

Multi-Omics Data Fusion and Artificial Intelligence for Phylogenetic Classification of Medicinal Plant Resources

Anel Melik

Department of Computer Science, Binghamton University, Binghamton, NY, USA.

helloanil@binghamton.edu

Zachary Fleming

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.

zachary.work@uab.edu

Karan C. Menon

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis,
OR, USA.

kcmemon@oregonstate.edu

Malcolm M. Jarvinen

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
KS, USA.

malcolmj@ku.edu

Abstract

Medicinal plant resources are central to pharmaceutical innovation, traditional medicine, and biodiversity conservation, yet their phylogenetic classification remains complicated by cryptic speciation, hybridization, phenotypic plasticity, and incomplete voucher-linked reference collections. This paper examines how multi-omics data fusion combined with artificial intelligence can transform phylogenetic classification from a narrow molecular barcoding task into a systems-level inference problem. It analyzes the structural trade-offs among early, intermediate, and late data fusion architectures and discusses how genomic, transcriptomic, metabolomic, and cell wall compositional features can be jointly represented for downstream classification. The discussion emphasizes system requirements such as provenance-aware data pipelines, modular model design, uncertainty quantification, validation across heterogeneous plant lineages, and continuous monitoring after deployment. The paper argues that technical accuracy alone is insufficient. Fairness, benefit-sharing, ecological sustainability, and policy coherence must be designed into the infrastructure rather than appended after model training. Drawing on cross-domain comparisons from biomedical and ecological data science, the paper provides a forward-looking perspective on resilient, governable, and ethically grounded computational infrastructure for medicinal plant phylogenetics. The conclusion highlights the need for international data governance frameworks, reproducible model registries, and participatory approaches that include Indigenous and local knowledge systems in the curation and interpretation of medicinal plant data.

Keywords

multi-omics data fusion; artificial intelligence; phylogenetic classification; medicinal plant resources; data governance; fairness.

1. Introduction

Medicinal plant resources present a distinctive challenge for modern computational biology because they are simultaneously objects of ecological study, pharmaceutical development, cultural heritage, and conservation policy. The same species may be understood through traditional therapeutic knowledge, molecular taxonomy, metabolomic profiling, and biogeographic distribution. This multiplicity of interpretive frames creates analytical complexity, but it also provides a rich foundation for data-driven classification systems. The core difficulty is not simply the absence of data, but the heterogeneity of data types, the unevenness of reference collections, and the need to reconcile evolutionary relatedness with functional and chemical similarity. In this context, phylogenetic classification is not only a question of evolutionary history; it is also a practical mechanism for verifying botanical identity, ensuring the safety of medicinal plant supply chains, and prioritizing conservation actions for threatened taxa.

Advances in deep learning have fundamentally changed the way high-dimensional biological data can be represented, transformed, and classified [1]. In parallel, multi-omics data fusion has emerged as an organizing principle for integrating signals that are distributed across different molecular layers and measurement technologies [2,3]. These developments intersect with a longer tradition of plant molecular systematics in which DNA barcoding has provided a rapid and reproducible basis for species identification [4]. Within medicinal plant research, there has been a gradual but decisive shift from morphology-only identification toward integrative approaches that combine anatomical, chemical, and molecular evidence [5]. International efforts to establish standardized barcode markers for land plants have strengthened the reference infrastructure for such integrative work, although they have also exposed the limitations of single-locus approaches in highly complex plant genera [6].

The central argument of this paper is that the next generation of medicinal plant classification systems should be designed as socio-technical infrastructures rather than as isolated machine learning models. This requires careful attention to data fusion architectures, phylogenetic inference methods, deployment environments, robustness under distribution shift, governance arrangements, and long-term sustainability. It also requires explicit recognition that medicinal plant data are entangled with questions of access, ownership, and traditional knowledge. In the following sections, these concerns are examined from a systems-level perspective.

