Assessing Environmental Similarity Integration in Herb-Disease Association Prediction: An Augmented HDAPM-NCP Framework

Towards Environment-Aware Computational Pharmacology

Tobenna Udeze[1] and Ann Mathew[2]

[1] Computer Science, Vanderbilt University, Nashville, TN

[2] Computer Science, Vanderbilt University, Nashville, TN

Created November 30th, 2025

ABSTRACT

Context. Computational prediction of herb-disease associations represents an important intersection of traditional medicine, pharmacology, and data science. While biochemical similarity measures have proven effective, the environmental context of medicinal plants remains largely unexplored in computational frameworks.
Aims. This study aimed to assess whether incorporating environmental similarity information derived from climate, biome, and geographic data improves herb-disease association prediction within the HDAPM-NCP framework.
Methods. We extended the HDAPM-NCP framework by integrating herb occurrence data from GBIF with Köppen-Geiger climate classifications. A climate similarity kernel was engineered using Köppen code frequency vectors and fused with existing biochemical kernels via weighted averaging. The Network Similarity Projection algorithme generated association scores evaluated using AUROC and AUPR metrics.
Results. Initial integration showed negligible improvement in predictive performance. Baseline AUROC and AUPR were approximately 0.9322 and 0.9033, while climate-fused models yielded similar values. Results confirmed a neutral to slightly negative effect, indicating that current climate signals do not enhance prediction accuracy.
Conclusions. While environmental similarity did not immediately boost prediction accuracy, the study established a reproducible pipeline for environmental data integration. Future work should focus on refining climate feature engineering and exploring advanced fusion strategies.

Key words. computational pharmacology – network medicine – environmental similarity – herb-disease associations – climate integration

1. Introduction

The computational prediction of herb-disease associations represents an important intersection of traditional medicine, pharmacology, and data science. Traditional Chinese Medicine (TCM) and other natural medicine systems have documented thousands of herb-disease relationships through a long history of empirical observation, but systematic validation and discovery of novel associations remains challenging. The emergence of network-based computational approaches has offered opportunities to scale association discovery while also providing tangible insights.

The foundational work by Chen et al. (2025) established the HDAPM-NCP (Herb-Disease Association Prediction Model via Network Consistency Projection) framework, which demonstrated superior performance in predicting herb-disease associations compared to previous methods. This model leverages similarity network projection to propagate known associations through herb-herb and disease-disease similarity neighborhoods, effectively predicting novel therapeutic relationships. The original implementation incorporated multiple herb similarity dimensions including target proteins, Gene Ontology (GO) enrichment, chemical ingredients, KEGG pathway enrichment, and gene targets, which are all biochemical or molecular in nature.

1.1. Environmental Context in Medicinal Plants

Environmental factors, particularly climate, fundamentally shape plant secondary metabolism and the production of bioactive compounds with therapeutic potential. Environmental stressors such as temperature extremes, drought, and UV radiation can trigger specific biochemical pathways that increase concentrations of valuable secondary metabolites like alkaloids, phenolics, and (7; 4). These environmentally-mediated variations suggest that plants adapting to similar environmental pressures may develop convergent biochemical profiles, potentially leading to shared medicinal properties.

Environmental niche theory and principles of convergent evolution provide a theoretical framework for understanding these patterns. Unrelated plant species occupying similar environmental niches often evolve comparable traits (including biochemical defenses) through independent adaptation to comparable selective pressures (6). This environmental convergence may extend to therapeutic properties, as plants in similar environments develop similar chemical arsenals against herbivores, pathogens, and abiotic stresses, some of which may also apply to human diseases.

Ethnobotanical research also consistently documents geographical patterning in traditional medicinal plant use that correlates with environmental patterns. Studies of indigenous knowledge systems reveal that plant therapeutic applications are not randomly distributed but instead show regional patterns that align with environmental zones and climate characteristics (5; 2).

1.2. Research Gap and Objectives

Despite these developments in environmental phytochemistry, ethnobotanical pattern recognition, and network pharmacology, integration of environmental context into computational herb-disease prediction is still relatively unexplored. The HDAPM-NCP framework and similar approaches focus exclusively on biochemical and molecular similarity measures, overlooking the environmental dimensions that may play a role in shaping the medicinal properties of plants.

