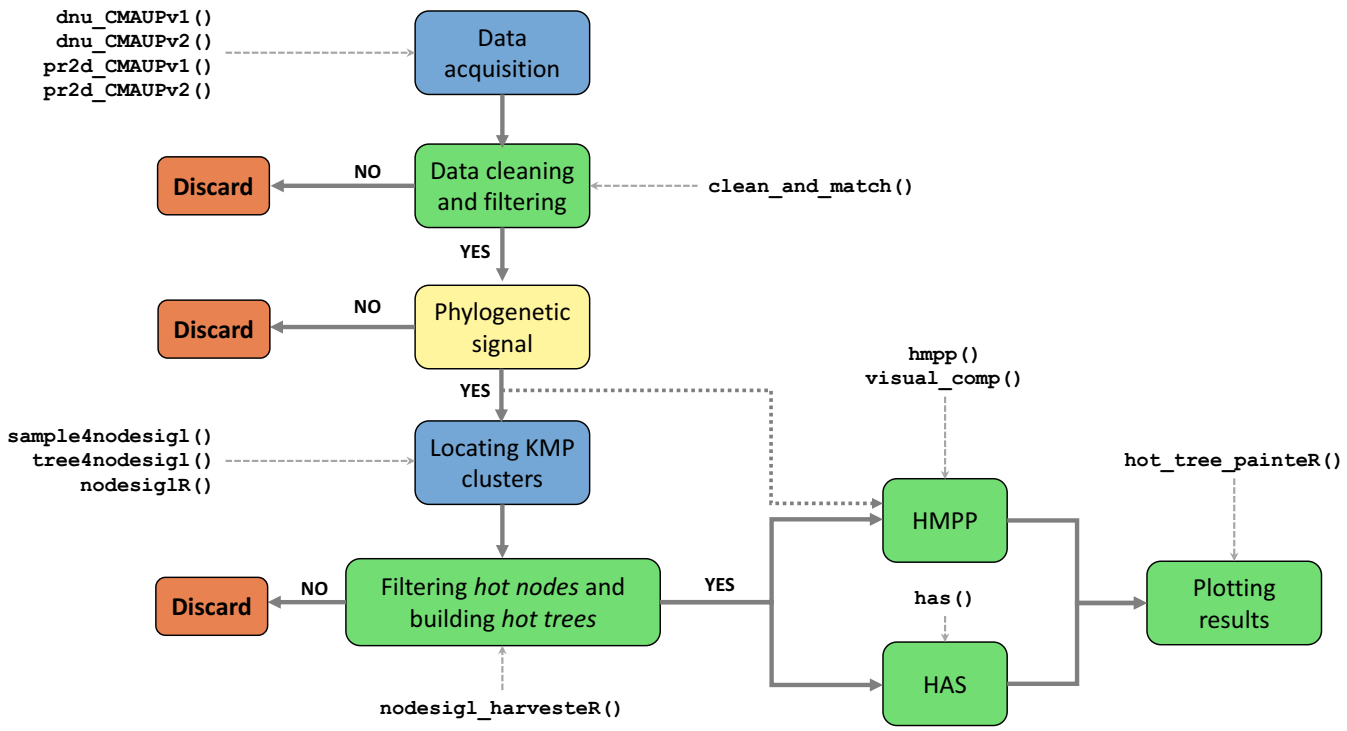

FIGURE 1 Illustration of the analysis pipeline described in the paper. The rounded rectangles connected by solid grey arrows depict the steps of the workflow. The blue rectangles indicate the stages in which the *pm4mp* package integrates with existing databases or software. The green rectangles denote the stages that are exclusive to the *pm4mp* package. The yellow rectangle shows the sole stage where the *pm4mp* package was not utilised (see the main text). Brick orange rectangles indicate that data filtering was performed in the process. The functions of the *pm4mp* package applied are indicated in Courier font and linked to the relevant workflow steps by dashed grey arrows. The dotted grey arrow indicates a possible alternative workflow that does not include hot trees detection. HAS, Hot Ancestry Score; HMPP, Hidden Medicinal Plant Prediction; KMP, known medicinal plants.

2 | MATERIAL AND METHODS

2.1 | Data collection

Data about known medicinal plants (KMPs) were obtained from the Collective Molecular Activities of Useful Plants database (CMAUP database; Zeng et al., 2019; Hou et al., 2024), which in its first release (i.e. CMAUP v1.0) included information on 5645 plants related to 656 human diseases through the interaction of their molecules with human target proteins. In order to illustrate the applicability of our workflow, we selected 10 different diseases from those present in CMAUP v1.0: allergy, Alzheimer's disease, arteriosclerosis, colon cancer, depression, hyperglycaemia, hypertension, insomnia, malaria and prostate hyperplasia. Disease-related keywords were used to query the database and to identify lists of medicinal plants linked to the diseases of interest. When the returned output included different items, these were merged to obtain the final list of medicinal plants related to a specific disease (Table S1). The custom functions *dnu_CMAUPv1()* and *pr2d_CMAUPv1()* were used to facilitate the download of the data of interest in '.csv' format. Raw data downloaded from CMAUP v1.0 are available in Data S1 (accessed on March 3, 2023).

2.2 | Reference phylogeny selection

We used the chronogram published by Zanne et al. (2014), based on seven loci (18S rDNA, 26S rDNA, ITS, *mat*K, *rbc*L, *atp*B and *trn*L-F) from 32,223 land plant species, as reference phylogeny. Given the high number of species present in the reference phylogeny, five monophyletic subtrees were extracted to be analysed separately, speeding up computation time. The R package *castor* was used for tree manipulations (Louca & Doebeli, 2018). The five extracted subtrees were Monocotyledoneae (MO, 7060 taxa), Magnoliidae (MA, 1047 taxa), Superrosidae (SR, 10,009 taxa), Superasteridae (SA, 11,324 taxa) and Basal Eudicots (BE, 948 taxa). Overall, a total of 30,388 species were retained for subsequent analyses. We will henceforth use 'reference phylogeny' to refer indifferently to each of these monophyletic subtrees, while we will use the plural 'reference phylogenies' to refer to two or more subtrees together. Reference phylogenies used in this work are available in Data S1.

2.3 | Data curation

The downloaded data underwent a cleaning process to conform the species names to the format used in the reference phylogenies. Several cleaning steps were performed, in particular: Leading and/or trailing whitespace were removed; internal whitespaces were replaced by underscores, and multiple consecutive underscores, when present, were replaced by one; information relating subspecies (including hybrid subspecies), variants, cultivars, etc. was discarded; all species names were written in the case sensitive format '*Genus_species*', except for hybrid species; the hybrid names were written using the format '*Genus_x*' after checking their identity and uniqueness against the names included in the original alignment file used to construct the reference chronogram; ambiguous or incomplete names, which could have produced multiple matches, such as those named '*Genus_sp/_spp/_st/_sp./_spp./_st.*', were discarded.

Subsequently, the lists of medicinal plants were compared with the tip names found in the reference phylogenies and exactly matching species were kept. Furthermore, fuzzy matching names were identified using the Full Damerau–Levenshtein distance implemented in the *stringdist* R package (van der Loo, 2014) and setting an acceptance threshold for string similarity score at 0.9. All fuzzy matches found were individually evaluated to decide whether to accept or reject them, and those retained were modified to agree with the species names found in Zanne et al. (2014). Non-matching species were discarded and excluded from the subsequent analyses. The entire data curation process was automatically carried out using the function *clean_and_match()* with the argument '*reftype*' set to 'z'. This newly implemented function allows the user to customise some of the data cleaning features described above according to specific needs.

2.4 | Preliminary investigation of the phylogenetic signal

The *D*-statistic (Fritz & Purvis, 2010) implemented in the *caper* R package (Orme et al., 2023), the δ statistic (Borges et al., 2019), the standard effect size mean pairwise distance (SES_{MPD}) and the standard effect size mean nearest taxon distance (SES_{MNTD}) implemented in the *picante* R package (Kembel et al., 2010) were applied to investigate the phylogenetic distribution of medicinal species within the reference phylogenies. In each single analysis, taxa in the reference phylogenies were labelled as '1' if they corresponded to a KMP based on CMAUP v1.0 database, such as '0' otherwise. To ensure at least some level of initial information, we tested only combinations of trees and diseases containing at least 10 KMPs.

The *D*-statistic is used with binary traits under the assumption that they have evolved according to Brownian motion's threshold model. In contrast, the δ statistic is based on Shannon entropy and has no specific requirements regarding the number of states or the evolutionary model of the trait. The SES_{MPD} and SES_{MNTD} are methods derived from the field of phylogenetic community structure analysis and are two measures of phylogenetic relatedness of species within a particular predefined group (community). All these statistics compare the observed value to the pattern expected under a null random model. The SES_{MPD} is generally considered to be more sensitive to tree-wide phylogenetic patterns, whereas the SES_{MNTD} is more sensitive to patterns located closer to the tips of the phylogeny. Further guidance on interpreting the results of these tests is provided in Data S2 and in the original methodological references. Because different approaches are based on different assumptions, all of the above-mentioned tests are expected to measure different aspects of phylogenetic structure. For this reason, we considered further investigation of all cases where at least one of δ, SES_{MPD} or SES_{MNTD} was

significant or where the *D*-statistic differed significantly from a random structure with a value of $D \leq 0.8$. No further analysis was conducted in all other cases.

2.5 | Stability evaluation of hot nodes

The *nodesigl* command in Phylocom v4.2 (Webb et al., 2008) was used to identify the location of phylogenetic clustering (i.e. the position of hot nodes) within reference phylogenies for the diseases that showed evidence of structuring in the previous analysis. Input files were prepared using the custom functions *tree4nodesigl()* and *sample4nodesigl()*. The *nodesigl* analysis identifies hot nodes by comparing the number of medicinal species descending from a given node with a null distribution of values obtained randomly drawing without replacement the taxa labelled as 'medicinal species' from the list of all species in the phylogeny pool (i.e. null model 2 implemented in Phylocom). The default setting is to use 999 randomisations (i.e. *−r* 999) to generate the reference null distributions, and it has been widely used in literature (Atienza-Barthelemy et al., 2024; Crum et al., 2024; Ernst et al., 2016; Saslis-Lagoudakis et al., 2012; Zaman et al., 2021). While a single analysis with 999 randomisations might be adequate under certain circumstances, phylogenies that include hundreds or thousands of taxa might require multiple replicates to properly evaluate the results. To assess to what extent the identified hot nodes were congruent between different software executions we iterated the *nodesigl* analysis with the default setting on many replicates using the custom function *nodesiglR()*. The optimal number of iterations was estimated in a preliminary test using data on colon cancer, hypertension and malaria diseases together with the five reference phylogenies. Once the optimal number of *nodesigl* analysis replicates was estimated in the preliminary test, it was applied in all the analyses.

