Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Non-leakage machine learning reveals a hypothesis-driven aromatic medicine module in Mongolian prescriptions for cardiovascular disorders

Xiaomin Chen
Inner Mongolia University for Nationalities
Narisu Dai
Baotou Medical College
Murilige Chao
Inner Mongolia University for Nationalities
Suyaoletu Wu
Inner Mongolia Medical University
Hongying Wu
Inner Mongolia University for Nationalities
Erguna E
Inner Mongolia University for Nationalities
Furong Chen
Inner Mongolia University for Nationalities
Long Bao
192247219270163.com

Inner Mongolia University for Nationalities

Article

Keywords: Mongolian medicine, natural products, heart disease, traditional prescription, non-leakage machine learning, medicine co-occurrence, Heyi

Posted Date: October 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11099759/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Supervised learning studies of traditional prescriptions can be inflated when formula names, indications or function terms leak disease-label information into the model. We analysed 7839 Mongolian medicine prescription records, including 216 heart disease-related records, using a non-leakage feature space composed only of medicine composition, dosage form and prescription size. Formula names, indications, functions, sources and notes were excluded from the primary model. Heart disease-related prescriptions showed a concentrated aromatic-warming and mind-calming module rooted in Mongolian Heyi theory (a concept analogous to vital energy and mind-body regulation), centred on nutmeg, *Choerospondias* fruit, agarwood, clove and Aucklandia root, and powders accounted for 68.5% of these records. The non-leakage random forest achieved a high AUC (0.925 ± 0.014) and recall (0.958 ± 0.025), but precision (0.064 ± 0.004) and F1 score (0.119 ± 0.007) were low, indicating a high false-positive burden rather than practical diagnostic utility. L1-regularized logistic regression had lower AUC but better precision and F1 performance. From a natural-products perspective, the identified module links phenylpropanoids, flavonoids and sesquiterpene-rich aromatic medicines to a literature-supported hypothesis involving inflammation, oxidative stress and circulation regulation. This study tests the hypothesis that heart disease-related prescriptions in Mongolian medicine contain a stable and interpretable natural-product module that can be detected after removal of direct disease-label leakage. The identified module provides a rational basis for future validation of aromatic medicine-derived compounds, including phenylpropanoids, flavonoids and sesquiterpenes, in regulating inflammation, oxidative stress and cardiovascular homeostasis. Therefore, the proposed framework should be considered a hypothesis-driven knowledge-discovery strategy rather than a clinical prediction model.

Highlights

A non-leakage feature space excludes disease-label text and still discovers structure.

216 heart disease-related prescriptions converge on an aromatic-warming, mind-calming core.

The core module centres on nutmeg, fruit, agarwood, clove and Aucklandia root.

High AUC (0.925) masked a high false-positive burden (precision 0.064, F1 0.119).

The workflow is hypothesis-generating, pending expert and external validation.

1. Introduction

Conventional pattern-discovery approaches for traditional medicine prescriptions may appear more informative than they are because models can capture explicit label information rather than underlying prescription structure. When formula names, indications or function descriptions contain terms such as heart disease, palpitation, chest pain or Heyi, a classifier can obtain a high AUC without providing meaningful knowledge about medicine composition. Reporting standards and risk-of-bias tools for

prediction modelling emphasize transparent feature provenance, appropriate validation and cautious interpretation of model performance [1, 2].

The present study was designed to answer a narrower and more useful question: whether heart disease-related Mongolian medicine prescriptions contain stable, interpretable natural-product modules when obvious textual leakage is removed. The key Mongolian medicine concepts are translated here as Heyi (a fundamental concept in Mongolian medicine, analogous to vital energy and mind-body regulation) and aromatic-warming, meaning the use of aromatic herbs to warm, unblock and regulate circulation. Many high-frequency medicines in this tradition are also natural-product sources rich in phenylpropanoids, flavonoids, lignans, volatile terpenoids and sesquiterpenes, providing a literature-based bridge between ethnomedical theory and natural-products pharmacology.

To address this, we developed a workflow that combines (1) a conservative non-leakage feature set, (2) a grouped cross-validation strategy, and (3) interpretable modelling, aiming to identify stable medicine modules rather than to maximize predictive accuracy. We therefore excluded formula names, indications, functions, sources and notes from the primary model and retained only medicine composition, dosage form and prescription size. Cross-validation was grouped by normalized formula name to reduce training-test contamination from duplicated formula-name records.

This leakage-control strategy is deliberately conservative but not complete. Mongolian prescription corpora contain same-name/different-composition records, different-name/same-composition records, modified formulas and optical-character-recognition artefacts. Grouping by formula name can reduce identical-name leakage, but it cannot prevent composition-level leakage among highly similar formulas with different names. We therefore frame formula-name grouping as a minimum leakage-control strategy and recommend future validation based on Jaccard or MinHash clustering of medicine composition.

Mongolian medicine prescriptions for heart disease should also not be treated as a simple extension of Chinese medicine prescriptions for chest impediment or Tibetan precious-pill systems. Chinese formula studies commonly discuss Gualou Xiebai Banxia decoction in relation to warming Yang, resolving phlegm and relieving chest pain [3, 4], whereas Tibetan Rannasangpei-type studies often focus on complex precious-pill systems, cerebrovascular disease and neuroprotection [5, 6]. The corpus analysed here emphasized Heyi theory, powder dosage forms and dense aromatic-medicine combinations. This combination of ethnomedical theory, preparation form and natural-product chemistry gives the study a distinct context for Chinese Journal of Natural Medicines and related natural-products journals.