2. The Data Landscape for Medicinal Plant Phylogenetics

An integrative view of medicinal plant biology must coordinate genomic, transcriptomic, proteomic, metabolomic, and anatomical data over large and often unevenly sampled lineages [7]. These data layers differ in their temporal stability, measurement cost, and phylogenetic information content. Genomic sequences provide a relatively stable evolutionary record, while metabolomic profiles may shift with developmental stage, tissue type, season, and environmental stress. Genomic sequences provide a relatively stable evolutionary record, while metabolomic profiles may shift with developmental stage, tissue type, season, and environmental stress. This variability does not diminish the value of metabolomic data but indicates that classification systems must account for conditional expression rather than treating chemical fingerprints as invariant species markers. Transcriptomic and proteomic layers add further complexity because they capture dynamic regulatory states that may be

shared across distantly related lineages under similar ecological pressures. Anatomical and cell wall compositional features occupy a different position in this landscape. They often integrate genetic, developmental, and environmental influences over longer temporal scales and therefore provide complementary evidence for resolving taxa that are poorly separated by sequence data alone [10]. The challenge for a system-level design is not to privilege one data layer but to construct an infrastructure that can absorb heterogeneous evidence without imposing a single biological ontology prematurely.

The unevenness of reference collections creates structural difficulties that go beyond ordinary missing data problems. Many medicinal plant species are represented in public repositories only by partial sequences, inconsistently annotated voucher specimens, or chemical profiles collected under varying experimental conditions [8]. Some lineages are heavily sampled because of commercial or pharmacological interest, while close relatives with subtle diagnostic differences remain underrepresented. This imbalance means that classification models trained on current repositories may learn to recognize abundant species accurately while systematically failing on rare or locally important taxa. A robust data landscape therefore requires not only more data but also better provenance tracking, standardized metadata, and deliberate sampling strategies that address phylogenetic blind spots. The importance of voucher-linked specimen data cannot be overstated, because downstream classification reliability depends on the quality of the reference taxonomy as much as on the learning algorithm.

3. Data Fusion Architectures

The selection of a data fusion strategy is one of the most consequential design decisions in any multi-omics classification system. Early fusion approaches concatenate raw or lightly processed features from different omics layers into a single matrix before model training. This strategy is conceptually simple and allows a model to learn cross-layer interactions directly, but it suffers when individual layers have very different dimensionality, noise characteristics, or missingness patterns. In medicinal plant phylogenetics, early fusion can be problematic because genomic alignments may contain tens of thousands of variable sites, while metabolomic profiles may consist of a few hundred quantified compounds, and cell wall compositional data may be represented by a small set of monosaccharide ratios. Without careful normalization and dimensionality reduction, the highest-dimensional layer tends to dominate the representation, and the specific phylogenetic signal from chemical or anatomical features may be lost.

Intermediate fusion architectures attempt to preserve layer-specific structure by first learning compressed representations for each data type and then combining those representations in a shared latent space. This approach has parallels in multimodal learning methods that are increasingly used in biomedical research to integrate imaging, molecular, and clinical data [2]. For medicinal plant resources, intermediate fusion can be operationalized through modality-specific autoencoders, kernel methods, or attention-based mechanisms that weigh the contribution of each layer according to its relevance for a given classification task. A well-designed intermediate fusion system can be more robust to missing data because each modality encoder can be trained independently before joint fine-tuning. However, this modularity comes with additional computational and engineering overhead, and it requires careful decisions about how many shared dimensions should be allocated to each omics layer.

Late fusion, sometimes called decision-level fusion, trains separate classifiers on each data layer and then combines their outputs through voting, stacking, or Bayesian model averaging.

Late fusion is attractive when the data layers are collected in different laboratories, at different times, or under different governance regimes, because it avoids the need to co-localize raw data. It also facilitates interpretability, since each layer-specific model can be inspected independently. A major limitation is that late fusion cannot capture cross-layer interactions that only become visible when features are considered jointly. In phylogenetic classification, such interactions may be important, for example when a chemical signature is strongly diagnostic only in combination with a particular genomic lineage. The choice among early, intermediate, and late fusion is therefore not a purely technical optimization problem; it reflects assumptions about the underlying biology and about the institutional conditions under which data can be shared and processed.

A pragmatic systems design may adopt hierarchical fusion strategies in which some layers are fused early because they share common provenance, while other layers are integrated later because they are governed by different access constraints. This hybrid arrangement can be more resilient than a single monolithic architecture, but it demands clear interface contracts and well-defined uncertainty propagation across fusion stages. For medicinal plant workflows that involve field sampling, herbarium vouchering, laboratory sequencing, and mass spectrometry, the data pipeline must be able to combine block-level and record-level identifiers without introducing silent misalignment errors. The integrity of the fusion process depends as much on data engineering discipline as on model selection.