2.6 | Method I: The nodesigl_harvesteR: From hot nodes to hot trees

For each investigated combination of disease and reference phylogeny, only hot nodes that were associated with a significant overabundance of KMPs in all replicates and whose set of descending tips contain at least a predetermined percentage (see below) of KMPs were kept (hereinafter we will refer to these nodes as *stable hot nodes*). Among the stable hot nodes, independent subsets of phylogenetically nested nodes were recognised, and within each subset, the node corresponding to the most recent common ancestor (MRCA) was determined. Independent subtrees rooted in the identified MRCAs and containing distinct subsets of the species were then extracted from the reference phylogenies. These trees were clades of the original reference phylogenies where KMPs and stable hot nodes were clumped, thus representing promising targets for the search for new medicinal species. For this reason, we called them *hot trees* (Figure 2a). Furthermore, stable hot nodes that consistently achieved the highest possible ranking across all replicates of the *nodesigl* analysis were identified. The processing of *nodesigl* outputs and the identification of the associated hot trees have been implemented in the function *nodesigl_harvesteR()*.

2.7 | Establishing priorities among taxa

Especially when dealing with large reference phylogenies, the hot trees obtained with the *nodesigl_harvesteR* method could still include hundreds of taxa. In order to reduce the number of potential target species and focus on those taxa with the highest medicinal potential, two new methods employing the hot trees as a starting point were proposed. These new methods, which made it possible to further exploit the information contained in hot trees, are outlined in the next two paragraphs.

2.8 | Method II: The Hot Ancestry Score

The *Hot Ancestry Score* (HAS) method was derived directly from the hot node concept. Since hot nodes have been defined as those nodes that are significantly over-represented by species related to a given disease, the portions of a hot tree in which hot nodes were phylogenetically nested were considered as the clades in which the likelihood of finding new medicinal plants was highest. According to this view, the greater the number of hierarchically nested hot nodes from which a species descended, the more likely it was to be medicinal. This allowed us for a straightforward species ranking once the number of ancestral hot nodes for each terminal taxon was calculated (Figure 2b). As this method is based on the hot nodes approach, it does not take into account information relating to branch lengths. The steps necessary to calculate taxa scores were implemented in the function *has()*.

2.9 | Method III: The Hidden Medicinal Plant Prediction

The *Hidden Medicinal Plant Prediction* (HMPP) method was based on a modification of the Hidden State Prediction (HSP) approach. HSP methods are algorithms for predicting unknown (hidden) traits based on a sample of known character states and a phylogenetic tree (Zaneveld & Thurber, 2014). We built on the implementation of HSP for a binary trait based on the binomial distribution implemented in the *castor* R package (i.e. the function *hsp_binomial()*) to extend its application to our case where only information about one of the two possible states was available at tips (i.e. no information about the 'non-medicinal' state was available). Information on branch lengths and the possibility of accounting for potential state measurement errors were employed to forecast the 'low medicinal potential' state for a certain set of taxa while incorporating a measure of the a priori uncertainty into the prediction.

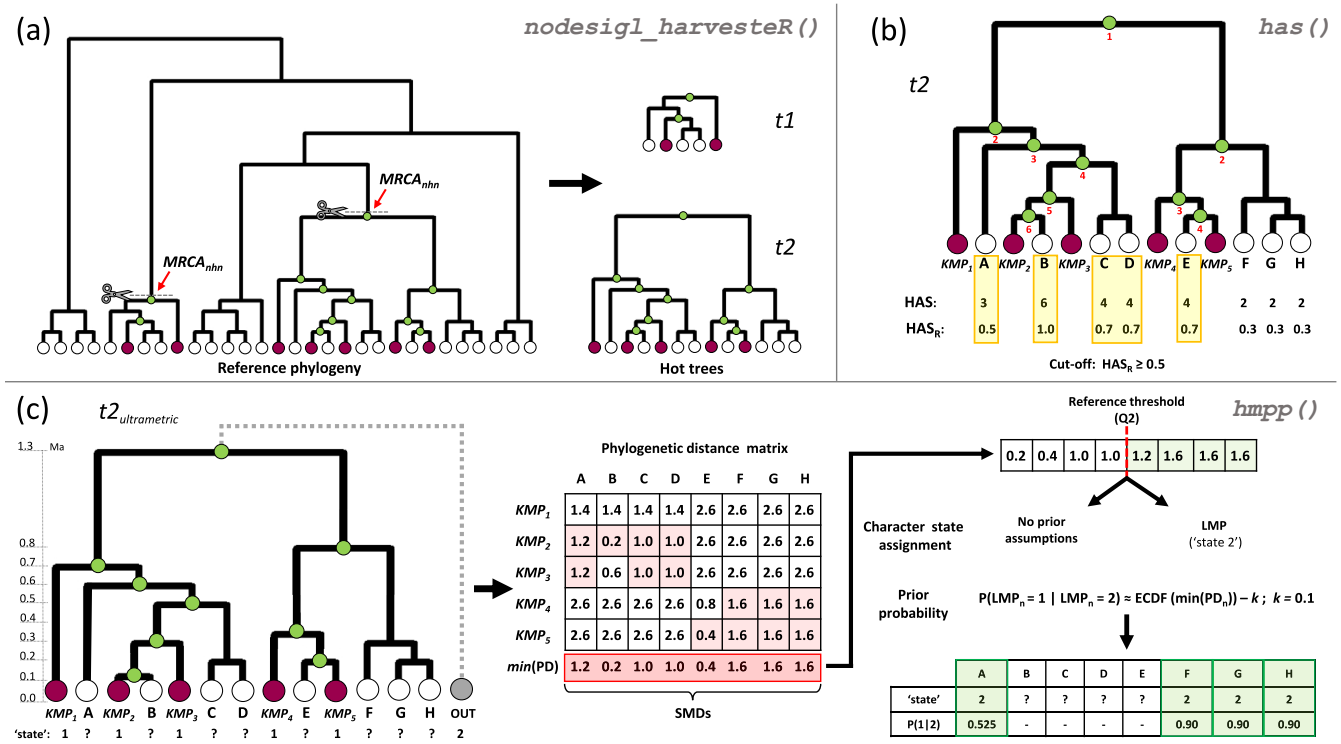

FIGURE 2 A schematic representation of the novel methods outlined in this paper, employing a simplified example to facilitate clarity and ease of understanding. Light green circles = stable hot nodes obtained from multiple *nodesigl* analysis iterations; burgundy circles = known medicinal plants; white circles = species with unknown character state. Panel (a) refers to the *nodesigl_harvesteR()* method. The figure shows the extraction of two hot trees (*t1* and *t2*) from the reference phylogeny subsequent to the identification of their corresponding independent sets of phylogenetically nested hot nodes. $MRCA_{nhn}$ = the most recent common ancestor of a set of phylogenetically nested hot nodes. Panel (b) illustrates an example of applying the *has()* method to the *t2* hot tree previously identified. The red numbers beneath the hot nodes indicate the level of phylogenetic nestedness of the individual nodes, with an increase from the root to the tips. KMP_{1-5} = known medicinal plants; A–H = species with unknown character state; HAS, Hot Ancestry Score. Only HAS for taxa with unknown status are displayed in figure; HAS_R, rescaled hot ancestry score, which is calculated by determining the ratio between the original value and the maximum HAS calculated in the tree (values rounded to the first decimal place). Taxa selected employing a HAS_R cut-off ≥0.5 are highlighted by yellow boxes. The *has()* method utilises solely tree topology information. Panel (c) provides a visual representation of the initial steps involved in the *hmpp()* method for preparing probability calculations. The figure on the left illustrates the ultrametric version of the *t2* hot tree. This was to simplify the visualisation. It is important to emphasise that in order to use the method, the branch lengths do not have to be proportional to time, but only to be present in the analysed tree. The character 'state 1' is assigned automatically to known medicinal plants (KMP_{1-5}), and the 'state 2' is designated to the randomly selected external outgroup (OUT, grey circle). '?' = unknown character state (taxa A–H). Time is expressed in million years ago (Ma). The central part of the figure exemplifies the construction of the phylogenetic distance matrix and the identification of the sample of minimum distances (SMDs). For each taxon with unknown character state the minimum phylogenetic distance (i.e. *min* [PD]) from its nearest KMP(s) is highlighted in red in the matrix and reported in the red box below that represents the SMDs. The right-hand side of the figure shows the process for designating the putative low medicinal potential (LMP) species (i.e. 'state 2') based on a predetermined phylogenetic distance threshold, incorporating an estimate of the uncertainty in the assignment. For purely illustrative purposes, the 50th percentile of the SMDs has been used as the reference threshold in the figure. The minimum phylogenetic distance from the nearest KMP is used as a proxy to estimate the degree of confidence with which the 'absence of status 1' has been 'measured' in putative LMP species (and hence the level of uncertainty with which they have been assigned 'status 2'). For this purpose, the quantile returned by the empirical cumulative distribution function (ECDF) built on the SMDs values, optionally decreased by a correction factor (*k*), is used as prior probability of the assigned state. The correction factor was set to 0.1 in the example. The values thus obtained (highlighted in green in the table in the bottom right corner) are subsequently utilised by the function *hmpp()* to calculate the final probabilities for all taxa in the tree, except the KMPs. The procedure described remains the same regardless of the criterion chosen (*single-threshold*, *multi-threshold*, *kmeans* or *pam*) to define the reference threshold. In the case of the *multi-threshold* criterion, the same procedure is repeated for each of the reference thresholds that have been set and the final probabilities are averaged over all computed values.

Given a hot tree, the proposed algorithm started calculating the phylogenetic ('patristic') distance matrix between taxa with unknown character state and the KMPs included in the tree. For each species with unknown state, the minimum distance among those calculated was selected, and all the values thus identified were used to define a 'sample of minimum distances' (SMDs). We then implemented four alternative procedures based on different criteria to identify taxa with putative low medicinal potential.

In the *single-threshold* criterion, a single user-defined phylogenetic distance threshold was applied by providing the probability corresponding to the desired sample quantile of the SMDs. Taxa with an unknown character state and a phylogenetic distance to the nearest KMP greater than the reference threshold were marked as having a (putative) low medicinal potential (i.e. 'state 2' in the function).

The *multi-threshold* criterion worked in a similar way but differed from the previous one in that it applied multiple user-defined and evenly spaced distance thresholds. These reference thresholds could be defined by providing the probabilities corresponding to the desired minimum and maximum sample quantile of the SMDs along with the interval between consecutive thresholds.

The *kmeans* and the *pam* criteria utilised the K-means and the PAM clustering algorithms implemented in the *stats* and *cluster* R packages, respectively (Maechler et al., 2025; R Development Core Team, 2023), to automatically determine the taxa to be included in the group of putative low medicinal potential species. First, the desired maximum number of clusters to be tested was specified (default setting $k = 10$). Then the optimal number of clusters was estimated using the 'average silhouette' method implemented in the *factoextra* R package (Kassambara & Mundt, 2020), and it was used to partition the SMDs accordingly. The species corresponding to the distances clustered around the largest centroid or medoid, depending on the chosen algorithm, were considered as putative low medicinal potential taxa.