2. Materials and methods

2.1 Data source and label definition

The dataset was derived from a compiled formulary of Mongolian medicine prescriptions. After field cleaning, 7839 prescription records were retained. A total of 216 records were labelled as heart disease-

related according to the disease-category field, corresponding to 2.76% of the corpus. Each record contained formula name, alias, source, dosage form, function, indication, prescription composition and notes. Medicine names were normalized with reference to the Pharmacopoeia of the People's Republic of China and the Ministry of Health drug standards for Mongolian medicines [23, 24] before feature construction. The binary outcome in this study was whether the record belonged to the heart disease-related category.

2.2 Feature engineering and non-leakage principle

The primary feature space contained three feature classes: binary medicine-composition variables, one-hot dosage-form variables and the number of medicines in a formula. Formula names, aliases, indications, functions, sources and free-text notes were excluded from the primary model. Function and indication text embeddings were reserved for sensitivity analysis because these terms may represent both legitimate covariates and direct label hints.

During parsing, a small number of medicine-composition tokens retained an attached quantity fragment (for example, a gram value written immediately after the medicine name in the source text) and were extracted as separate dose-encoded variants alongside the corresponding plain medicine variable, rather than being merged during cleaning. These dose-encoded variants are reported as extracted in the interpretability results (Section 3.4) and should be interpreted as the same underlying medicine, not as a distinct ingredient.

2.3 Grouped validation and leakage control

Cross-validation was grouped by normalized formula name so that records sharing the same normalized name were assigned to the same fold. This reduces contamination from repeated formula-name records but cannot eliminate leakage arising from different names with highly similar compositions. A preliminary audit found many same-name records with different composition sets, which may reflect source variation, formula modifications or data-cleaning artefacts. Future analyses should group records by prescription-composition similarity, for example by assigning formulas with Jaccard similarity above 0.80 or 0.85 to the same validation group.

2.4 Models and performance metrics

L1-regularized logistic regression and random forests were compared. L1-regularized logistic regression was used to obtain sparse and directionally interpretable medicine contributions [7]. Random forests were used to capture non-linear feature combinations [8]. Model performance was assessed by AUC, average precision, F1 score, precision, recall, Brier score, confusion-matrix patterns and predicted-positive rate. Because the positive class represented only 2.76% of records, AUC was not considered sufficient evidence of practical usability.

2.5 Interpretation, co-occurrence and graph-aggregation baseline

Random-forest feature importance and SHAP-based rankings were used to summarize influential features [9]. L1-regularized logistic regression coefficients were used as a direction-aware complement because mean absolute SHAP values can obscure whether a feature increases or decreases the probability of the positive class. Medicine co-occurrence networks were used to visualize the internal structure of heart disease-related prescriptions. A previous description of the graph baseline as GraphSAGE-style was revised to 'medicine co-occurrence neighbourhood mean-aggregation baseline'. The baseline averages first-order co-occurrence neighbourhood features before logistic regression and does not contain trainable aggregation, non-linear message passing or multi-layer graph representation learning; therefore it is not a true GraphSAGE or graph neural network [10].

3. Results

3.1 Core medicine module in heart disease-related prescriptions

Among the 216 heart disease-related prescriptions, high-frequency medicines were concentrated in an aromatic-warming and Heyi-calming module. Nutmeg occurred in 141 prescriptions (65.3%), *Choerospondias* fruit in 106 (49.1%), agarwood in 94 (43.5%), clove in 85 (39.4%) and Aucklandia root in 66 (30.6%). These medicines were not isolated high-frequency items; they formed a stable core in medicine-pair and co-occurrence network analyses.

Table 1
High-frequency medicines in heart disease-related Mongolian medicine prescriptions.

Medicine	Frequency	Proportion of heart disease-related prescriptions (%)
Nutmeg (Rou-dou-kou; Myristica fragrans)	141	65.3
Choerospondias fruit (Guang-zao; Choerospondias axillaris)	106	49.1
Agarwood (Chen-xiang; Aquilaria spp.)	94	43.5
Clove (Ding-xiang; Syzygium aromaticum)	85	39.4
Aucklandia root (Mu-xiang; Aucklandia lappa)	66	30.6
Cardamom (Bai-dou-kou; Amomum spp.)	48	22.2
Asafoetida (A-wei; Ferula spp.)	46	21.3
Bai-yun-xiang (Mongolian medicine name; botanical source to be confirmed)	43	19.9
Chebulae Fructus (He-zi; Terminalia chebula)	42	19.4
Dried ginger (Gan-jiang; Zingiber officinale)	39	18.1
White sandalwood (Bai-tan-xiang; Santalum album)	37	17.1
Red salt (Chi-yan; halite/mineral medicine)	36	16.7
Safflower (Hong-hua; Carthamus tinctorius)	35	16.2
Long pepper (Bi-bo; Piper longum)	35	16.2
Aconite root (Cao-wu; Aconitum spp.)	32	14.8

Bai-yun-xiang is a Mongolian medicine name whose botanical source remains to be confirmed; it is retained in the table as recorded in the source corpus.

3.2 Dosage-form structure and data-quality correction

Powders were the dominant dosage form. Among the 216 heart disease-related records, 148 were powders (68.5%), 31 were pills (14.4%) and 28 were decoctions (13.0%); the remaining 9 records (4.1%) used other or mixed dosage forms. This pattern was consistent with the dense use of aromatic medicines in the corpus and differed from the decoction-centred narrative often seen in Chinese chest-impediment formula studies.