This study addresses this gap by augmenting the established HDAPM-NCP framework with environmental similarity measures derived from climate, biome, and geographic data. We hypothesize that **herbs growing in similar climates are more likely to share disease associations than those from dissimilar environments**. This environmental perspective is intended to inform rather than replace biochemical approaches, offering additional information that could enhance predictive accuracy and biological interpretability.

Building on the existing HDAPM-NCP framework, our research aims to:

1. Extend the existing model by incorporating environmental similarity measures derived from climate, biome, and geographic data
2. Develop a reproducible pipeline for harvesting and processing environmental data relevant to medicinal plants
3. Engineer a climate similarity kernel using Köppen-Geiger classifications
4. Implement and evaluate fusion strategies for integrating environmental with biochemical similarity measures
5. Assess whether environmental integration enhances predictive performance across multiple evaluation metrics
6. Analyze how the network is restructured by environmental augmentation to understand environmental patterns in herb-disease associations

2. Materials and Methods

2.1. Overview of HDAPM-NCP Framework

The baseline HDAPM-NCP framework (1) uses Network Consistency Projection (NCP, also referred to as Network Similarity Projection) to predict herb-disease associations. The core inputs include:

1. **Herb similarity matrix (H)**: Combines multiple biochemical similarity measures:
   - Target similarity: Based on shared protein structures targeted in the human body
   - GO enrichment similarity: Based on shared Gene Ontology term enrichment patterns
   - Ingredient similarity: Based on shared chemical components
   - KEGG enrichment similarity: Based on shared pathway enrichment
   - Gene target similarity: Based on shared gene expression targets
2. **Disease similarity matrix (D)**: Computed from disease-gene associations and semantic similarity
3. **Herb-disease adjacency matrix (A)**: Binary matrix indicating known associations from the HERB database (3)

The NSP algorithm projects association information through both similarity spaces:

$$S = \frac{A \cdot D \cdot \text{norm}_{\text{row}} + H \cdot A \cdot \text{norm}_{\text{col}}}{||H_{\text{row}}||_2 + ||D_{\text{col}}||_2} \tag{1}$$

where:

- S is the predicted herb-disease association score matrix
- A is the binary herb-disease adjacency matrix (known associations)
- D is the disease similarity matrix
- H is the herb similarity matrix
- norm_{row} denotes row-wise normalization
- norm_{col} denotes column-wise normalization
- $||\cdot||_2$ is the L2-norm
- H_{row} and D_{col} represent row-normalized herb similarity and column-normalized disease similarity matrices respectively

and normalization controls for degree effects and the denominator ensures consistent scaling.

2.2. Environmental Data Integration

2.2.1. Data Sources

The HERB database (herb.ac.cn) served as the primary source for known herb-disease associations. This comprehensive database integrates experimental evidence, literature mining, and traditional knowledge to document therapeutic relationships.

Geographic occurrence records for herb species were retrieved from the Global Biodiversity Information Facility (GBIF) using the pygbif Python library. For each herb species with existing taxonomic identification, we collected occurrence records including decimal latitude and longitude coordinates, applying basic quality filters for coordinate validity and temporal relevance.

We derived climate zone assignments using the Köppen-Geiger classification system, which categorizes global climates based on temperature, precipitation, and seasonal patterns. We employed the lookupCZ function to map each occurrence coordinate to its corresponding Köppen code (e.g., Cfa for temperate humid climate, Dfb for cold snow-forest climate).