Once the putative low medicinal potential species have been determined by the method selected by the user, the phylogenetic distance information was used to weight the uncertainty in the predicted states in the form of a priori probabilities. The function *hsp_binomial()* from the *castor* package allows accounting for potential state-measurement errors using the parameter *state1_probs*, keeping the 'state 1' (i.e. the known medicinal state) as the reference. To this end, for each putative low medicinal potential taxon, the distance to the nearest KMP was passed to the empirical cumulative distribution function constructed on the SMDs to obtain the corresponding percentile. The returned value was then used to express the probability of correctly measuring 'state 1' (conditional upon its true state and conditional it is a non-hidden state). In other words, the closer the distance between a putative low medicinal potential taxon and its nearest KMP was to the predefined distance threshold, the greater the likelihood for that taxon that its 'state 1' had been erroneously 'measured' (and thus wrongly assigned to 'state 2'). Optionally, it was possible to apply a correction factor to avoid assigning 'state 2' with certainty to taxa belonging to a hot tree (which might be desirable under certain circumstances).

To include at least one taxon with error-free 'state 2', we implemented an additional step. Starting from the hot tree under consideration, the proposed algorithm moved back one split in the reference phylogeny towards the root, thus identifying a new ancestral node. Among the tips descending from this node, a taxon was randomly chosen with the constraint that it did not belong to the hot tree nor to the set of KMPs for the disease of interest. The selected taxon was then merged with the hot tree at the previously determined ancestral node, which served as the new root, and the added outgroup was automatically marked as low medicinal potential species, assuming no state-assignment error. Finally, the algorithm assigned the 'state 1' to all KMPs, assuming an error-free measurement, while tips not assigned to any state up to this point were treated as having hidden states (Figure 2c).

From this point onwards, the calculation to predict unknown traits proceeded as implemented in the function *hsp_binomial()*, except when the *multi-threshold* method was used. In this case, the final probabilities were obtained by taking the average of all the calculated values with different distance thresholds. The described approach was implemented in the function *hmpp()*.

2.10 | Rescaling values

The values obtained by the HAS and HMPP methods were relative to a particular hot tree, and although suitable for comparing taxa within the same tree, they could not be used for comparisons between different trees. Therefore, for each metric, we also provided its rescaled version, obtained by dividing the original values calculated for a given hot tree by the maximum value calculated within that hot tree. As the KMP probabilities were fixed, KMPs were not included in the determination of the estimated maximum values in the HMPP analysis.

2.11 | Graphical outputs

We provided two functions to support the inspection and display of outputs. The function *visual_comp()* was used to plot and visually compare the probabilities obtained from the different criteria implemented in the function *hmpp()*. The function *hot_tree_painteR()* was utilised to graphically summarise all results related to a specific hot tree.

2.12 | Testing the effectiveness of new methods

Verifying the effective predictive ability of a method involves the possibility of testing a posteriori its predictions against new lines of evidence. This ideally requires that the question of interest is addressed by independent methods and that the results obtained are compared with the predictions made. The whole process normally takes a long time to complete. However, the CMAUP v2.0 database has recently been released (Hou et al., 2024), which incorporates more species than the first database release. These new species were included on the basis of independent research, thus offering the possibility of directly evaluating our approach more quickly. The same disease-related keywords used with the previous version of the database were also used to query the CMAUP v2.0 database, and new lists of medicinal plants were assembled, as explained above. The custom functions

dnu_CMAUPv2() and *pr2d_CMAUPv2()* were used to download the data of interest. Raw data downloaded from CMAUP v2.0 are available in Data S1(accessed on 6 November 2023).

Taxa already present in CMAUP v1.0 were therefore discarded from the new lists obtained from CMAUP v2.0. Then, for each disease, we defined two sets of newly identified taxa:

- *validation set #1*, including all new medicinal plants present in the reference phylogenies, used to assess the effectiveness of the *nodesigl_harvesteR()* method;
- *validation set #2*, including all new medicinal plants present in the identified hot trees with at least 100 taxa, used to assess the effectiveness of the HAS and HMPP methods.

As these two methods were primarily (although not exclusively) designed to guide prioritisation in medium and large trees, only hot trees with a minimum of 100 taxa were included in their validation sets (i.e. *cut_off* = 100 in both functions). This choice also facilitated the assurance of a minimum number of medicinal plants in each of the hot trees under consideration, a requirement necessary for the HMPP method to estimate probabilities at the tips in a sensible way.

In order to assess the effectiveness of the new methods, we applied a twofold validation strategy. First, the validation procedure was conducted through the implementation of randomisation tests. Subsequently, the performance of the HAS and HMPP methods was evaluated via post hoc tests with repeated stratified subsampling.

2.13 | Randomisation tests

For each disease, we calculated how many times the number of species from validation set #1 included in the identified hot trees (observed result) was less than or equal to that found in 9999 samples of the same size, randomly sampled from taxa included in the reference phylogenies (random results).

To define the set of target species identified by the HAS and HMPP methods, we proceeded as follows. For each combination of method and disease, all species with a rescaled value greater than or equal to an established cut-off were included in the corresponding set of target species.

The selected thresholds were 0.5 and 0.75 for the HAS and HMPP methods, respectively. These values were determined empirically based on available data, representing a compromise between the need to reduce the number of target species identified by the methods and the need to predict a reasonable number of new medicinal species in multiple diseases simultaneously. We next calculated how many times the number of species from validation set #2 included in the set of target species (observed result) was less than or equal to that found in 9999 samples of the same size, randomly sampled from the taxa present in the considered hot trees (random results). Throughout these analyses, the function *hmpp()* was called with the following additional parameters: *method* = 'multi_thr', *thr_level* = c(0.75, 0.95), *by* = 0.01, *min_revealed* = 2, *max_STE* = 0.35, cf = 0. All random samplings were conducted setting seed to 1.

For each test, the non-parametric *p*-value was calculated in order to test the significance of the method's prediction against the generated null distribution using the following formula:

$$p = T + 1/N + 1 \quad (1)$$

where *T* was the number of times the condition *observed result* ≤ *random result* was true within a null distribution of values obtained from *N* random samples. We added 1 to the numerator to assure that a *p*-value of 0 was never reached, thus providing a conservative estimation of *p*-values. For a 5% nominal alpha, *p*-values below 0.05 were taken as evidence of a significant result (one-tailed test). Since our primary interest was to focus on new potentially medicinal plants (PMPs), in all tests, KMPs obtained from CMAUP v1.0 were excluded prior to calculations from both the sets of target species identified by the methods tested and the species pools used to generate the reference null distributions.

Additionally, for each combination of method and disease, the observed result was compared with the number of taxa in the validation set that would be expected to be found randomly in a sample of the same size (Saslis-Lagoudakis et al., 2012). To this aim, the mean of the generated null distribution was taken as the expected value and was used to calculate the gain in percentage of positive hits compared with random.

2.14 | Post hoc validation tests

The lack of true negatives precluded the calculation of a confusion matrix and prevented us from using the metrics typically employed to evaluate predictive models. In order to circumvent this issue, the following strategy was implemented, which enabled the execution of post hoc validation tests based on repeated stratified subsampling. In accordance with the randomisation tests, the focus was exclusively on hot trees with more than 100 taxa, and the rescaled values predicted by the models were employed. All species that were not recognised as medicinal plants in neither CMAUP v1.0 nor in CMAUP v2.0 database were labelled as pseudo-negative cases, while species included in the validation set #2 were designated as true positive cases. The minimum distance between each pseudo-negative and the KMPs present in the same hot tree was calculated. The phylogenetic ('patristic') distance was used in the case of HMPP method, while the path distance (i.e. the number of branches separating two taxa) was used in the case of the HAS method. To account for the risk of pseudo-negative contamination (i.e. true positive being incorrectly labelled as a pseudo-negative), the resulting set of minimum distances was divided into four zones based on distance percentiles, defined as follows: Z1 = [0%, 30%], representing the maximum risk of contamination; Z2 = [30%, 55%], identifying a high risk of contamination; Z3 = [55%, 80%], reflecting a moderate risk of contamination; Z4 = [80%, 100%], indicating the lowest risk of contamination. Subsequently, the species

corresponding to each of the four zones were identified. The asymmetric zone boundaries were selected to reflect the measured distances, which, as expected, exhibited a bias towards small values. The validation was performed using 1000 repeated subsampling iterations, applying an overall imbalance ratio of ~1:9 between positive samples and pseudo-negative samples. This proportion was selected to replicate the conditions encountered during phylogenetic bioprospecting, where the number of medicinal plants is expected to be substantially lower than the number of non-medicinal species. In each iteration, a random subsample of 75% of the true positives was selected, while pseudo-negatives were randomly subsampled according to two different sampling schemes. In the first scheme, hereafter referred to as 'Favourable Case' (FC), species belonging to the zones Z1, Z2, Z3 and Z4 were subsampled to represent 10%, 25%, 30% and 35% of the pseudo-negative samples, respectively. In the second scheme, henceforth designated as 'Unfavourable Case' (UC), species belonging to the zones Z1, Z2, Z3 and Z4 were subsampled to represent 35%, 30%, 25% and 10% of the pseudo-negative samples, respectively. The FC was designed to mitigate the impact of pseudo-negatives contamination. In doing so, it mimics the evolution of a conserved trait characterised by a strong phylogenetic signal. Conversely, the UC was designed to exacerbate the effect of pseudo-negative contamination, thus mimicking the evolution of a labile/convergent trait. Any non-integer numbers produced by calculations in the subsampling procedure were rounded to their nearest integer.

For each combination of disease, model and sampling strategy, the average area under the precision–recall curve (AUC_{PR}), and its non-parametric 95% confidence interval (CI) were derived from 1000 subsampling iterations. This was done to measure the average enrichment and global prediction efficacy of new methods. By focusing exclusively on the positive class, whilst disregarding the large proportion of pseudo-negatives, and being particularly sensitive to false positives, the AUC_{PR} provides a direct measurement of 'the cost of searching'. This makes it a more suitable metric for imbalanced data analysis than other metrics, such as the area under the receiver operating characteristic (ROC) curve. Subsequently, to assess the operational utility of the tested models for candidate prioritisation, the associated lift charts were built by averaging the results across the subsampling iterations. The objective of a lift chart is to compare the model's predictions with a random selection and demonstrate the extent to which the model is more (or less) effective in identifying positive results within a given percentage of the species pool considered compared to random predictions. To this end, the values predicted by the methods were initially sorted in descending order and then grouped according to a subdivision into successive deciles. The average lift score and its 95% non-parametric CI were calculated for each decile based on the values obtained during the 1000 subsampling iterations. The R packages PRROC (Grau et al., 2015) and ROCit (Khan & Brandenburger, 2024) were utilised to calculate the AUC_{PR} values and the lift index within percentage groups.

3 | RESULTS AND DISCUSSION

3.1 | Preliminary results of phylogenetic signal tests

The results of the data curation process and the investigation of phylogenetic structuring of KMPs within the reference phylogenies are summarised in Table 1. Detailed results of all tests are available in Data S2.