A data-quality audit corrected the interpretation of several Jiuwei Chenxiang San-related records. One source record contained a cleaning error in which two medicine names were merged, causing the number of medicines to be underestimated. Other records retained nine components, whereas one

record required confirmation against the original book image. These records were not forced into a uniform nine-medicine category; instead, they were retained with explicit data-quality flags.

3.3 Model performance: high AUC did not indicate practical screening value

The non-leakage random forest achieved AUC 0.925 ± 0.014, average precision 0.376 ± 0.078 and recall 0.958 ± 0.025. However, precision was only 0.064 ± 0.004 and the F1 score was only 0.119 ± 0.007. This pattern represents high-recall, low-precision failure rather than a high-performance screening model.

The confusion-matrix pattern showed that the random forest identified approximately 59 true positives per fold but also produced about 1018 false positives per fold. The model labelled 41.6% of test prescriptions as positive despite a true positive rate of only 2.76%. If used for screening, it would therefore return a large number of non-heart-disease prescriptions as candidates.

This failure mode still has a constructive use. The high recall of the random forest makes it suitable as a hypothesis-generating ranking tool before expert review, provided that it is not treated as a definitive classifier. In a practical workflow, the model could prioritize a calibrated top-ranked subset of candidate prescriptions for Mongolian medicine expert review, thereby reducing manual workload while keeping missed positives low. The exact review fraction, such as the top 10% or 20%, should be selected only after threshold calibration on an external or composition-clustered validation set.

L1-regularized logistic regression had a lower AUC (0.887 ± 0.041), but its precision (0.134) and F1 score (0.229) were higher than those of the random forest. This result shows that, under extreme class imbalance, model selection cannot rely on AUC and recall alone. Threshold tuning, calibration and cost-sensitive learning should be prioritized if the objective is practical candidate prioritization [11, 12].

Table 2
Model performance under non-leakage and sensitivity-analysis feature settings.

Model	AUC	AP	F1	Precision	Recall	Brier	Predicted positive (%)
Random forest	0.925 ± 0.014	0.376 ± 0.078	0.119 ± 0.007	0.064 ± 0.004	0.958 ± 0.025	0.066	41.6
L1-regularized logistic regression	0.887 ± 0.041	0.329	0.229	0.134	0.796	0.057	16.4
Random forest plus function-text sensitivity features	0.936	0.519	0.248	0.144	0.898	0.035	17.2

3.4 Positive and negative interpretable features

The revised interpretation reports both positive and negative discriminating features. Positive features included nutmeg, *Choerospondias* fruit, wild hare heart and agarwood, indicating enrichment in heart

disease-related prescriptions within this corpus. Negative features included Xiao-mi-la, Lv-rong-hao, Shui-chang-pu, licorice, Phellodendron bark, sulfur, rhubarb and related medicines, suggesting that these ingredients were more frequently associated with non-heart-disease formulas in the present feature space.

Negative coefficients do not mean that a medicine cannot be used in heart disease. They indicate only that, in this labelled corpus and feature representation, the medicine was less specific for the heart disease-related category. Licorice, for example, is a broadly used harmonizing medicine and therefore has weak discriminative value for this task.

Table 3
Direction-aware negative discriminating features from the L1-regularized logistic model.

Negative feature	Coefficient	Overall frequency	Heart-disease frequency	Proportion positive when present (%)
Xiao-mi-la	-2.18168	186	1	0.54
Lv-rong-hao	-1.88272	375	0	0.00
Shui-chang-pu	-1.51764	405	1	0.25
Licorice (Gan-cao)	-1.27551	743	10	1.35
Mao-lian-cai	-1.22526	35	1	2.86
Phellodendron bark (Huang-bai)	-1.13148	307	0	0.00
Sulfur (Liu-huang)	-1.12113	280	0	0.00
Rhubarb (Da-huang)	-1.11694	241	0	0.00
Bai-gu-mu	-1.01821	80	4	5.00
Cao-wu-hua	-0.94134	246	1	0.41

Xiao-mi-la, Lv-rong-hao, Shui-chang-pu, Mao-lian-cai, Bai-gu-mu and Cao-wu-hua are transcribed directly from the source corpus; their botanical identification is preliminary and requires expert verification against the original formulary before final submission. Licorice, Phellodendron bark, sulfur and rhubarb correspond to pharmacopoeial medicines with established botanical or mineral sources.

3.5 Function-text embedding sensitivity analysis

Adding function and indication text embeddings increased the random-forest AUC from 0.925 to 0.936, with fold-wise increments of 0.002, 0.020 and 0.011. Because only three folds were available, the permutation test was not significant (P = 0.333). The 0.011 AUC gain was therefore not interpreted as statistically significant evidence of leakage. A more cautious interpretation is that function-text embeddings carry strong discriminating information that may reflect either legitimate prescription-function associations or direct disease-label hints.

To visualize this association, the free-text function field was reduced to a small set of single- or double-character keyword tags (for example, Heyi, Heat, Pain, Palpitation/anxiety) and cross-tabulated against the high-frequency medicines (Fig. 4A). These tags are keyword-derived groupings used for exploratory visualization and sensitivity analysis; they are not a formal Mongolian medicine diagnostic classification and were not used as inputs to the primary non-leakage model.

3.6 Medicine co-occurrence neighbourhood mean-aggregation baseline

The medicine co-occurrence neighbourhood mean-aggregation baseline achieved an AUC of 0.893, close to the 0.887 AUC of L1-regularized logistic regression. This did not show a stable advantage. In the presence of only 216 positive samples and automatically derived labels, the corpus is not yet sufficient to support reliable training of a graph neural network such as GraphSAGE. Future graph-based modelling would require a larger corpus, expert-reviewed labels and external validation.