2.2.2. Data Processing Pipeline

The environmental data processing pipeline comprised several interconnected aspects:

1. **Climate Data Harvest**: Query GBIF for occurrence records per herb species and map coordinates to Köppen climate codes
2. **Herb Set Filtering and Climate Vectorization**:
   - Load climate data and herb identification files
   - Identify herbs missing GBIF occurrence data (6 out of 25 original herbs lacked data)
   - Drop corresponding rows/columns from existing biochemical kernels
   - Generate filtered files and climate-only kernel using cosine similarity over normalized frequency vectors of climate codes per herb

3. **Climate Kernel Construction**:
   - For each herb, compile occurrence counts per Köppen-Geiger climate code
   - Normalize counts to probability distributions (summing to 1), capturing multi-climate herbs naturally
   - Compute climate similarity between herbs using cosine similarity of these distribution vectors

   The strategy for this step was to perform a simple lookup using a python API to first obtain the scientific name for each herb according to their ID using a web-scrapper, perform a search on the GBIF site with pygbif to determine if these herbs have data occurrences on world climatic databases that will give us geographical coordinates of highly concentrated populations of the species. From there, we can categorize each herb according to the köppen-geiger climatic classification for the biome descriptors. Table 2 illustrates how the climatic information was looked after it had been included into the herb kernel.

4. **Kernel Fusion and Prediction**:
   - Implement weighted fusion of environmental and biochemical kernels
   - Execute NSP algorithm with fused kernels
   - Generate comparative predictions for baseline and climate-integrated models

2.2.3. Kernel Fusion Strategy

The climate distribution kernel was fused with the baseline biochemical kernel using a weighted average:

$$K_{\text{fused}} = \alpha \cdot K_{\text{climate}} + (1 - \alpha) \cdot K_{\text{baseline}} \tag{2}$$

where α represents the weight given to the climate kernel. To systematically evaluate the contribution of environmental similarity, we performed a parameter sweep across α values ranging from 0.0 to 0.9 in increments of 0.1.

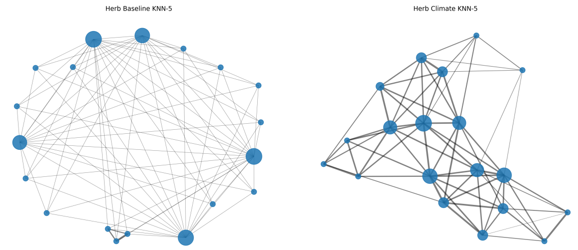

Fig. 1: ***k*-NN (*k* = 5) herb similarity network using the climate-derived kernel Vs. baseline kernel.** *The baseline kernel creates a globally connected network, providing the primary predictive signal. In contrast, the climate kernel yields a more modular structure with distinct ecological clusters.*

2.3. Evaluation Methodology

2.3.1. Performance Metrics

1. **Area Under Receiver Operating Characteristic Curve (AUROC)**: Measures the model's ability to rank positive associations higher than negative ones across all threshold settings.
2. **Area Under Precision-Recall Curve (AUPR)**: Particularly valuable under class imbalance as it focuses on the model's performance on the positive class (existing associations).
3. **Average Precision (AP)**: The weighted mean of precisions at each threshold, with the increase in recall from the previous threshold used as weight.

2.3.2. Experimental Design

1. **Baseline Evaluation**: NSP with original biochemical kernels only ($\alpha = 0.0$)
2. **Climate-Integrated Evaluation**: NSP with fused kernels at $\alpha = 0.3$

The ground truth consisted of a filtered association matrix with 3,106 positive associations and 4,494 negative associations (total 7,600 pairs).

2.3.3. Implementation

Analyses were conducted in Python using standard scientific computing libraries. The implementation focused on creating a reproducible pipeline for climate data integration and evaluation.

3. Results

3.1. Predictive Performance Analysis

Integration of climate information into the HDAPM-NCP framework resulted in minimal changes to predictive performance. A comparative run with the climate kernel weight parameter α set to 0.3 yielded the following metrics against the established baseline:

The observed differences (ΔAUROC = -0.0006, ΔAUPR = -0.0010) indicate a slight decrease in performance. Evaluation was performed on a filtered dataset containing 3106 positive herb-disease associations and 4494 negative associations (total 7600 pairs).