4. Artificial Intelligence and Phylogenetic Inference

Artificial intelligence models for phylogenetic classification can be broadly divided into supervised approaches, which learn a mapping from molecular or chemical features to predefined taxonomic labels, and unsupervised or semi-supervised approaches, which learn latent structures that may correspond to evolutionary relationships without relying on fixed taxonomies [9]. In supervised settings, deep neural networks have demonstrated strong performance on problems such as DNA barcode classification and species identification from image or spectral data [1]. These models can learn nonlinear feature interactions that are difficult to specify manually, and they can scale to large reference sets. However, supervised models are fundamentally constrained by the quality and completeness of the label space. When the reference taxonomy is contested or when cryptic species remain undescribed, a supervised model may learn to reproduce existing annotations without revealing the underlying phylogenetic uncertainty.

Unsupervised and self-supervised methods offer a different pathway. Approaches such as variational autoencoders, contrastive learning, and graph neural networks can learn representations that reflect similarity structure in multi-omics data without requiring exhaustive taxonomic labels [3]. These representations can then be used to generate hypotheses about species boundaries, hybrid origins, or convergent chemical evolution. In phylogenetic classification, such models are useful when the goal is not only to assign an unknown sample to an existing species but also to place that sample within a broader evolutionary context. Graph-based methods are particularly relevant because phylogenetic relationships are inherently relational, and a graph representation can capture reticulate evolution, incomplete lineage sorting, and hybridization more naturally than a strictly hierarchical tree [11].

The integration of phylogenetic priors into neural network architectures is an active area of research. One approach constrains the learned latent space to be consistent with known phylogenetic distances, while another approach uses tree-aware loss functions that penalize

classifications inconsistent with a reference topology. These methods can improve generalization to taxa that were not present in the training set, which is a critical requirement for medicinal plant resources where new species and chemotypes are continuously being described. Transfer learning strategies, in which models pretrained on large general plant datasets are fine-tuned on smaller curated medicinal plant collections, have shown promise in reducing the amount of labeled data needed for accurate classification [11]. However, transfer learning can also introduce subtle biases when the source domain is dominated by agricultural or model plant species whose genomic and chemical characteristics differ from those of wild medicinal taxa.

Uncertainty quantification is essential for any system that is expected to support regulatory decisions, conservation prioritization, or clinical safety assessments. Deep learning models are often poorly calibrated, meaning that their predicted probabilities do not reliably reflect true confidence [20]. Bayesian approximation methods, ensemble techniques, and conformal prediction can provide more reliable uncertainty estimates, but they increase computational cost and may complicate deployment in resource-limited settings. For phylogenetic classification, uncertainty must be conveyed not only at the species assignment level but also at deeper nodes in the tree, because an ambiguous placement at a particular taxonomic rank may be more informative than a single high-confidence but misleading label. This requires model outputs that are structured as distributions over trees or as sets of credible placements rather than as point estimates.

5. Infrastructure and Deployment

The deployment of multi-omics classification systems for medicinal plant resources requires a data infrastructure that is fundamentally different from a one-time research pipeline. In many current workflows, sequence data are generated in one facility, metabolomic data in another, and anatomical data in a third, with limited coordination across institutions. The result is a fragmented data ecosystem in which sample identifiers are not consistently linked, metadata are incomplete, and provenance information is lost during downstream processing. A well-designed infrastructure must treat data curation as a first-class activity, with versioned reference databases, standardized ontologies for plant organs and tissues, and automated checks for taxonomic consistency [14]. The FAIR principles of findability, accessibility, interoperability, and reusability provide a useful baseline, but they are not sufficient for systems that must operate across national borders and diverse legal regimes.

The computational requirements of multi-omics fusion are substantial, particularly when models are retrained frequently or when large genomic and metabolomic matrices are processed. Cloud-based platforms offer scalability and ease of access, but they raise concerns about data sovereignty and long-term cost sustainability. Federated learning has been proposed as a way to train models across distributed data repositories without centralizing raw data [13]. This approach is attractive for medicinal plant databases that contain commercially sensitive or culturally significant information, but it introduces new challenges in model synchronization, communication efficiency, and auditability. In practice, a hybrid architecture may be most suitable, in which common reference data are centrally maintained while locally sensitive datasets contribute to model training through privacy-preserving protocols.