4. Discussion

4.1 Main finding and defensible claim boundary

The most reliable finding is not the model score but the convergence of three evidence streams: high-frequency analysis, co-occurrence structure and interpretable feature analysis all pointed to a core module centred on nutmeg, *Choerospondias* fruit, agarwood, clove and Aucklandia root. In Mongolian medicine theory, this module is consistent with an aromatic-warming, Heyi-calming and circulation-regulating prescription structure. In natural-products terms, it is also a coherent module of phenylpropanoid-, flavonoid- and sesquiterpene-rich aromatic medicines. The modelling conclusion must remain bounded: the random forest is useful for candidate ranking and knowledge discovery, but it is not suitable as a direct screening tool.

4.2 Comparison with Chinese medicine studies of chest impediment

Studies of Chinese medicine for chest impediment, stable angina and coronary heart disease frequently discuss Gualou Xiebai Banxia decoction, usually framed by warming Yang, resolving phlegm, relieving chest constraint and improving pain [3, 4]. The chemical and pharmacological discussion of this formula system often emphasizes saponins, alkaloids, polysaccharides and other constituents associated with phlegm-resolving and Yang-warming actions. In contrast, the high-frequency medicines in the present Mongolian medicine corpus were aromatic and Heyi-calming medicines, with a chemical emphasis on volatile oils, flavonoids, phenylpropanoids and lignan-like aromatic constituents. The difference is therefore not only theoretical; it reflects different natural-product utilization strategies and different preparation traditions.

4.3 Comparison with Tibetan Rannasangpei systems

Tibetan Rannasangpei and related seventy-flavour precious-pill systems are often discussed in modern studies in relation to cerebrovascular disease, neuroprotection and neurodegenerative conditions [5, 6]. Their complexity is typically associated with precious substances, mineral or metal-containing components, animal medicines and broad multi-target effects. The Mongolian medicine module observed here has a different chemical and pharmaceutical signature: it is dominated by organic small molecules from aromatic medicines, including volatile terpenoids, flavonoids and phenylpropanoids, and it is frequently delivered as powders. This pairing of small aromatic molecules with powder dosage forms may be relevant to rapid oral exposure and sensory-pharmacological effects, although absorption and blood-brain-barrier penetration remain hypotheses that require experimental testing.

4.4 Natural-products evidence is hypothesis-supporting rather than mechanism-validating

The core module can be translated into a testable natural-products hypothesis. Nutmeg contains phenylpropanoids and lignan-related aromatic constituents [13], *Choerospondias* fruit provides flavonoid-rich cardioprotective leads [14], and agarwood contains sesquiterpenes and related aromatic constituents, while the cited review concerns agarwood in periodontitis rather than the present formula [15]. Clove and eugenol add another phenylpropanoid-rich component with antioxidant and anti-inflammatory evidence [16]. We hypothesize that the combined module may act through convergent regulation of inflammation, oxidative stress, vascular tone and neurohumoral stress responses, which offers a biochemical interpretation of the Mongolian medicine concepts of aromatic-warming and Heyi-calming. This is a literature-grounded mechanism hypothesis, not a validated mechanism. Network pharmacology, molecular docking, metabolite profiling and targeted experimental validation should be added only when actually performed [17–22].

4.5 External validation and expert review are the next hard gate

The weakest part of the current study is the validation chain. At minimum, future work should include blinded review of disease-category labels by two Mongolian medicine experts, reporting Cohen's kappa; validation groups based on composition similarity rather than formula names alone; and an independent corpus from national or regional Mongolian medicine standards [24], hospital preparations or commonly used clinical formulas. Without these steps, the model may still partially learn source-book editorial structure rather than generalizable prescription principles.

5. Limitations

This study has seven major limitations. First, there were only 216 positive records, and the positive rate was 2.76%, which caused a high false-positive burden under the default random-forest threshold. Second, formula-name grouping reduces same-name leakage but cannot prevent different-name/same-composition or near-duplicate formula leakage. Third, synonymy, OCR errors and medicine-name normalization issues, despite reference to pharmacopoeial standards [23, 24], may still affect medicine

counts and co-occurrence structure. Fourth, negative discriminating features were inferred from L1-regularized logistic regression and should be checked against direction-aware SHAP values from tree models. Fifth, the AUC increase produced by function-text embeddings was not statistically significant and should not be described as definitive leakage evidence. Sixth, the co-occurrence mean-aggregation baseline is not a graph neural network and does not establish the superiority of graph learning. Seventh, no external validation, network-pharmacology workflow, docking analysis or wet-lab validation has yet been completed.

6. Conclusions

Using 7839 Mongolian medicine prescription records, this study identified 216 heart disease-related prescriptions and found a core medicine module centred on nutmeg, *Choerospondias* fruit, agarwood, clove and Aucklandia root. By excluding direct textual leakage features and reporting model failure modes, the study reframes high AUC as insufficient evidence of practical utility while preserving the model's value as a hypothesis-generating expert-review prioritization tool. The module also presents a coherent natural-products profile, linking phenylpropanoid-, flavonoid- and sesquiterpene-rich aromatic medicines to a literature-grounded hypothesis involving inflammation, oxidative stress and circulation regulation. The most appropriate positioning is therefore a non-leakage machine-learning workflow for discovering ethnomedical natural-product modules and generating testable mechanism hypotheses. Expert label review, validation grouped by composition similarity, external prescription validation, metabolite profiling and experimental pharmacology are needed before stronger translational claims can be made.