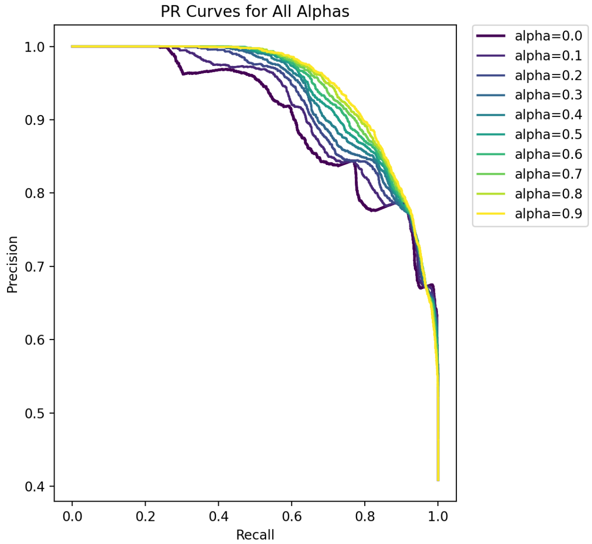

Fig. 2: ***Precision–Recall curves for all α values in the fused herb kernel.*** *Each curve represents a fused kernel $K_{fused} = \alpha K_{climate} + (1 - \alpha)K_{baseline}$. PR performance remains nearly identical across α, indicating that the current climate-based similarity (derived from Köppen–Geiger climate-zone distributions) contributes no additional precision at higher recall.*

Table 1: Comparison of Network Topology Metrics Reveals Climate Kernel Produces More Modular Structure than Biochemical Kernel

Graph	Density	Avg. Clust.	Avg. Degree	Edges	Transitivity
herb_baseline	0.45	0.01	8.105	77	0.509
herb_climate	0.392	0.422	7.053	67	0.562
disease	0.025	0.287	9.835	1,967	0.068

The climate kernel (herb_climate) produces a more modular network structure than the biochemical kernel (herb_baseline), evidenced by its significantly higher average clustering coefficient (0.422 vs. 0.01) despite having a lower density. This indicates that herbs cluster into distinct ecological groups based on shared climate zones, consistent with visual inspection of Fig. 1. The disease network exhibits sparse, highly connected topology typical of semantic similarity graphs.

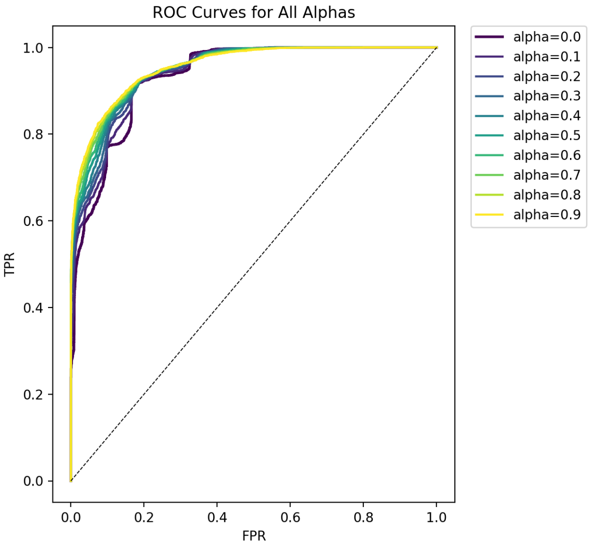

Fig. 3: ***Receiver Operating Characteristic curves for all α values in the fused herb kernel.*** *ROC curves exhibit strong overlap from $\alpha = 0$ to 0.9, demonstrating that the climate-based similarity kernel does not alter global ranking behavior. AUROC differences remain minimal, confirming that fusion produces no measurable gain over the baseline biochemical kernels.*

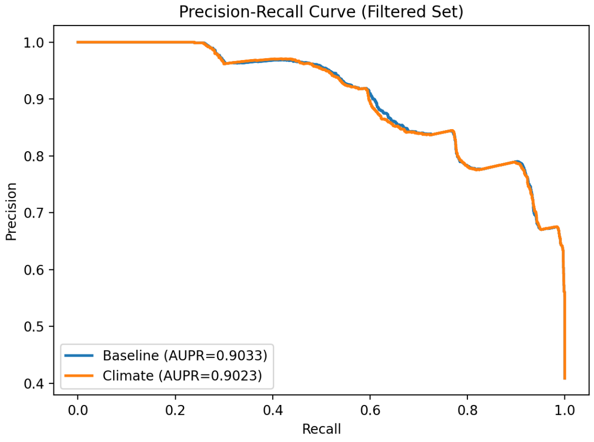

Fig. 4: ***Combined Precision–Recall curve across all α values.*** *The curves cluster tightly with no consistent trend as α increases, indicating that the climate kernel—constructed from Köppen–Geiger climate-zone frequency vectors—does not contribute additional predictive signal under high class imbalance.*