Deployment environments for medicinal plant classification range from high-throughput pharmaceutical quality control laboratories to field-based biodiversity surveys in remote regions. These environments differ dramatically in their connectivity, computing resources, and user expertise. A classification system intended for field use must be robust to

intermittent network access, must operate on low-power devices, and must provide interpretable outputs that can be acted upon by non-specialist users. This argues for compact model variants, on-device inference, and careful attention to user interface design. At the same time, the system must retain the ability to synchronize with central repositories when connectivity becomes available, so that field observations can contribute to ongoing model improvement.

Continuous monitoring is another essential component of deployment. Once a model is released, its performance will drift as new data are collected, as reference taxonomies are revised, and as environmental conditions change. A monitoring framework should track the distribution of input features over time, detect shifts in model confidence, and trigger retraining or recalibration when necessary [17]. For medicinal plant applications, such monitoring is particularly important because the chemical composition of plants can shift due to climate change, cultivation practices, or adulteration in commercial supply chains. A system that was accurate when initially deployed may become silently unreliable if it is not maintained as a living infrastructure.

6. Robustness, Fairness, and Governance

Robustness in multi-omics classification must be evaluated across multiple dimensions, including biological variation, technical batch effects, and adversarial or accidental contamination. Batch effects are pervasive in metabolomics and transcriptomics because measurements are influenced by instrument calibration, reagent lots, and laboratory protocols. A model that has not been explicitly trained to be invariant to these nuisance factors may learn to classify samples by their batch of origin rather than by their true biological identity. Domain adaptation and harmonization methods can mitigate these problems, but they must be validated on independent datasets that reflect realistic variation [3]. In the context of medicinal plants, the presence of closely related species that share many chemical and genetic features makes robustness especially difficult to achieve, because small measurement perturbations can flip a classification across a species boundary.

Fairness in medicinal plant classification is not typically framed in the same terms as fairness in social applications of machine learning, but it is equally important. A classification system may be highly accurate on average while systematically failing for taxa that are rare, geographically restricted, or culturally significant to specific communities [12]. These failures can reinforce existing biases in conservation funding, pharmaceutical research, and international trade regulation. Fairness-aware evaluation should therefore disaggregate performance by taxonomic group, geographic region, and data source, rather than relying on global accuracy metrics. Where traditional knowledge plays a central role in medicinal plant identification, the design process must include mechanisms for communities to contest incorrect classifications and to contribute authoritative knowledge without being subsumed into a purely Western taxonomic framework.

Governance arrangements must address the fact that medicinal plant data are not neutral resources. They are entangled with intellectual property claims, access and benefit-sharing obligations under the Nagoya Protocol, and the rights of Indigenous peoples and local communities [15]. A classification infrastructure that ignores these dimensions may inadvertently facilitate biopiracy or undermine local stewardship. The CARE principles for Indigenous data governance emphasize collective benefit, authority to control, responsibility, and ethics [15]. These principles suggest that data governance should not be treated as a legal compliance afterthought but as a foundational requirement for system design. This includes

transparent data use agreements, community review processes, and benefit-sharing mechanisms that are appropriate to the scale and purpose of the research.

7. Sustainability and Policy Implications

Sustainability in medicinal plant research is often discussed in ecological terms, but the computational infrastructure itself must also be sustainable over time. Many research data platforms are built through short-term project funding and then become orphaned when grants end. A classification system that is intended to support long-term regulatory or conservation functions requires an institutional home, predictable funding, and a governance structure that can survive personnel turnover. Open-source software and open data standards can reduce the risk of vendor lock-in, but they do not by themselves solve the maintenance problem. Sustainable infrastructure requires a community of practice that includes taxonomists, data engineers, policy makers, and end users who have a stake in the system's continued operation [14].

The policy implications of AI-based phylogenetic classification extend into several domains. Regulatory agencies responsible for herbal medicine safety need to know whether a classification system is reliable enough to support enforcement actions. Conservation organizations need to know whether a system can detect mislabeled trade specimens or identify populations at risk. Patent offices and access and benefit-sharing authorities need to know whether a system can provide defensible evidence of species identity in cases involving traditional knowledge and genetic resources. In all these contexts, model explainability and documentation are as important as predictive accuracy [21], because decisions must be justified to courts, ethics committees, and the public. Model cards and datasheets for datasets provide a structured way to communicate the strengths, limitations, and intended uses of a classification system [16].