Declarations

Declaration of Competing Interest

The authors declare that they have no competing interests.

Funding

This work was supported by (1) the First-Class Discipline Scientific Research Special Project of Inner Mongolia Autonomous Region, "Analysis and Scientific Evaluation of Mongolian Medicine Property Theory in the Context of Big Data" (No. YLXKZX-NMD-001); (2) the Natural Science Foundation of Inner Mongolia Autonomous Region, "Research on Hemostatic Mongolian Medicine Medication Patterns and Knowledge Graph Construction Based on Data Mining Techniques" (No. 2026MS0631); and (3) the Science and Technology Plan Project of Inner Mongolia Autonomous Region, "2026 Inner Mongolia Autonomous Region Mongolian Medicine Standardization Key Laboratory — Terminology Compilation Research on Mongolian Medical History and Literature" (No. 2026KYPT0054).

Author Contribution

BL and CX conceptualized the study and developed the research framework. CX and BL designed the methodology, including the non-leakage machine-learning strategy and validation approach. CX, DN and CMR contributed to data curation, prescription database construction, data preprocessing and terminology normalization. CX performed computational analysis, model evaluation and visualization. CX prepared the original manuscript draft. BL, DN, CMR, WSLT, WHY, EEG and CFR contributed to manuscript review, interpretation of results and critical revision. BL supervised the project and coordinated the overall research process. All authors read and approved the final manuscript.

Acknowledgement

The authors thank colleagues in Mongolian medicine research and data management for valuable discussions and technical assistance during the preparation of this study.

Data Availability

https://doi.org/10.6084/m9.figshare.32906909

Code Availability

The model-training and figure-generation scripts have been deposited in the same Figshare repository under the DOI cited above, with versioned software dependencies included to ensure reproducibility.

References

1. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
2. Wolff, R. F. et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).
3. Chen, M. et al. A systematic review of the effectiveness and safety of Chinese herbal medicine formula Gualou Xiebai Banxia (GLXBBX) decoction for the treatment of stable angina pectoris. *Med. (Baltim).* **98** (51), e18375. 10.1097/MD.0000000000018375 (2019).
4. Lin, P. et al. Qualitative and quantitative analysis of the chemical profile for Gualou-Xiebai-Banxia decoction, a classical traditional Chinese medicine formula for the treatment of coronary heart disease, by UPLC-Q/TOF-MS combined with chemometric analysis. *J. Pharm. Biomed. Anal.* **197**, 113950. 10.1016/j.jpba.2021.113950 (2021).

5. Wu, P. et al. Rannasangpei is a therapeutic agent in the treatment of vascular dementia. *Evid. Based Complement. Alternat Med.* **2016**, 2530105. 10.1155/2016/2530105 (2016).
6. Du, Y., Wang, Y. L., Chen, L., Li, Q. E. & Cheng, Y. Anti-depressant-like effects of rannasangpei and its active ingredient crocin-1 on chronic unpredictable mild stress mice. *Front. Pharmacol.* **14**, 1143286. 10.3389/fphar.2023.1143286 (2023).
7. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R Stat. Soc. Ser. B*. **58**, 267–288 (1996).
8. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
9. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30**, 4765–4774 (2017).
10. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. *Adv. Neural. Inf. Process. Syst.* **30**, 1024–1034 (2017).
11. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
12. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
13. Ha, M. T. et al. Phytochemical and pharmacological properties of *Myristica fragrans* Houtt.: an updated review. *Arch. Pharm. Res.* **43**, 1067–1092 (2020).
14. Li, C. et al. Preventive effect of total flavones of Choerospondias axillaries on ischemia/reperfusion-induced myocardial infarction-related MAPK signaling pathway. *Cardiovasc. Toxicol.* **14** (2), 145–152. 10.1007/s12012-013-9238-7 (2014).
15. Xie, C. et al. Pharmacology and therapeutic potential of agarwood and agarwood tree leaves in periodontitis. *Front. Pharmacol.* **15**, 1468393. 10.3389/fphar.2024.1468393 (2024).
16. Batiha, G. E. et al. *Syzygium aromaticum* L. (Myrtaceae): traditional uses, bioactive chemical constituents, pharmacological and toxicological activities. *Biomolecules* **10**, 202 (2020).
17. Ru, J. et al. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *J. Cheminform*. **6**, 13 (2014).
18. Wu, Y. et al. SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping. *Nucleic Acids Res.* **47**, D1110–D1117 (2019).
19. Fang, S. et al. HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine. *Nucleic Acids Res.* **49**, D1197–D1206 (2021).
20. Daina, A., Michielin, O. & Zoete, V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res.* **47**, W357–W364 (2019).
21. Szklarczyk, D. et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).

22. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking. *J. Comput. Chem.* **31**, 455–461 (2010).
23. Chinese Pharmacopoeia Commission. *Pharmacopoeia of the People's Republic of China* (China Medical Science, 2020).
24. Pharmacopoeia Commission of the Ministry of Health of the People's Republic of China. Drug Standards of the Ministry of Health of the People's Republic of China: Mongolian Medicines [Chinese]. [Publisher and exact edition details require verification from the title page of the cited copy.] (1998).