Our results confirmed that the different metrics employed possess different sensitivity to phylogenetic clustering (with the δ statistic being the most restrictive), suggesting their combined use as a viable option in circumstances similar to those of this study. Overall, 22 combinations of reference phylogenies and diseases showed evidence of phylogenetic signals, most of which involved the Superasteridae and Superrosidae trees. Surprisingly, no evidence was found for the Monocotyledoneae reference phylogeny. This could be the outcome of a phylogenetically unbundled distribution of KMPs within monocots.

3.2 | Stability assessment of hot nodes

In the preliminary test carried out to estimate the optimal number of replications, 3000 *nodesigl* analyses (*−r* 999) were performed for each combination of the three investigated diseases and the five reference phylogenies. The number of hot nodes recorded in the first run was taken as a reference and the number of hot nodes that remained stable over the course of the repetitions was recalculated every 250 runs and expressed as a percentage of the initial reference value. Our analysis showed that the number of hot nodes became nearly constant regardless of the combinations studied after 1500 replicates, with sporadic residual reductions of less than 5% of the initial reference value (Figure S1). Accordingly, in all subsequent analyses, the number of replicates was fixed at 1500. Depending on the combination of disease and reference phylogeny, the loss of nodes before the steady phase varied between 11% and 12% and 50% of the initial number. This preliminary investigation yielded no definitive evidence to substantiate a correlation between the loss of hot nodes and tree size, nor between the loss of hot nodes and the ratio of the number of KMPs to the number of taxa in the phylogeny. Nevertheless, our findings strongly advised against relying on a single execution of the *nodesigl* command or a small number of its replicates, especially in the context of large phylogenies, and suggested that the stability of hot nodes should be carefully assessed. From this perspective, we suggest that utilising a greater number of independent replications, with a reduced number of randomisations (i.e. 999 randomisations per replicate), constitutes a better strategy in comparison to employing a smaller number of replicates with a larger number of randomisations.

TABLE 1 Distribution of known medicinal plants within the reference phylogenies after the data curation process and the results of the corresponding phylogenetic signal tests.

Disease	Reference phylogeny	KMPs	*D*-statistic	SES_{MPD}	SES_{MNTD}	δ statistic	Selected
Allergy	MO	105					
	SA	488					
	SR	396					
	MA	55					
	BE	49					
Alzheimer's disease	MO	141					
	SA	617					
	SR	524					
	MA	76					
	BE	67					
Arteriosclerosis	MO	83					
	SA	346					
	SR	283					
	MA	35					
	BE	28					
Colon cancer	MO	73					
	SA	384					
	SR	345					
	MA	42					
	BE	30					
Depression	MO	94					
	SA	416					
	SR	380					
	MA	53					
	BE	44					
Hyperglycaemia	MO	2	–	–	–	–	
	SA	11					
	SR	29					
	MA	1	–	–	–	–	
	BE	4	–	–	–	–	
Hypertension	MO	74					
	SA	266					
	SR	265					
	MA	35					
	BE	37					
Insomnia	MO	125					
	SA	563					
	SR	480					
	MA	72					
	BE	63					
Malaria	MO	58					
	SA	187					
	SR	167					
	MA	22					
	BE	32					

TABLE 1 (Continued)

Disease	Reference phylogeny	KMPs	*D*-statistic	SES_{MPD}	SES_{MNTD}	δ statistic	Selected
Prostate hyperplasia	MO	52					
	SA	221					
	SR	208					
	MA	25					
	BE	21					

Note: Light green indicates a significant value of one of the δ, SES_{MPD} and SES_{MNTD} statistics or the rejection of a random structure with a value of $D \leq 0.8$ in the case of the *D*-statistic; red = all other cases; dark green (last column) = indicates the 22 combination of reference phylogenies and diseases selected for further analyses; '—' = no analysis was performed as the number of KMPs in the reference phylogeny was less than 10. For detailed test results, see Data S2.

Abbreviations: BE, basal eudicots; KMPs, number of known medicinal plants; MA, Magnoliidae; MO, Monocotyledoneae; SA, Superasteridae; SES_{MNTD}, standard effect size mean nearest taxon distance; SES_{MPD}, standard effect size mean pairwise distance; SR, Superrosidae.

3.3 | Detecting hot trees

For each disease, only hot nodes that were consistently present in all replicates of the *nodesigl* analysis were kept. These stable hot nodes were further filtered based on the percentage of KMPs included in their descending tips. First, we apply a 5% threshold setting the *fract* parameter to 5 in the function *nodesigl_harvesteR()*. This approach worked well with most of the 22 studied phylogeny–disease combinations. However, seven of the analysed combinations included large hot trees among those identified (with the number of tips ranging from 1338 to 7788). For those seven combinations, we increased the KMPs threshold to 10%. The entire process resulted in the identification of 464 hot trees, containing a total of 2791 hot nodes (Table S2). Of the 464 hot trees, 290 had less than 15 tips, 60 comprised more than 100 taxa, and the largest had 868 species. Newick files of all identified hot trees and the plotted version of the 174 hot trees with more than 14 taxa are available in Data S2. Additional filtering performed on the stable hot nodes was done to reduce the size of the identified hot trees by focusing on hot nodes located within the regions of the reference phylogenies where KMPs were most clustered. The search for a balance between information retrieval and practical usability was achieved at the expense of some of the results obtained from the *nodesigl* analysis. However, the majority of the loss was limited to deeper hot nodes, and the approach taken allowed the preservation of most information-rich clades with less than 1000 taxa. Notably, the switch from considering hot nodes to hot trees has enabled the preservation of additional information embedded in the reference phylogenies. This included the hierarchical pattern of hot nodes and branch lengths that were then used to identify the most promising subsets of PMPs among those present in the hot trees according to the new HAS and HMPP methods, respectively (see below).

3.4 | Evaluating the performance of new methods

When the same keywords used to query the previous version of the database were used in CMAUP v2.0, no matches were found for 'atherosclerosis' and 'hyperglycaemia' (Table S3). This was due to the fact that the disease definitions used in the two versions of the database did not fully overlap. Consequently, the effectiveness of the proposed methods was assessed using data on the remaining eight diseases.

3.5 | Randomisation tests

First, we verified whether the results obtained from the function *nodesigl_harvesteR()* were able to predict the species in validation set #1 significantly better than random sampling. We therefore verified whether the HAS and HMPP methods could achieve significantly better results than random sampling using validation set #2, thus allowing us to take a further step forward in identifying the most promising candidate species. The performance of the three methods tested against the null random distributions generated for each disease is shown in Figure 3. Uncorrected probability values and the significance levels following the Benjamini and Hochberg (1995) adjustment method to control the false discovery rate across multiple comparisons are shown in Table 2. In all cases, the correction was performed within table columns (i.e. by method).

The tests conducted using the validation set #1 were significant for all the diseases studied (Table 2). This indicated that the identified hot trees effectively encompassed more medicinal plants than random samples of the same size extracted from the reference phylogenies. Overall, approximately 25% of the species included in the validation sets #1 was recovered by the identified hot trees (the percentage relative to each disease is shown in Table S4). The observed value was comparable to the average percentage of medicinal flora predicted by hot nodes shown in Saslis-Lagoudakis et al. (2012) (see Figure 2; C and D, 'among regions' comparisons, orange bars). Their results provided an appropriate comparison, as both the reference phylogenies and the validation sets used in our analyses included plants from different regions. On the other hand, the average percentage gain considering all eight diseases analysed was ~121%, ranging from ~66% for colon cancer to ~222% for prostate hyperplasia (Table S5). Remarkably, this average percentage was close to

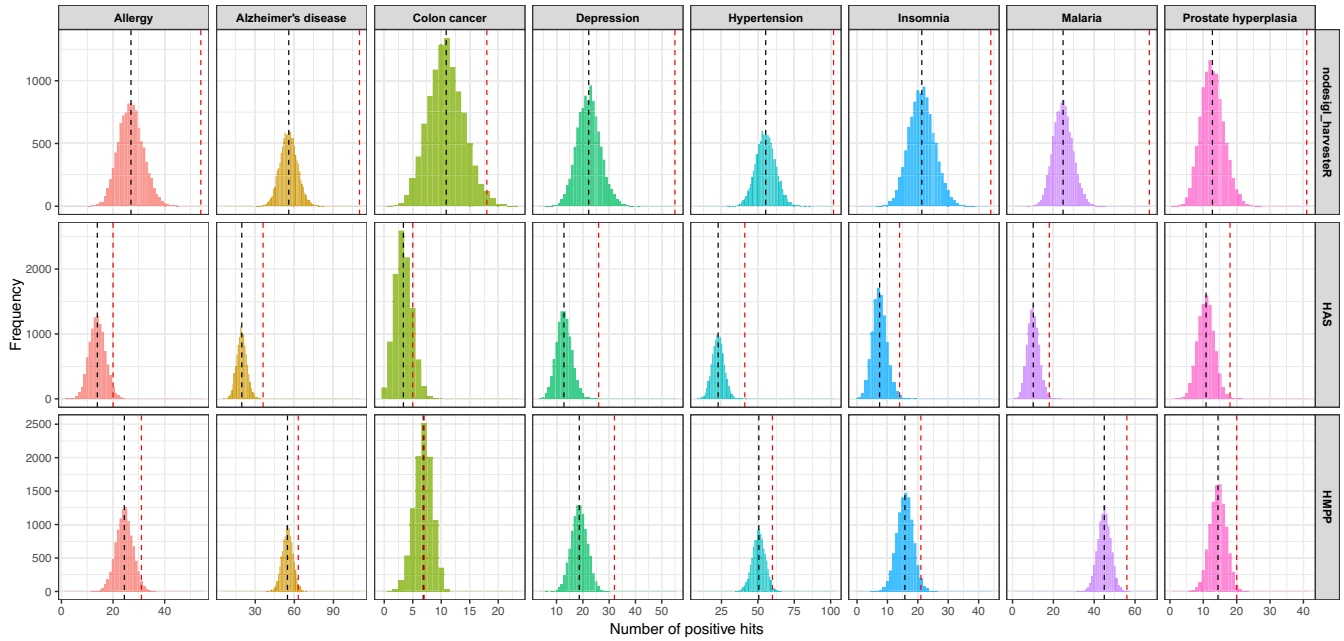

FIGURE 3 Randomisation tests to assess the performance of the proposed new methods for the diseases examined. The number of positive hits obtained by a method (i.e. the number of species belonging to the validation set that has been correctly predicted by that method) is shown as vertical dashed red line. Bell-shaped distributions represent null reference distributions generated by random sampling. The mean value of each null distribution is shown as vertical dashed black line. Each test was based on 9999 random replicates.