International coordination is essential but difficult. Medicinal plant species do not respect national boundaries, and many are distributed across multiple countries with different legal frameworks. A classification system that is authoritative in one jurisdiction may not be accepted in another unless there is mutual recognition of data standards and model evaluation procedures. The development of regional and global initiatives for plant biodiversity informatics can help align metadata standards and reference taxonomies, but such coordination must be attentive to power asymmetries between high-income and low-income countries. Countries that are rich in medicinal plant biodiversity but limited in computational infrastructure often bear the costs of data collection while receiving few of the benefits of downstream AI applications [15]. Policy frameworks should therefore include capacity-building provisions and equitable access to trained models, not just open access to raw data.

The integration of cell wall component characteristics into medicinal plant classification illustrates how a relatively underexplored data layer can contribute to more robust identification when combined with genomic and metabolomic evidence [10]. This is not a call to replace established barcode markers but rather to recognize that different data layers capture different aspects of evolutionary and functional divergence. Policy instruments that support multi-omics data generation should therefore avoid funding structures that reward only the most fashionable molecular technologies. Long-term value will come from maintaining a diversity of data types and from building the computational capacity to fuse them responsibly.

8. Conclusion

Multi-omics data fusion and artificial intelligence hold significant promise for phylogenetic classification of medicinal plant resources, but their success depends on more than algorithmic innovation. This paper has argued that the challenge should be understood as a systems problem that spans data infrastructure, model architecture, deployment logistics, robustness evaluation, fairness, governance, and policy. The structural trade-offs among early, intermediate, and late fusion strategies reflect deeper tensions between modularity and interaction, interpretability and predictive power, and local data sovereignty and global model sharing. A resilient classification system must be able to absorb heterogeneous data without erasing provenance, to express uncertainty honestly, and to evolve as reference taxonomies and environmental conditions change.

The inclusion of cell wall compositional features alongside more commonly used omics layers illustrates the value of maintaining methodological pluralism. No single data type can resolve all phylogenetic questions in complex medicinal plant lineages, and the most informative classification systems will be those that can weigh evidence from multiple sources in a transparent and contestable way. At the same time, technical pluralism must be matched by institutional pluralism. The governance of medicinal plant data and classification models must include the voices of the communities who have stewarded that knowledge over generations, and it must ensure that the benefits of AI-enabled classification accrue to those communities as well as to commercial and regulatory actors.

Looking forward, the field needs more comparative evaluations of fusion architectures on well-curated benchmark datasets that include challenging taxonomic groups, more research on uncertainty-aware inference in phylogenetic settings, and more sustained investment in the maintenance of reference collections and data standards. It also needs honest conversations about the limits of automation in biodiversity science. Artificial intelligence can accelerate identification and reveal patterns that are invisible to human inspection, but it cannot replace the careful taxonomic work, field knowledge, and ethical deliberation that underpin meaningful classification. The goal is not to build a perfect classifier but to build a learning infrastructure that can serve science, conservation, and society with integrity.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
2. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanese, L. (2016). Methods for the integration of multi-omics data: Mathematical aspects. *BMC Bioinformatics*, 17(Suppl 2), 15.
3. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), e8124.
4. CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31), 12794–12797.
5. Mishra, P., Kumar, A., Nagireddy, A., Mani, D. N., Shukla, A. K., Tiwari, R., & Sundaresan, V. (2016). DNA barcoding: An efficient tool to overcome authentication challenges in the herbal market. *Plant Biotechnology Journal*, 14(1), 8–21.
6. Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S.

- W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S. C. H., van den Berg, C., Bogarin, D., ... Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE*, 6(5), e19254.
7. Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: Large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6), 817–836.
8. Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50, 71–91.
9. Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Suppl 6), 153.
10. Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
11. Wang, H., Cimen, E., Singh, N., & Buckler, E. (2020). Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, 54, 34–41.
12. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 115.
13. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119.
14. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
15. Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19, 43.
16. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
17. Fahlgren, N., Gehan, M. A., & Baxter, I. (2015). Lights, camera, action: High-throughput plant phenotyping is ready for a close-up. *Current Opinion in Plant Biology*, 24, 93–99.
18. Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1), 1–10.
19. Posey, D. A., & Dutfield, G. (1996). Beyond intellectual property: Toward traditional resource rights for indigenous peoples and local communities. *International Development Research Centre*.

20. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.
21. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
22. Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620–1633.