Figures

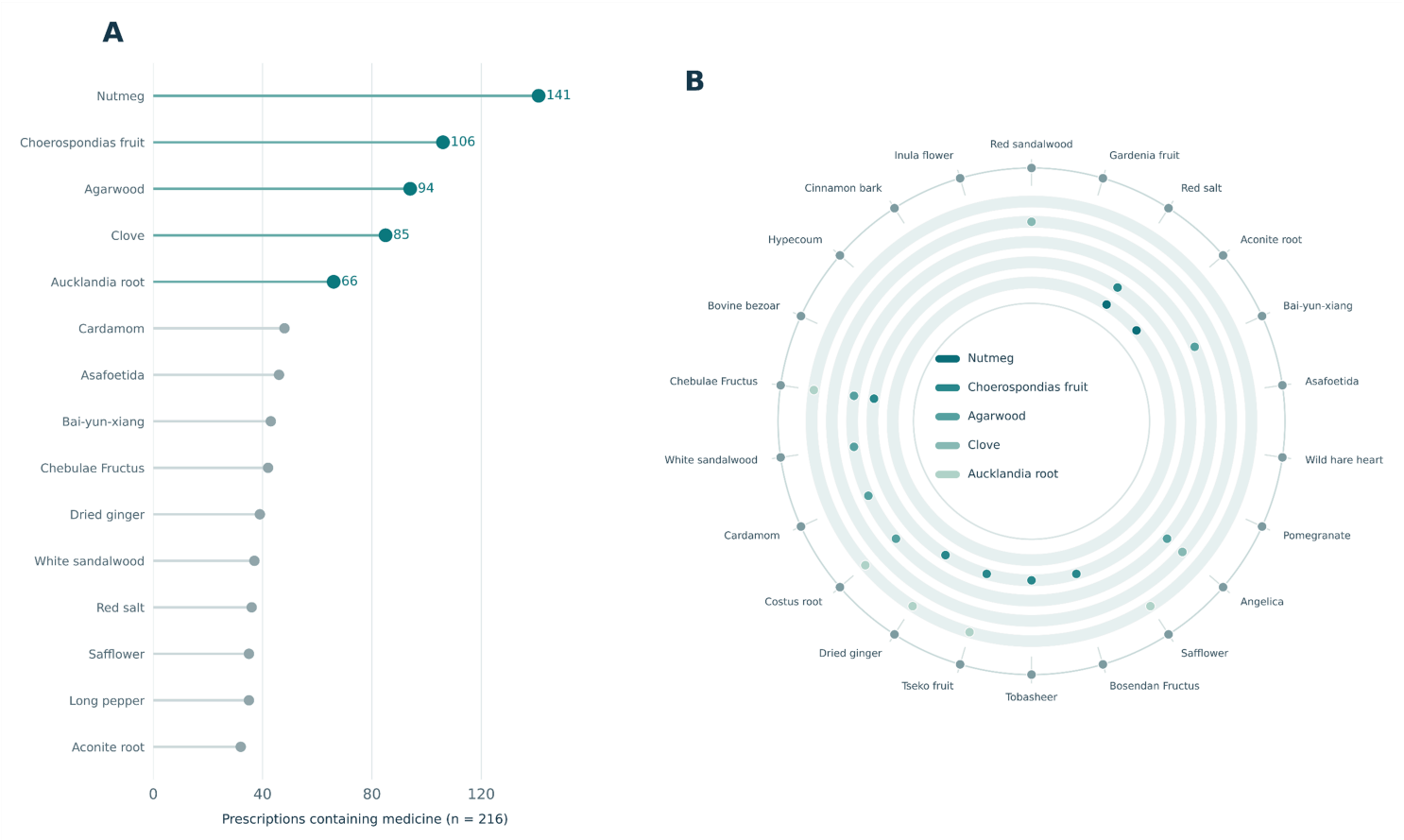


Figure 1

Medicine frequencies and the aromatic core. A, counts for the 15 most frequent medicines among 216 heart disease-related prescriptions. B, circular schematic of the five core medicines; circle size is proportional to individual frequency. The schematic does not encode pairwise co-occurrence strength.