Table 2: Climate Zone Distribution for Select Medicinal Herbs

Herb ID	**Species**	**Biome**	**K-Zone**
HERB000189	Ginkgo biloba	Temperate	Cfb
HERB000523	Lepidium apetalum	Continental	Dwb
HERB001023	Glycine max	Dry	BSk
HERB001104	Allium sativum	Tropical	Aw
HERB001210	Angelica sinensis	Temperate	Cwa
HERB001333	Ophiocordyceps sinensis	Ocean	Ocean
HERB006931	Cinnamomum camphora	Temperate	Csb

Representative climate zones for medicinal herbs in the dataset. Many herbs occur in multiple climate zones; this table shows one representative zone per herb.

Table 3: Alpha sweep results, including baselines without climate fusion.

Model	**Alpha**	**ROC AUC**	**PR AUC**	**Avg Precision**
Baseline (no climate fusion)	—	0.93220	0.90331	0.90332
Climate-only (no fusion)	—	0.93163	0.90234	0.90236
Fused	0.0	0.93220	0.90331	0.90332
Fused	0.1	0.93824	0.91457	0.91458
Fused	0.2	0.94263	0.92280	0.92282
Fused	0.3	0.94565	0.92807	0.92808
Fused	0.4	0.94821	0.93233	0.93234
Fused	0.5	0.95036	0.93591	0.93592
Fused	0.6	0.95207	0.93886	0.93887
Fused	0.7	0.95352	0.94142	0.94143
Fused	0.8	0.95469	0.94350	0.94351
Fused	0.9	0.95560	0.94517	0.94518

Data summary: Positives = 3106, negatives = 4494, total $n = 7600$.
Notes: "Baseline (no climate fusion)" equals $\alpha = 0.0$
$K_{\text{fused}} = \alpha K_{\text{climate}} + (1 - \alpha)K_{\text{baseline}}$.

3.2. Network Restructuring by Climate Similarity

Climate similarity introduced clear structural changes to the herb network. The climate-derived kernel produced a more modular, clustered topology (Fig. 1) compared to the densely connected, homogeneous network from the biochemical kernel. Herbs grouped into distinct clusters based on shared Köppen-Geiger climate zones.

Despite this reorganization, the new clusters did not improve predictive performance. Precision-Recall and ROC curves remained nearly identical across all fusion weights (Figs. 2-5), as reflected in the consistent metrics across α values in Table 3. This indicates that while climate reshapes network topology, it does not add discriminative signaling for herb-disease association prediction under the current representation.

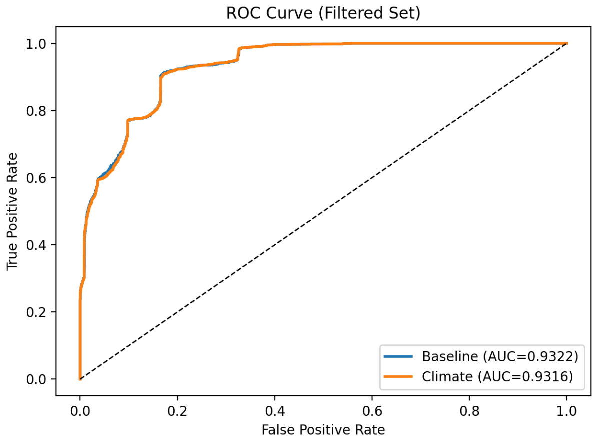

Fig. 5: ***Combined ROC curves across all α values used in the fusion sweep.*** *Each curve corresponds to a fused similarity $K_{fused} = \alpha K_{climate} + (1-\alpha)K_{baseline}$. ROC performance remains effectively unchanged across α, confirming that environmental similarity does not meaningfully improve discriminative power under the current feature construction.*

3.3. Data Characteristics and Limitations

Analysis of the implementation revealed several factors potentially contributing to the neutral performance outcome:

1. **Climate Kernel Characteristics**: The climate distribution kernel, based on Köppen code frequencies, was noted for potential sparsity and noise. Herbs occurring in multiple climate zones produced diffuse frequency vectors, which may impact discriminative structural information when using cosine similarity.
2. **Feature Redundancy**: Moderate correlation was suspected between the new climate similarity and the existing biochemical kernels (target, GO, ingredient, etc.), suggesting the environmental signal may already be captured in a way by the existing feature space.
3. **Data Coverage Bias**: The reliance on GBIF occurrence data introduced a selection bias. Well-documented species had richer climate profiles, while herbs with sparse or missing records were filtered out, potentially limiting the variance and novel signal the climate data could possibly introduce.

3.4. Climate Zone Distribution

Of the original 25 herbs, 19 had available occurrence records in GBIF, while 6 herbs lacked occurrence data and were excluded from climate-integrated analyses. The 19 herbs with occurrence data showed the following distribution across climate zones:

- Cfb (Temperate): 14 herbs
- Cfa (Temperate): 13 herbs
- Cwa (Temperate): 13 herbs
- Csb (Temperate): 12 herbs
- Aw (Tropical): 9 herbs
- BSk (Dry): 9 herbs

Individual herbs frequently occurred in multiple climate zones, with representation across 22 distinct Köppen-Geiger classifications.

4. Discussion

The minimal impact of climate integration on predictive performance, demonstrated by almost identical AUROC/AUPR values between baseline and climate-fused models, can be attributed to several key implementation-specific factors.

First, the climate kernel derived from Köppen code frequency vectors was inherently sparse and noisy, particularly for herbs occurring in multiple climate zones, diluting its discriminative power. The cosine similarity measure applied to these normalized frequency vectors may not effectively capture meaningful environmental relationships when herbs exhibit broad climate tolerances. Additionally, the ecoregion layer used in this study (climate data from RESOLVE) was relatively coarse and may not have captured finer-scale environmental variations that could influence herb biochemistry. Future work should incorporate more detailed, high-resolution climate datasets to better represent the environmental niches of medicinal plants.

Second, the exclusion of herbs lacking GBIF records may have biased the analysis toward widely distributed species with comprehensive occurrence data, potentially limiting any novel signal that climate data could introduce. This selection bias toward well-documented species may have reduced the environmental variance available for predictive modeling.

Third, moderate correlation between climate similarity and existing biochemical features suggests potential redundancy. The environmental influences on medicinal properties may already be indirectly captured through biochemical similarity measures, as plants from similar environments often develop comparable secondary metabolite profiles through convergent evolution.

These results, supported by the nearly identical PR and ROC curves across all α values (Figs. 2-5), suggest that while environmental theory indicates the existence of strong environment-herb-disease relationships, translating this into computational predictive gains requires more nuanced feature engineering than rough climate zone frequencies. The environmental connection appears to either be redundant with existing biochemical similarities or too weakly correlated with therapeutic associations to improve the model in a significant way under the current implementation.

Future work could benefit from incorporating insights from multi-view learning frameworks, such as those described by Li et al. (2014), which integrate annotational, functional, and topological similarity measures. While applying such a framework in this study would have reduced our dataset size and limited interpretability, future studies with expanded herb coverage and more detailed environmental data could leverage these approaches to better understand environment–herb–disease relationships. These approaches could potentially capture the complementary signals observed in our network analysis (Fig. 1), where climate similarity created distinct ecological clusters that were not leveraged for predictive gain.