TABLE 2 Probabilities and significance levels of randomisation tests for each of the proposed new methods.

	nodesigl_harvesteR		*HAS*[a]		*HMPP*[a]	
Disease	***p***	**FDR**	***p***	**FDR**	***p***	**FDR**
Allergy	.0001	***	.0349	*	.0288	*
Alzheimer's disease	.0001	***	.0002	***	.0276	*
Colon cancer	.0225	*	.225	n.s.	.599	n.s.
Depression	.0001	***	.0001	***	.0001	***
Hypertension	.0001	***	.0002	***	.0207	*
Insomnia	.0001	***	.0072	**	.0336	*
Malaria	.0001	***	.0066	**	.0007	**
Prostate hyperplasia	.0001	***	.0033	**	.0144	*

Note: Each test was performed using 9999 random samples to generate its reference null distribution. [§] = only hot trees with ≥100 taxa were analysed; *p* = uncorrected *p*-value obtained in randomisation test (see the main text for calculation details). FDR = significance level after applying the Benjamini & Hochberg adjustment method for multiple comparisons to control the false discovery rate. In all cases the correction was performed within table columns (i.e. by method). Significance levels: *** = [0.0001, 0.001]; ** = [0.001, 0.01]; * = [0.01, 0.05]; n.s. = [0.05, 1]. Further details on tests are provided in Tables S5 and S6.

Abbreviations: HAS, Hot Ancestry Score; HMPP, Hidden Medicinal Properties Prediction.

[a]Only hot trees with ≥100 taxa were analysed.

the results reported by Saslis-Lagoudakis et al. (2012) for the data with plants grouped by condition categories (see Figure 2; B, 'within same region' comparisons, green bars). Hence, our findings indicate that the dual filtering procedure implemented in the *nodesigl_harvesteR()* function (i.e. stable hot node-based and informative hot node-based filtering), despite its substantial impact on reducing the number of finally retained hot nodes, yielded only a marginal decrease in the overall retrieval power of the approach. Moreover, the subsequent identification of the hot trees enabled the organisation of the remaining hot nodes into independent informative sets, thereby facilitating their interpretation. Ultimately, this strategy enhanced the reliability of the results and led to a noticeable gain in medicinal hits compared with random.

At the established cut-offs, both the HAS and HMPP methods predicted the species present in validation sets #2 better than random sampling for all diseases except colon cancer (Table 2).

The observed result can likely be attributed to the reduced overlap between the taxa in the CMAUP v2.0 database and those in the reference phylogenies, resulting in smaller validation sets for colon cancer than for other diseases (Table S6). Nonetheless, the result obtained with only one cut-off value per method was noteworthy, as theoretically any disease could require a targeted setting. Both the HAS and HMPP methods reduced the set of target species with respect to the function *nodesigl_harvesteR()*, always resulting in a gain in percentage of positive hits compared with random sampling. In significant tests, the magnitude of the percentage gain varied depending on the disease, ranging from ~44% to ~103% and from ~15% to ~76% for the HAS and HMPP methods, respectively (Table S6 and Figure S2). Surprisingly, an increase in percentage of positive hits was also observed in the non-significant test (HAS: ~49%; HMPP: 2%), highlighting a weak concentration trend for the species included in validation set #2 (Table S6 and Figure S2). This suggested that more targeted cut-off values might be necessary in this case. We repeated the analysis with the colon cancer data assuming stricter cut-off values for the calculated ratios of HAS and HMPP (i.e. 0.75 and 0.95, respectively) and indeed achieved significant results (HAS: $p = .0006$; HMPP: $p = .0427$; Figure S3).

Taking the output of the *nodesigl_harvesteR()* function as a reference, both HAS and HMPP methods proved effective in identifying reduced sets of target species with a high concentration of medicinal species. However, the relative performance of the two methods could not be easily compared directly. A first challenge in comparing results from the two methods was that, since they were based on different criteria, the same cut-off value had a different level of stringency on them, making comparison difficult. Secondly, although our results showed a higher gain in percentage of positive hits for HAS than for HMPP in the tests performed, the HMPP method was consistently superior when the positive hits were taken as the yardstick (Table S6 and Figure S2). The malaria case study was particularly illustrative in this regard, as from an initial validation pool of 60 taxa, the HMPP method predicted 56 species (~93%), while the HAS method identified only 18 (~30%; Table S6 and Figure S2). Indeed, in many cases the taxa in validation set #2 predicted by the HAS method were a subset of those predicted by the HMPP method (Figure S4). A further aspect to consider when comparing HAS and HMPP methods is that the sets of new PMPs identified by the two approaches (which should be their main focus) were not completely overlapping, probably reflecting the inherent differences in the criteria underlying the two methods (Figure S5).

3.6 | Post hoc validation tests

The outcomes of the post hoc validation assessments conducted on the HAS and HMPP methods are presented in Table 3 and Figure 4. The results obtained with the FC sampling strategy suggest that both new methods provide considerable overall predictive power and a relevant average enrichment in the true positive class when a phylogenetic signal is present, regardless of the disease considered. In particular, the HAS method demonstrated a global average percentage increase over the null model, which was superior to the HMPP method (Table 3). From this perspective, it is noteworthy that the threshold-independent approach adopted during post hoc testing produced results consistent with those obtained in randomisation testing, which instead used a pre-established cut-off threshold. As expected, the UC sampling strategy proved to be more challenging. However, even in this case, the HAS method demonstrated a high level of overall performance in all the diseases considered, while the HMPP method proved effective in the cases of colon cancer, depression and prostate hyperplasia (Table 3).

A more thorough assessment of the performance of the HAS and HMPP methods was facilitated by the analysis of lift charts, which provided additional information (Figure 4). The ideal model should demonstrate effectiveness in concentrating positive cases within a smaller portion of the data. This would assist in determining what percentage of the highest-scoring species should be targeted to optimise the return on effort in research in terms of correctly identified true positives. Our findings revealed that both methods exhibited the highest lift scores at the first or second decile (i.e. the deciles characterised by the highest-scoring predictions), after which there was a decline in lift scores. The most evident exception to this general pattern was represented by malaria in combination with the HAS method. A comprehensive comparison of the graphs presented in Figure 4 revealed that the HMPP method outperformed the HAS method in the initial decile of model predictions, with a lift score ranging from ~2.8 (malaria) to ~6.3 (colon cancer). Only in the case of hypertension the two methods achieved comparable outcomes. Two points are worthy of note. Firstly, the findings obtained from the implementation of the two sampling strategies demonstrated consistent results. This observation suggests that the confounding effect of pseudo-negative contamination, simulating the evolution of a labile/convergent trait (i.e. with week or none phylogenetic signal), did not impact the performance of the method in the initial decile of its predictions. Secondly, when the focus was exclusively on the top-ranking predictions, the HMPP method yielded useful results, even for the malaria (lift score: ~2.8) and hypertension (lift score: ~3.3) diseases, which were the two most recalcitrant diseases for this method.

Taken together, these results indicate that the HAS method, thanks to its overall effectiveness, is ideal for a preliminary screening aimed at identifying an initial pool of candidate species. On the other hand, the HMPP method has shown to be highly effective and robust in targeting top-ranking species. These findings support its use in prioritising target species, further narrowing the candidate pool and directing attention towards taxa with high predicted but still unexplored medicinal potential while reducing overall effort (see next paragraph).

3.7 | HAS and HMPP methods: Practical guidelines and remarks

In this paper, we have presented two methods called HAS and HMPP useful to effectively narrow down the species encompassed by hot trees, focusing on the species with the highest medicinal potential.

TABLE 3 Results of post hoc validation tests conducted for the Hot Ancestry Score and the Hidden Medicinal Plant Prediction methods.

Sampling strategy	Method	Disease	*N*	Mean AUC_{PR}	95% CI	Mean % improvement
FC	HAS	Allergy	330	0.2287	0.161–0.324	107.9
	HAS	Alzheimer's disease	620	0.2114	0.167–0.264	92.2
	HAS	Colon cancer	80	0.4275	0.196–0.633	288.6
	HAS	Depression	280	0.3658	0.258–0.479	232.5
	HAS	Hypertension	630	0.2281	0.176–0.286	107.4
	HAS	Insomnia	220	0.2647	0.174–0.36	140.6
	HAS	Malaria	450	0.2563	0.189–0.325	133.0
	HAS	Prostate hyperplasia	190	0.2258	0.162–0.325	105.3
	HMPP	Allergy	330	0.1530	0.131–0.179	39.1
	HMPP	Alzheimer's disease	620	0.1637	0.146–0.183	48.8
	HMPP	Colon cancer	80	0.2014	0.135–0.299	83.1
	HMPP	Depression	280	0.2205	0.182–0.266	100.5
	HMPP	Hypertension	630	0.1487	0.132–0.17	35.2
	HMPP	Insomnia	220	0.1700	0.145–0.203	54.5
	HMPP	Malaria	450	0.1284	0.115–0.143	16.7
	HMPP	Prostate hyperplasia	190	0.1970	0.154–0.247	79.1
UC	HAS	Allergy	330	0.1639	0.123–0.217	49.0
	HAS	Alzheimer's disease	620	0.1559	0.129–0.19	41.7
	HAS	Colon cancer	80	0.3558	0.151–0.575	223.5
	HAS	Depression	280	0.2529	0.181–0.343	129.9
	HAS	Hypertension	630	0.1681	0.137–0.205	52.8
	HAS	Insomnia	220	0.1856	0.128–0.26	68.7
	HAS	Malaria	450	0.1869	0.139–0.243	69.9
	HAS	Prostate hyperplasia	190	0.1662	0.13–0.22	51.1
	HMPP	Allergy	330	0.1118	0.098–0.127	1.6
	HMPP	Alzheimer's disease	620	0.1163	0.105–0.128	5.7
	HMPP	Colon cancer	80	0.1496	0.105–0.22	36.0
	HMPP	Depression	280	0.1625	0.138–0.192	47.7
	HMPP	Hypertension	630	0.1046	0.094–0.116	−4.9
	HMPP	Insomnia	220	0.1228	0.106–0.145	11.6
	HMPP	Malaria	450	0.0914	0.083–0.1	−16.9
	HMPP	Prostate hyperplasia	190	0.1431	0.115–0.175	30.1

Note: Each post hoc validation test was based on 1000 iterations of repeated stratified subsampling. The expected value was based on the prevalence of true positives (i.e. ~10%). Negative values indicate a decrease compared to the null model.

Abbreviations: 95% CI, the non-parametric 95% confidence interval of the AUC_{PR} value, calculated from 1000 subsampling iterations; FC, favourable case; HAS, Hot Ancestry Score; HMPP, Hidden Medicinal Properties Prediction; Mean % improvement, the average percentage improvement over the expected value, calculated based on 1000 subsampling iterations; mean AUC_{PR}, the mean area under the precision–recall curve, averaged over 1000 subsampling iterations; *N*, sample size used in each iteration (i.e. the sum of true positives and pseudo-negatives); UC, unfavourable case.