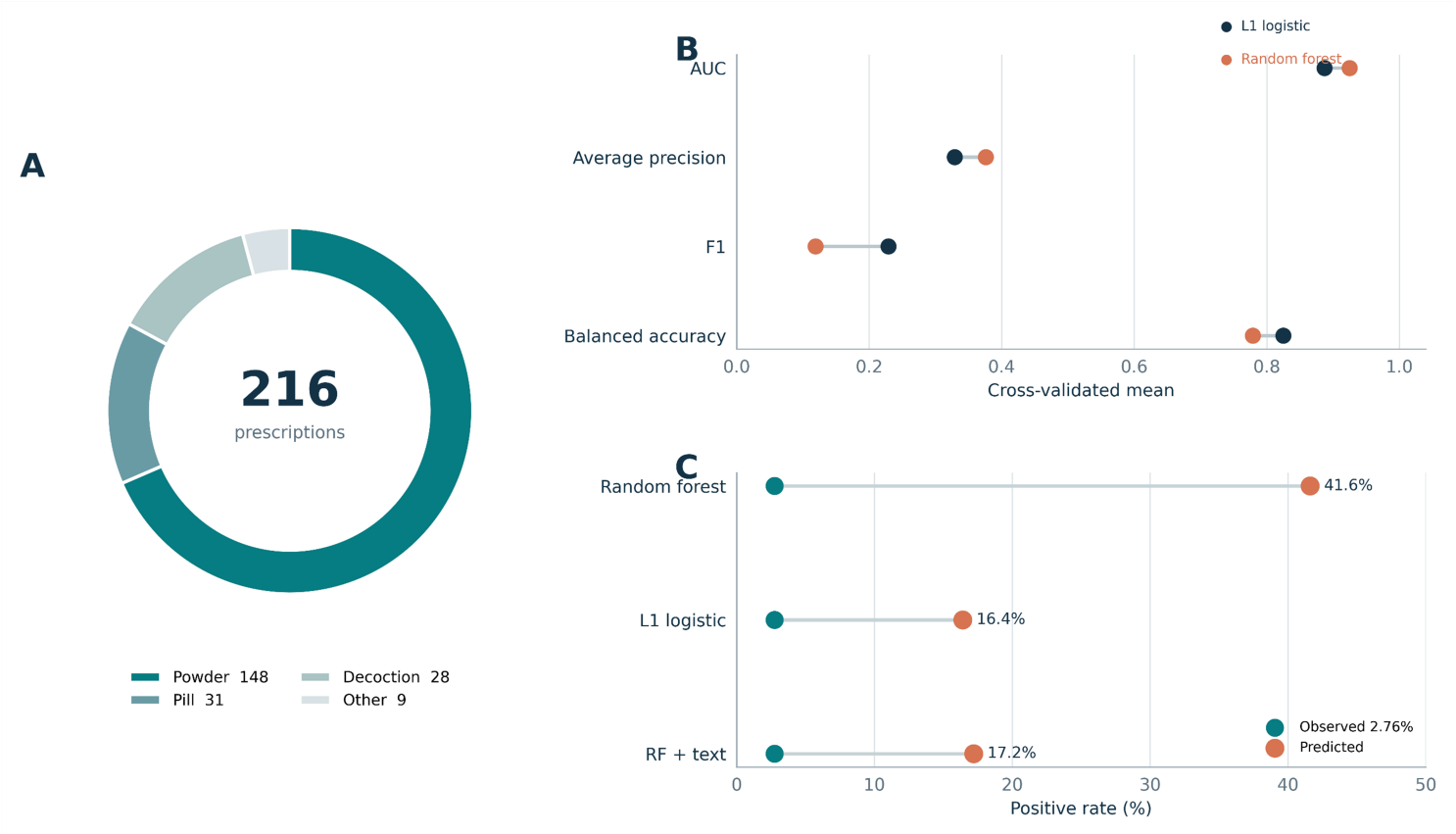


Figure 2

Dosage forms and model behaviour. A, distribution of dosage forms (other combines the remaining nine records). B, cross-validated model metrics from Table 2. C, observed positive prevalence compared with predicted positive rates. RF + text is a sensitivity analysis.