Despite the neutral predictive result, this study established a reproducible and modular pipeline for integrating environmental data into network-based association prediction. The framework supports kernel fusion and systematic evaluation, which could provide a solid foundation for future refinement. The methodological contributions include:

1. A complete data processing pipeline for harvesting and integrating environmental occurrence data
2. Climate similarity kernel construction using Köppen-Geiger classifications

3. Weighted fusion strategies for combining environmental and biochemical similarity measures
4. Systematic evaluation framework for environmental augmentation in network-based prediction

4.1. Future Directions

Future work should focus on data enrichment and methodological improvements:

1. **Continuous Climate Variables**: Incorporating continuous climate variables (temperature, precipitation, seasonality indices) rather than categorical climate zones might allow for more discriminative environmental features.
2. **Advanced Fusion Strategies**: Other fusion strategies, such as attention-based weighting or graph neural networks, could potentially capture complementary signals between environmental and biochemical similarity spaces in a more effective way.
3. **Multi-Scale Environmental Features**: Combining climate data with soil characteristics, altitude, and other ecological variables could provide a more comprehensive environmental representation.
4. **Expanded Herb Coverage**: Improving data coverage through manual curation or alternative data sources could reduce selection bias and increase environmental variance.
5. **Non-linear Similarity Measures**: Exploring non-linear similarity measures or deep learning approaches might better capture complex environment-biochemistry-therapeutic relationships.

5. Conclusion

This study attempted to augment the HDAPM-NCP framework with environmental similarity measures to test whether environmental data could enhance herb-disease association prediction. Initial integration using Köppen climate codes resulted in minimal performance changes, indicating that the current representation does not add significant predictive value beyond the existing biochemical kernels under the implemented framework.

However, our work provides important frameworks and diagnostic insights for future research. The augmented pipeline results in reproducible climate data integration, and the analysis identifies specific limitations, such as kernel sparsity, feature redundancy, and data coverage bias. These insights can serve to guide future improvements in environmental feature engineering for computational pharmacology.

The neutral predictive outcome should not be interpreted as evidence against environment-herb-disease relationships, but rather as an indication that current environmental representations may not capture the relevant signals for association prediction, or that these signals are already captured indirectly through biochemical features. More sophisticated environmental modeling and integration strategies may be required to realize the potential of environmental data in computational pharmacology.

While we were unable to identify direct predictive gains, this work establishes an essential foundation for more sophisticated environmental integration in computational pharmacology, ideally allowing future research to move toward models that reflect the complex interactions between environment, plant biochemistry, and therapeutic potential. The reproducible pipeline and methodological framework developed here can facilitate continued exploration of environmental dimensions in network-based association prediction.

Acknowledgements. We acknowledge the Global Biodiversity Information Facility (GBIF) for providing species occurrence data and the HERB database for herb-disease association data. We thank Dr. You Chen and Xinmeng Zhang for helpful discussions and feedback during the project progress reports.

References

[1] Chen, L., Zhang, S., & Zhou, B. (2025). Herb-disease association prediction model based on network consistency projection. *Scientific Reports*, *15*, 3328. https://doi.org/10.1038/s41598-025-87521-7

[2] Gaoue, O. G., Coe, M. A., Bond, M., Hart, G., Seyler, B. C., & McMillen, H. (2017). Theories and Major Hypotheses in Ethnobotany. *Economic Botany*, *71*(3), 269–287. https://doi.org/10.1007/s12231-017-9389-8

[3] HERB Database. (2020). *HERB: A High-Throughput Experiment- and Reference-Guided Database of Traditional Chinese Medicine*. http://herb.ac.cn

[4] Kleinwächter, M., & Selmar, D. (2013). Stress enhances the synthesis of secondary plant products: the impact of stress-related over-reduction on the accumulation of natural products. *Plant and Cell Physiology*, *54*(6), 817–826. https://doi.org/10.1093/pcp/pct054

[5] Moerman, D. E. (1996). An analysis of the food plants and drug plants of native North America. *Journal of Ethnopharmacology*, *52*(1), 1–22. https://doi.org/10.1016/0378-8741(96)01393-1

[6] Palacio, S., Paterson, E., Hester, A. J., Nogués, S., Lino, G., Anadon-Rosell, A., Maestro, M., & Millard, P. (2020). No preferential carbon-allocation to storage over growth in clipped birch and oak saplings. *Tree Physiology*, *40*(5), 621–636. https://doi.org/10.1093/treephys/tpaa011

[7] Yang, L., Wen, K.-S., Ruan, X., Zhao, Y.-X., Wei, F., & Wang, Q. (2018). Response of Plant Secondary Metabolites to Environmental Factors. *Molecules*, *23*(4), 762. https://doi.org/10.3390/molecules23040762