Moreover, by assigning a value to each species, both methods provide a criterion for prioritising the taxa under consideration. These approaches are based on different assumptions and produce not completely identical results. We regard these differences as an advantage because researchers may use one or both approaches according to their preferences. Depending on the objective of their study, researchers may consider merging or intersecting the sets of new PMPs identified by the HAS and HMPP approaches, or even using one method to refine results of the other. The cut-offs applied in randomisation tests (i.e. 0.5 for the HAS method and 0.75 for the HMPP method) have proven effective in most cases, with the caveat that they may need to be refined in some circumstances, such as the case of colon cancer data. We suggest that these cut-off values be used as preliminary exploratory analyses to gain an understanding of the available data. Subsequently, if deemed necessary, users can identify more targeted values based on their research objectives, available financial resources, the disease being studied, the method employed and the size of the reference phylogeny used. Increasing the value of

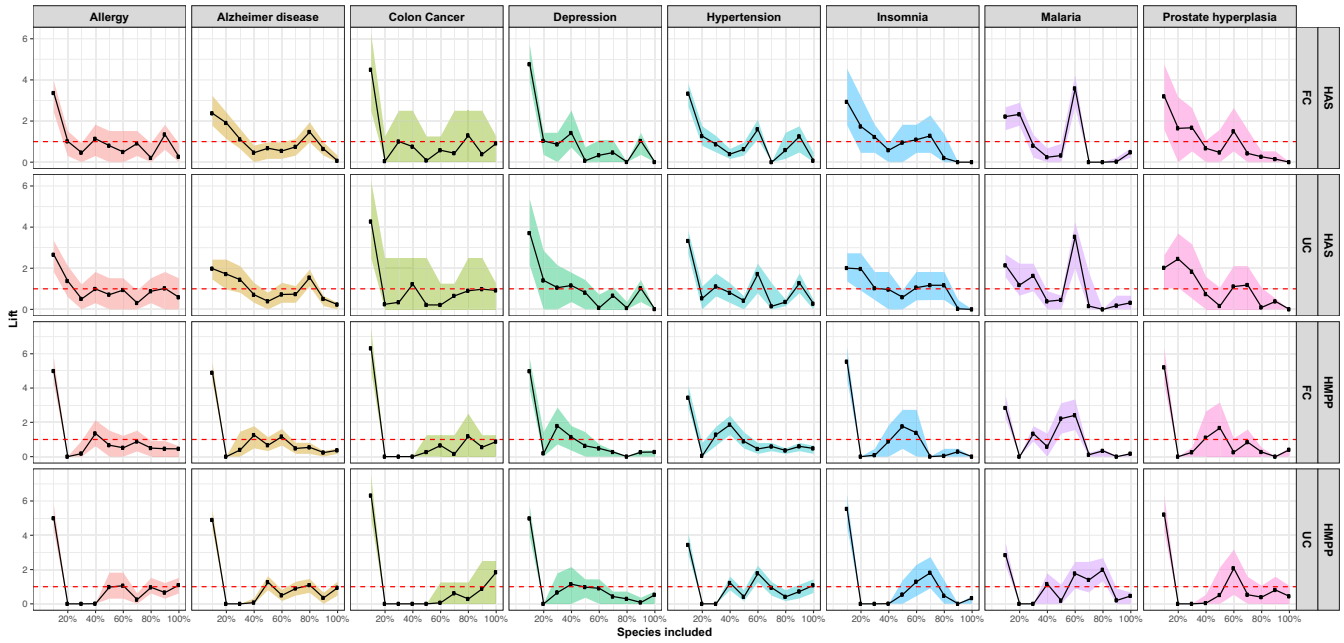

FIGURE 4 Lift charts utilised to assess the operational effectiveness of the Hot Ancestry Score (HAS) and Hidden Medicinal Plant Prediction (HMPP) methods in prioritising candidate medicinal plants. The charts were generated by calculating the mean lift index and its 95% non-parametric confidence interval (black dots and coloured area in the graphs, respectively) for each percentage class based on 1000 subsampling iterations. The deciles were used as percentage classes. Two distinct sampling strategies were employed: the favourable case (FC), to simulate the evolution of a conserved trait with a strong phylogenetic signal, and the unfavourable case (UC), to simulate the evolution of a labile/convergent trait. The expected result from the null model (i.e. random sampling) is indicated by a red dashed line in the graphs and corresponds to a lift index of 1 (i.e. ~10% of true positives predicted based on the prevalence of this class in the sampled cases).

these thresholds has the potential to render the methods more selective, resulting in smaller target species sets that are rich in species with high medicinal potential, according to the models. However, it should be noted that this process may also result in the exclusion of a proportion of genuinely medicinal plants from subsequent analyses. Conversely, the adoption of excessively lowered thresholds has the potential to engender diminished discriminatory capacity, thus giving rise to target species sets that are, in effect, too extensive to be analysed in their entirety. Indeed, in the context of large phylogenies, achieving a high level of selectivity often becomes a practical necessity in order to facilitate the effective allocation of resources to the most promising clades. In such cases, we propose two potential strategies for reducing the size of the target set.

i. In the first strategy, only those species that received the maximum support from either of the two methods are included in the final candidate set. This approach was illustrated in the case study presented in the next paragraph.
ii. The second strategy involves a two-step procedure that leverages the outcomes of post hoc tests. Initially, the HAS method is employed with a cut-off value greater than or equal to 0.5, depending on the desired level of selectivity, to perform an initial screening. Subsequently, among the species identified in the first phase, only those that have also obtained the highest scores from the HMPP method (for example, an HMPP ratio ≥0.99) are retained in the final selection of candidates.

It is not possible to assert that either of these two strategies is in general objectively superior to the other. Due to the potential for non-overlapping results, the selection of which to implement is dependent on the particular features of each study and must be evaluated on a case-by-case basis.

Lastly, a relevant implication of our findings relates to the circumstances under which HAS and HMPP methods can be successfully applied. While conducting bioprospecting in the presence of a clear phylogenetic signal increased the likelihood of ultimate success, our results suggest that, within certain limits, these new methods remain effective even in the presence of weak or absent phylogenetic signal. This would considerably broaden their applicability. However, this aspect deserves further exploration in future studies.

3.8 | A case study

To illustrate the potential of newly implemented methods, we used one of the hot trees identified by the function *nodesigl_harvesteR()* during the analysis of data on Alzheimer's disease and the reference phylogeny of Superasteridae. The selected hot tree encompassed 537 species, 67 of which were KMPs, and was obtained by setting the *fract* parameter to 10 to retain only stable hot nodes with at least 10% of KMPs included in their descending tips.

First, we executed the *hmpp()* function using the next four settings: (1) *method* = 'single_thr', *thr_level* = 0.75; (2) *method*

= 'multi_thr', *thr_level* = c(0.7, 0.95), *by* = 0.01; (3) *method* = 'kmeans', *kmax* = 10; (4) *method* = 'pam', *kmax* = 10. The remaining parameters were set as follows: *min_revealed* = 2, *max_STE* = 0.3, cf = 0. Then, the function *visual_comp()* was used to visually inspect the outputs produced by different methods implemented in the function *hmpp()*, showing that the *multi_thr* method was the one that displayed the greatest diversification among the calculated HMPP probabilities (Figure S6). Finally, the function *has()* was run to calculate the HAS values, and all results were plotted using the function *hot_tree_painteR()*. The input and output files of the analyses related to the illustrated case study are available in Data S3.

Figure 5 shows the selected hot tree with branches coloured according to the number of nested hot nodes from which they descend. For terminal branches, these values are equivalent to the HAS values calculated for their respective terminal taxa. Cool and warm colours represent the lowest and highest scores, respectively. Branches and arrows coloured in 'Burgundy' identify KMPs. Hot nodes are displayed as black dots, and the hot nodes that achieved the highest possible ranking across all replicates of the *nodesigl* analysis are shown as black stars. The numbers at the end of species names correspond to the computed HMPP probabilities.

The maximum HAS and the maximum probability calculated for the new PMPs were 16 and 1.0, respectively. Using the same cut-off values applied to evaluate the performance of the methods (i.e. HAS = 0.5; HMPP = 0.75), a reduced set of 227 new PMPs and a reduced set of 345 new PMPs were retained by the HAS and HMPP methods, respectively. These findings were a clear narrowing of the pool identified by the function *nodesigl_harvesteR()*.

To further illustrate the benefits of the proposed methods, we have focused our attention on the clades that comprise the species of