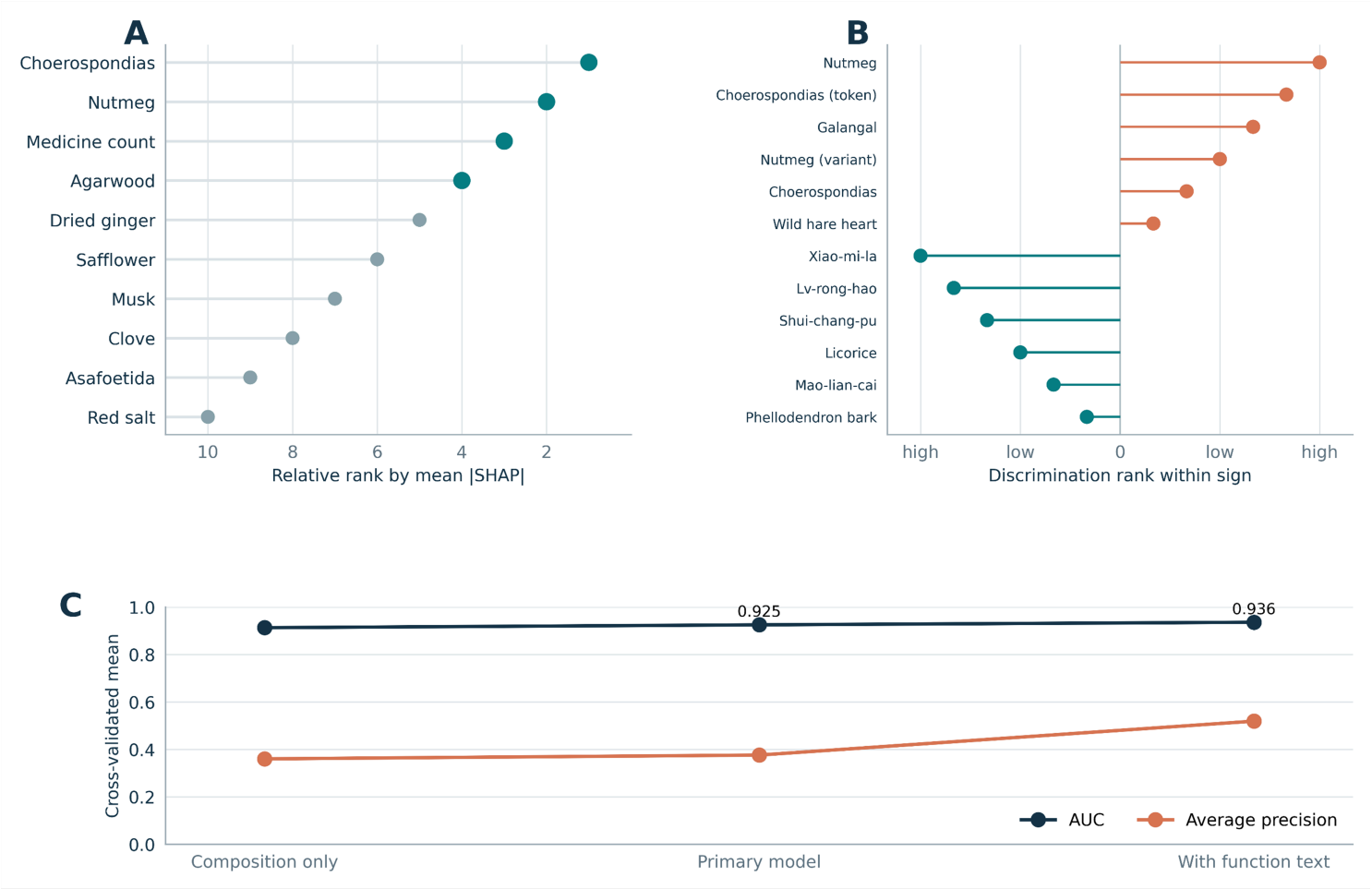


Figure 3

Feature ranks and function-text sensitivity. A, rank of mean absolute SHAP importance, and B, discriminating features ranked within positive and negative directions as displayed in the source plots; ranks do not represent coefficient magnitudes. C, AUC and average precision under three feature sets. The composition-only values were read from the original chart; no significance is implied by differences between plotted means.

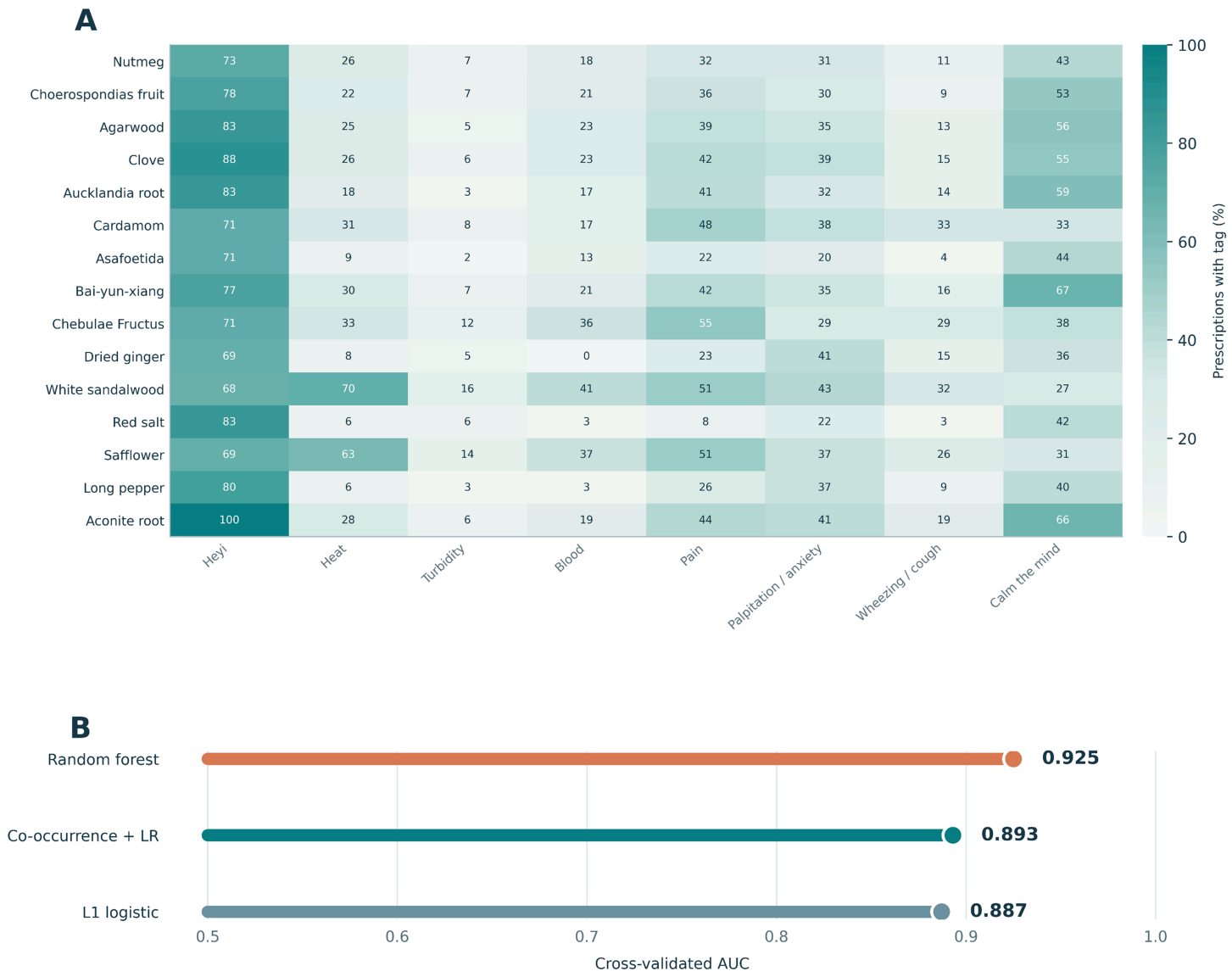


Figure 4

Function-text tags and baseline comparison. A, the 15 × 8 percentages transcribed from the original medicine–function heatmap. Tags are exploratory keywords, not diagnostic classes. B, AUC for L1 logistic regression, the co-occurrence mean-aggregation baseline and random forest; baseline value is reported in Results 3.6.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.tiff