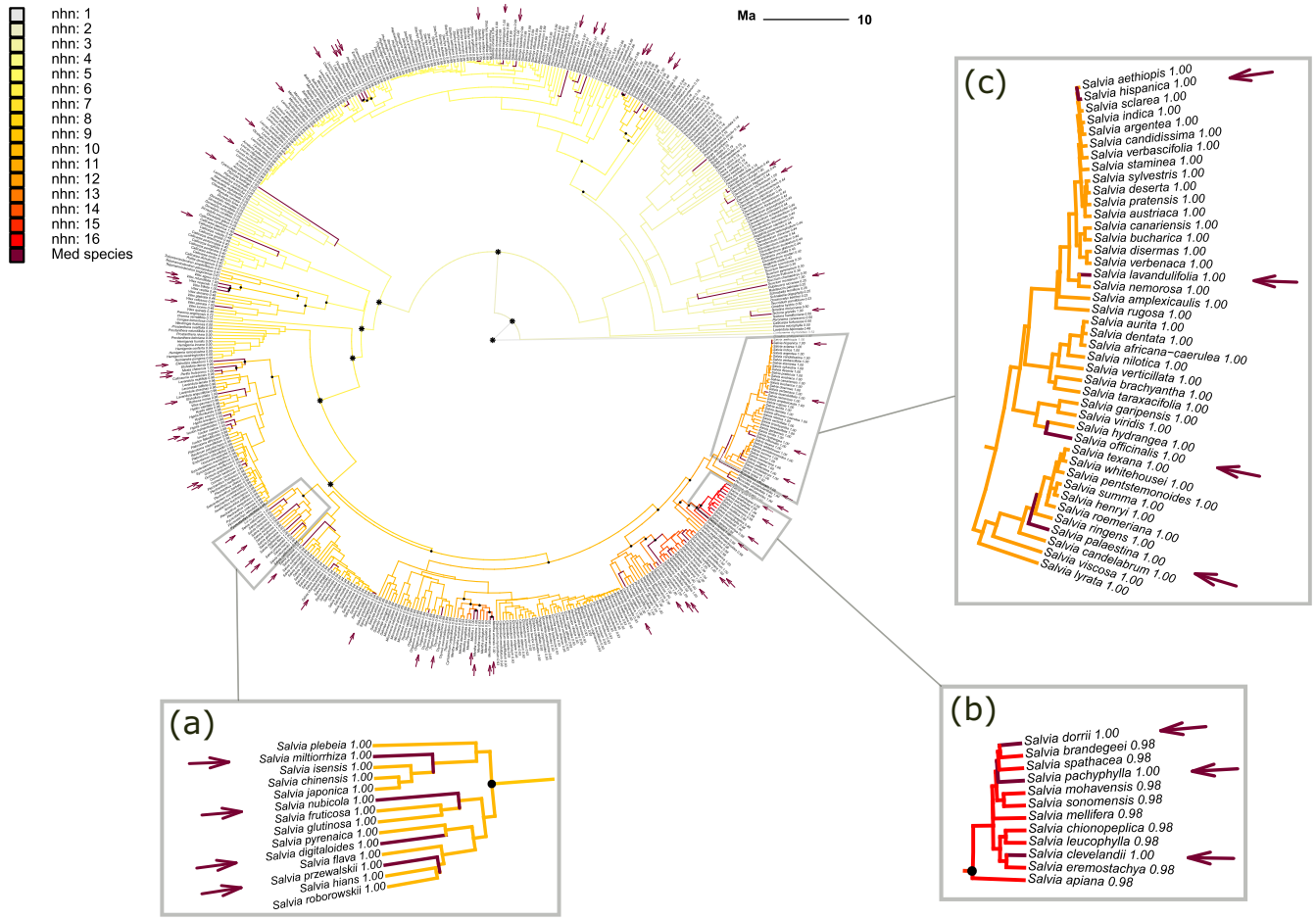

FIGURE 5 The *hot tree* selected for illustrative purposes. The tree was identified as relevant by cross-analysing Alzheimer's disease data and the Superasteridae reference phylogeny. Branches are coloured according to the number of nested hot nodes (referred as *nhn* in figure) from which they descend. For terminal branches, these values are equivalent to the Hot Ancestry Scores (HASs) calculated for the corresponding taxa. Cool to warm colours are used to represent the increasing number of nested hot nodes, as explained in the figure. Branches and arrows coloured in 'burgundy' identify already known medicinal plants (KMPs). Stable *hot nodes* and those among them that achieved the highest possible ranking across all replicates of the *nodesigl* analysis are shown as black dot and black stars, respectively. The numbers at the end of the species names correspond to the probabilities calculated using the Hidden Medicinal Plant Prediction (HMPP) approach. The grey boxes 'a', 'b' and 'c' illustrate an enlargement of the three distinct clades of genus *Salvia* encompassing the 57 species that received the most support either from the HAS or HMPP method.

the genus *Salvia*. *Salvia* is one of the largest genera of the family Lamiaceae consisting of more than 900–1000 species, and it is recognised to be polyphyletic or broadly paraphyletic (Walker & Sytsma, 2007; Will & Claßen-Bockhoff, 2017). Anti-Alzheimer's disease properties and cognitive-enhancing potential have been reported for several species of the genus, the best known of which are *Salvia miltiorrhiza* Bunge, *Salvia officinalis* L. and *Salvia lavandulaefolia* Vahl (Sharifi-Rad et al., 2018). The selected tree included 101 species of genus *Salvia*, 15 of which were KMPs. At the cut-off values applied, both the HAS and HMPP approaches identified all clades containing species of the genus *Salvia* as highly relevant for their medicinal potential. More importantly, both methods provided a support value for each species, thus enabling the further refinement of the pool of species with the highest medicinal potential. For example, if we were to consider only those species that received the maximum support from one or the other of the two methods, we would reduce the set of plants with the highest medicinal potential to 57 taxa divided into three distinct clades of 10, 9 and 38 species (clades 'a', 'b' and 'c' in Figure 5).

An additional advantage of our approach is particularly evident when dealing with non-monophyletic groups. Taxa from several other genera were found interspersed among the species of the genus *Salvia*, many of which received high support from our methods. Despite their high potential, these taxa would normally be overlooked in a screening based on a purely taxonomic approach. This is the case of *Melissa officinalis* L. (HAS = 0.75; HMPP = 1.0) and of the three species *Lepechinia fragrans* (Greene) Epling (HAS = 0.75; HMPP = 0.98), *Lepechinia calycina* (Benth.) Epling ex Munz (HAS = 0.75; HMPP = 0.98) and *Lepechinia chamaedryoides* (Balb.) Epling (HAS = 0.875; HMPP = 0.98). Extracts of *Melissa officinalis* have been shown to be useful in preventing the worsening of neuropsychiatric symptoms in patients with mild dementia due to Alzheimer's disease (Noguchi-Shinohara et al., 2020). The strong anticholinesterase activity, combined with the remarkable antioxidant and anti-inflammatory properties of the essential oils found in several species of the genus *Lepechinia*, have led some authors to propose their use as functional foods or as an adjuvant therapy for Alzheimer's disease (Calva et al., 2022; Panamito et al., 2021). These results were published after the release of the CMAUP v1.0 database. They confirmed the validity of the predictions made by our methods and highlighted the importance of using a broad phylogenetic perspective in bioprospecting.

4 | CONCLUSION

New tools for phylogeny-based bioprospecting of medicinal plants were introduced in this study. The presented approaches have proven to be effective in identifying species assemblages with high medicinal potential, and the HAS and HMPP methods were also able to formulate priorities among the identified taxa. Consistent with the terminology used for hot nodes and hot trees, we propose to use the term *hot species* for new PMPs identified by these methods that meet user-selected cut-off thresholds. As with any predictive method, the quality and completeness of the input data are crucial for accurate predictions. Users are therefore advised to critically examine their input data before performing the analysis. The selection of reference phylogeny is a particular aspect that necessitates meticulous evaluation. Along with Zanne et al.'s (2014) phylogeny used here for demonstration purposes, other mega-phylogenies have been published, for example, Smith and Brown (2018) and updates. We stimulate users to utilise the reference phylogeny that best suits their specific needs, constructing a customised one if necessary. Furthermore, several functions have user-settable parameters that can make the methods more or less stringent in the number of plants detected. There is therefore a trade-off between the practical need to reduce the final number of target species and the attempt not to leave out relevant taxa. In illustrating the new features introduced, we have provided a few examples of possible settings that have proven to be effective. However, we encourage users to find the parameter settings that best suit their purposes.

Overall, our methods represent a significant improvement over existing phylogenetic predictive techniques, and we expect that they will greatly accelerate the identification of new medicinal plants by narrowing the scope of subsequent laboratory analyses. In addition, HAS and HMPP methods, which provide a summary value for each species under analysis, also have the potential to be used to extract phylogenetic features from trees, which can then be integrated with multiple information sources and analysed directly with artificial intelligence-based methods, providing a further boost to medicinal plant identification. It is worth noting that, unlike the other two methods, the HMPP method can be applied alone, independent of *nodesigl* analysis, provided that a reference phylogeny and sufficient data on disease-related plants are available.

Finally, although our methods were developed with medicinal plant identification in mind, it is easy to imagine their potential application in other fields, for example, in the study of plant resistance to disease or in the bioprospecting of other organisms such as bacteria, fungi or animals.

AUTHOR CONTRIBUTIONS

Fabrizio Grassi, Massimo Labra and Giovanni Zecca conceived the ideas. Fabrizio Grassi and Giovanni Zecca designed the methodology. Giovanni Zecca developed and implemented the R package with contributions from Elisa Toini. Elisa Toini and Giovanni Zecca tested code components, collected and analysed the data. Elisa Toini, Fabrizio Grassi and Giovanni Zecca led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication. Fabrizio Grassi and Massimo Labra provided resources.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the anonymous referees for their constructive comments, which have resulted in a significant enhancement of the quality of the manuscript. Open access publishing facilitated by Universita degli Studi di Milano-Bicocca, as part of the Wiley - CRUI-CARE agreement.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

All code included in the *pm4mp* R package is freely available on GitHub at https://github.com/gzecca/pm4mp. The raw data and detailed results that support the findings of this study are available in Data S1–S3. Additionally, CMAUP v1.0 database can be accessed at https://bidd.group/CMAUP-2019/index.html; CMAUP v2.0 database can be accessed at https://bidd.group/CMAUP/; the complete phylogeny published by Zanne et al. (2014) can be downloaded from https://doi.org/10.5061/dryad.63q27.

ORCID

Giovanni Zecca https://orcid.org/0000-0003-2334-8897
Elisa Toini https://orcid.org/0000-0001-5668-9976
Massimo Labra https://orcid.org/0000-0003-1065-5804
Fabrizio Grassi https://orcid.org/0000-0003-3606-6469

REFERENCES

Albuquerque, I. P., Ramos, M. A., & Melo, J. G. (2012). New strategies for drug discovery in tropical forests based on ethnobotanical and chemical ecological studies. *Journal of Ethnopharmacology*, *140*, 197–201. https://doi.org/10.1016/j.jep.2011.12.042

Atienza-Barthelemy, D., Macía, M. J., & Molina-Venegas, R. (2024). Hot node limitations and impact of taxonomic resolution on phylogenetic divergence patterns: A case study on Ecuadorian ethnomedicinal flora. *Plants, People, Planet*, *7*, 644–653. https://doi.org/10.1002/ppp3.10594

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Borges, R., Machado, J. P., Gomes, C., Rocha, A. P., & Antunes, A. (2019). Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*, *35*, 1862–1869. https://doi.org/10.1093/bioinformatics/bty800

Buenz, E. J., Verpoorte, R., & Bauer, B. A. (2018). The ethnopharmacologic contribution to bioprospecting natural products. *Annual Review of Pharmacology and Toxicology*, *58*, 509–530. https://doi.org/10.1146/annurev-pharmtox-010617-052703

Calva, J., Cartuche, L., González, S., Montesinos, J. V., & Morocho, V. (2022). Chemical composition, enantiomeric analysis and anticholinesterase activity of Lepechinia betonicifolia essential oil from Ecuador. *Pharmaceutical Biology*, *60*, 206–211. https://doi.org/10.1080/13880209.2021.2025254

Crum, A. H., Philander, L., Busta, L., & Yang, Y. (2024). Traditional medicinal use is linked with apparency, not specialized metabolite profiles in the order Caryophyllales. *American Journal of Botany*, *111*, e16308. https://doi.org/10.1002/ajb2.16308

Davis, C. C., & Choisy, P. (2024). Medicinal plants meet modern biodiversity science. *Current Biology*, *34*, R158–R173. https://doi.org/10.1016/j.cub.2023.12.038

Domingo-Fernández, D., Gadiya, Y., Mubeen, S., Bollerman, T. J., Healy, M. D., Chanana, S., Sadovsky, R. G., Healey, D., & Colluru, V. (2023). Modern drug discovery using ethnobotany: A large-scale cross-cultural analysis of traditional medicine reveals common therapeutic uses. *iScience*, *26*, 107729. https://doi.org/10.1016/j.isci.2023.107729

Ernst, M., Saslis-Lagoudakis, C. H., Grace, O. M., Nilsson, N., Simonsen, H. T., Horn, J. W., & Rønsted, N. (2016). Evolutionary prediction of medicinal properties in the genus *Euphorbia* L. *Scientific Reports*, *6*, 30531. https://doi.org/10.1038/srep30531

Fathifar, Z., Kalankesh, L. R., Ostadrahimi, A., & Ferdousi, R. (2023). New approaches in developing medicinal herbs databases. *Database*, *2023*, baac110. https://doi.org/10.1093/database/baac110

Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *24*, 1042–1051. https://doi.org/10.1111/j.1523-1739.2010.01455.x

Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: Computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, *15*, 2595–2597. https://doi.org/10.1093/bioinformatics/btv153

Gurib-Fakim, A. (2006). Medicinal plants: Traditions of yesterday and drugs of tomorrow. *Molecular Aspects of Medicine*, *27*, 1–93. https://doi.org/10.1016/j.mam.2005.07.008

Hou, D., Lin, H., Feng, Y., Zhou, K., Li, X., Yang, Y., Wang, S., Yang, X., Wang, J., Zhao, H., Zhang, X., Fan, J., Lu, S., Wang, D., Zhu, L., Ju, D., Chen, Y. Z., & Zeng, X. (2024). CMAUP database update 2024: Extended functional and association information of useful plants for biomedical research. *Nucleic Acids Research*, *52*, D1508–D1518. https://doi.org/10.1093/nar/gkad921

Kassambara, A., & Mundt, F. (2020). factoextra: Extract and visualize the results of multivariate data analyses. https://doi.org/10.32614/CRAN.package.factoextra

Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., & Webb, C. O. (2010). picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, *26*, 1463–1464. https://doi.org/10.1093/bioinformatics/btq166

Khafagi, I. K., & Dewedar, A. (2000). The efficiency of random versus ethno-directed research in the evaluation of Sinai medicinal plants for bioactive compounds. *Journal of Ethnopharmacology*, *71*, 365–376. https://doi.org/10.1016/s0378-8741(00)00164-1

Khaiwa, N., Maarouf, N. R., Darwish, M. H., Alhamad, D. W. M., Sebastian, A., Hamad, M., Omar, H. A., Orive, G., & Al-Tel, T. H. (2021). Camptothecin's journey from discovery to WHO essential medicine: Fifty years of promise. *European Journal of Medicinal Chemistry*, *223*, 113639. https://doi.org/10.1016/j.ejmech.2021.113639

Khan, M., & Brandenburger, T. (2024). ROCit: Performance assessment of binary classifier with visualization. https://doi.org/10.32614/CRAN.package.ROCit

Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, *34*, 1053–1055. https://doi.org/10.1093/bioinformatics/btx701

Lulekal, E., Kelbessa, E., Bekele, T., & Yineger, H. (2008). An ethnobotanical study of medicinal plants in Mana Angetu district, southeastern Ethiopia. *Journal of Ethnobiology and Ethnomedicine*, *4*, 10. https://doi.org/10.1186/1746-4269-4-10

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2025). cluster: Cluster analysis basics and extensions. R package version 2.1.8.1. https://doi.org/10.32614/CRAN.package.cluster

Noguchi-Shinohara, M., Ono, K., Hamaguchi, T., Nagai, T., Kobayashi, S., Komatsu, J., Samuraki-Yokohama, M., Iwasa, K., Yokoyama, K., Nakamura, H., & Yamada, M. (2020). Safety and efficacy of *Melissa officinalis* extract containing rosmarinic acid in the prevention of Alzheimer's disease progression. *Scientific Reports*, *10*, 18627. https://doi.org/10.1038/s41598-020-73729-2

Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., Fritz, S., Isaac, N., & Pearse, W. (2023). caper: Comparative analyses of phylogenetics and evolution in R. R package version 1.0.3. https://doi.org/10.32614/CRAN.package.caper

Panamito, M. F., Bec, N., Valdivieso, V., Salinas, M., Calva, J., Ramírez, J., Larroque, C., & Armijos, C. (2021). Chemical composition and anticholinesterase activity of the essential oil of leaves and flowers from the Ecuadorian plant *Lepechinia paniculata* (Kunth) Epling. *Molecules*, *26*, 3198. https://doi.org/10.3390/molecules26113198

Pellicer, J., Saslis-Lagoudakis, C. H., Carrió, E., Ernst, M., Garnatje, T., Grace, O. M., Gras, A., Mumbrú, M., Vallès, J., Vitales, D., & Rønsted, N. (2018). A phylogenetic road map to antimalarial *Artemisia*

species. *Journal of Ethnopharmacology*, *225*, 1–9. https://doi.org/10.1016/j.jep.2018.06.030

Piel, W. H., Chan, L., Dominus, M. J., Ruan, J., Vos, R. A., & Tannen, V. 2009. TreeBASE v. 2: A database of phylogenetic knowledge. e-BioSphere. https://www.treebase.org/treebase-web/home.html

R Development Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reinaldo, R., Albuquerque, U., & Medeiros, P. (2020). Taxonomic affiliation influences the selection of medicinal plants among people from semi-arid and humid regions – A proposition for the evaluation of utilitarian equivalence in northeast Brazil. *PeerJ*, *8*, e9664. https://doi.org/10.7717/peerj.9664

Rønsted, N., Savolainen, V., Mølgaard, P., & Jager, A. K. (2008). Phylogenetic selection of *Narcissus* species for drug discovery. *Biochemical Systematics and Ecology*, *36*, 417–422. https://doi.org/10.1016/j.bse.2007.12.010

Rønsted, N., Symonds, M. R. E., Birkholm, T., Christensen, S. B., Meerow, A. W., Molander, M., Mølgaard, P., Petersen, G., Rasmussen, N., van Staden, J., & Stafford, G. I. (2012). Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. *Bmc Evolutionary Biology*, *12*, 182. https://doi.org/10.1186/1471-2148-12-182

Saslis-Lagoudakis, C. H., Klitgaard, B. B., Forest, F., Francis, L., Savolainen, V., Williamson, E. M., & Hawkins, J. A. (2011). The use of phylogeny to interpret cross-cultural patterns in plant use and guide medicinal plant discovery: An example from *Pterocarpus* (Leguminosae). *PLoS ONE*, *6*, e22275. https://doi.org/10.1371/journal.pone.0022275

Saslis-Lagoudakis, C. H., Savolainen, V., Williamson, E. M., Forest, F., Wagstaff, S. J., Baral, S. R., Watson, M. F., Pendry, C. A., & Hawkins, J. A. (2012). Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 15835–15840. https://doi.org/10.1073/pnas.1202242109

Sharifi-Rad, M., Ozcelik, B., Altın, G., Daşkaya-Dikmen, C., Martorell, M., Ramírez-Alarcón, K., Alarcón-Zapata, P., Morais-Braga, M. F. B., Carneiro, J. N. P., Borges Leal, A. L. A., Alves Borges Leal, A. L., Coutinho, H. D. M., Gyawali, R., Tahergorabi, R., Ibrahim, S. A., Sahrifi-Rad, R., Sharopov, F., Salehi, B., del Mar Contreras, M., … Sharifi-Rad, J. (2018). *Salvia* spp. plants-from farm to food applications and phytopharmacotherapy. *Trends in Food Science & Technology*, *80*, 242–263. https://doi.org/10.1016/j.tifs.2018.08.008

Smith, S. A., & Brown, J. W. (2018). Constructing a broadly inclusive seed plant phylogeny. *American Journal of Botany*, *105*, 302–314. https://doi.org/10.1002/ajb2.1019

Souza-Neto, A. C., Cianciaruso, M. V., & Collevatti, R. G. (2016). Habitat shifts shaping the diversity of a biodiversity hotspot through time: Insights from the phylogenetic structure of Caesalpinioideae in the Brazilian Cerrado. *Journal of Biogeography*, *43*, 340–350. https://doi.org/10.1111/jbi.12634

Thompson, J. B., & Hawkins, J. A. (2025). Phylogeny and bioprospecting: The diversity of medicinal plants used in cancer management. *Plants, People, Planet*, *7*, 147–158. https://doi.org/10.1002/ppp3.10566

Tu, Y. (2011). The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nature Medicine*, *17*, 1217–1220. https://doi.org/10.1038/nm.2471

van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *The R Journal*, *6*, 111–122. https://doi.org/10.32614/RJ-2014-011

Walker, J. B., & Sytsma, K. J. (2007). Staminal evolution in the genus *Salvia* (Lamiaceae): Molecular phylogenetic evidence for multiple origins of the staminal lever. *Annals of Botany*, *100*, 375–391. https://doi.org/10.1093/aob/mcl176

Wani, M. C., & Horwitz, S. B. (2014). Nature as a remarkable chemist: A personal story of the discovery and development of taxol. *Anti-Cancer Drugs*, *25*, 482–487. https://doi.org/10.1097/CAD.0000000000000063

Webb, C. O., Ackerly, D. D., & Kembel, S. W. (2008). Phylocom: Software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics*, *24*, 2098–2100. https://doi.org/10.1093/bioinformatics/btn358

Will, M., & Claßen-Bockhoff, R. (2017). Time to split *Salvia* s.l. (Lamiaceae) – New insights from old world *Salvia* phylogeny. *Molecular Phylogenetics and Evolution*, *109*, 33–58. https://doi.org/10.1016/j.ympev.2016.12.041

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry*, *64*, 3–19. https://doi.org/10.1016/s0031-9422(03)00300-5

Yessoufou, K., Daru, B. H., & Muasya, A. M. (2015). Phylogenetic exploration of commonly used medicinal plants in South Africa. *Molecular Ecology Resources*, *15*, 405–413. https://doi.org/10.1111/1755-0998.12310

Zaman, W., Ye, J., Saqib, S., Liu, Y., Shan, Z., Hao, D., Chen, Z., & Xiao, P. (2021). Predicting potential medicinal plants with phylogenetic topology: Inspiration from the research of traditional Chinese medicine. *Journal of Ethnopharmacology*, *281*, 114515. https://doi.org/10.1016/j.jep.2021.114515

Zamani, S., Fathi, M., Ebadi, M.-T., & Máthé, A. (2025). Global trade of medicinal and aromatic plants. A review. *Journal of Agriculture and Food Research*, *21*, 101910. https://doi.org/10.1016/j.jafr.2025.101910

Zaneveld, J. R. R., & Thurber, R. L. V. (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. *Frontiers in Microbiology*, *5*, 431. https://doi.org/10.3389/fmicb.2014.00431

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G., McGlinn, D. J., O'Meara, B. C., Moles, A. T., Reich, P. B., Royer, D. L., Soltis, D. E., Stevens, P. F., Westoby, M., Wright, I. J., Aarssen, L., Bertin, R. I., Calaminus, A., Govaerts, R., … Beaulieu, J. M. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature*, *506*, 89–92. https://doi.org/10.1038/nature12872

Zeng, X., Zhang, P., Wang, Y., Qin, C., Chen, S., He, W., Tao, L., Tan, Y., Gao, D., Wang, B., Chen, Z., Chen, W., Jiang, Y. Y., & Chen, Y. Z. (2019). CMAUP: A database of collective molecular activities of useful plants. *Nucleic Acids Research*, *47*, D1118–D1127. https://doi.org/10.1093/nar/gky965

Zhu, F., Qin, C., Tao, L., Liu, X., Shi, Z., Ma, X., Jia, J., Tan, Y., Cui, C., Lin, J., Tan, C., Jiang, Y., & Chen, Y. (2011). Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 12943–12948. https://doi.org/10.1073/pnas.1107336108

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zecca, G., Toini, E., Labra, M., & Grassi, F. (2026). Accelerating the prioritisation of plant species with underexplored medicinal potential: The *pm4mp* (Phylogenetic Methods for Medicinal Plants) R package. *Plants, People, Planet*, 1–19. https://doi.org/10.1002/ppp3.70189