

Cell Wall-Derived Biomass Fingerprinting for Sustainable Identification and Valorization of Medicinal Plants

Nicolas Kerlsson

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
nicolas909@buffalo.edu

Ole R. Marshall

Department of Computer Science, George Mason University, Fairfax, VA, USA.
ole.marshall@gmu.edu

Abstract

Medicinal plant value chains face persistent identification ambiguity, adulteration risk, and unsustainable harvesting pressure, yet conventional authentication approaches often rely on DNA barcoding or metabolomic profiling that can be costly, environmentally sensitive, or poorly suited to processed biomass. This paper examines cell wall-derived biomass fingerprinting as a systems-oriented strategy for sustainable identification and valorization of medicinal plants. Rather than treating cell wall chemistry as a narrow analytical problem, the discussion positions it within integrated hardware, data infrastructure, machine learning, and governance arrangements. The structural trade-offs among spectroscopic acquisition, chromatographic validation, data harmonization, and model generalization are analyzed. The paper argues that cell wall components, including polysaccharide composition, lignin-related residues, and structural glycoprotein patterns, offer a stable, taxonomically informative signal for authentication and quality assessment. However, realizing this potential requires modular architectures that connect laboratory instruments, edge deployments, centralized repositories, and regulatory workflows. Fairness concerns arise when model training sets overrepresent commercial species while underrepresenting wild relatives or regionally important taxa. Policy mechanisms, incentive alignment, and ongoing robustness auditing are therefore necessary to prevent fingerprinting systems from reproducing asymmetries in global botanical supply chains. The analysis concludes that cell wall-derived fingerprinting can support more transparent and resilient medicinal plant systems only when technical validation, data governance, and livelihood considerations are jointly designed.

Keywords

Cell wall fingerprinting; medicinal plant authentication; biomass valorization; machine learning; data governance; sustainable supply chains.

1. Introduction

The sustainable identification and valorization of medicinal plants has become a pressing systems problem at the intersection of biodiversity conservation, pharmaceutical supply chains, and digital verification infrastructure. Medicinal plant materials are often traded as dried roots, bark, seeds, or fragmented biomass that retains limited morphological identity. This creates opportunities for substitution, adulteration, or unintentional misidentification, particularly when multiple species share common vernacular names or visual characteristics.

Recent advances in natural product research have expanded the analytical toolbox for characterizing such materials, but many approaches remain fragmented across laboratory workflows, regional quality standards, and computational platforms [1]. The result is a structural gap between the increasing capacity to measure molecular features and the ability to deploy those measurements in decentralized, low-resource supply chains where medicinal plant authentication is most needed.

A systems perspective is necessary because identification is not merely a technical classification task. It is embedded in agronomic practices, harvest regulation, postharvest processing, trade networks, and quality assurance regimes. Deep metabolome annotation has improved chemical profiling of complex plant matrices, yet the full potential of such annotation depends on standardized data models and interoperable repositories that can accommodate heterogeneous samples and instrumentation [2]. Cell wall-derived biomass fingerprinting offers a complementary strategy because cell wall polymers are relatively stable under drying, storage, and moderate processing. Unlike many secondary metabolites that may degrade or fluctuate with environmental conditions, structural carbohydrates and associated wall components can preserve taxonomically informative patterns over longer temporal horizons. This stability is particularly relevant for processed medicinal plant products that have lost morphological and genomic markers.

The safety and efficacy of herbal medicinal products require integrating ethnopharmacological knowledge, phytochemical analysis, and reproducible analytical science [3]. However, reproducibility is difficult to achieve when datasets are generated through proprietary protocols, stored in isolated databases, or described with incomplete metadata. The FAIR Guiding Principles provide a useful foundation for data stewardship by emphasizing findability, accessibility, interoperability, and reuse across scientific and industrial boundaries [4]. Applying FAIR principles to cell wall fingerprinting would require explicit description of biomass preparation, instrument parameters, feature extraction methods, and model validation procedures. Without such governance, even technically sound fingerprinting systems may fail to scale across institutions and jurisdictions.

This paper examines cell wall-derived biomass fingerprinting as an integrated sociotechnical infrastructure. It considers analytical architecture, data integration, machine learning, governance, robustness, and policy implications. The emphasis is on structural trade-offs rather than on specific chemical protocols. By linking biomass chemistry, digital infrastructure, and institutional design, the discussion aims to clarify how fingerprinting systems can support sustainable identification without reproducing existing asymmetries in botanical supply chains.

2. Conceptual Foundations and Systems Framing

Cell wall-derived fingerprinting differs from conventional phytochemical authentication because its primary signal arises from structural biomass rather than from extractable metabolites alone. Plant cell walls are composed of cellulose, hemicelluloses, pectins, lignin or lignin-like polymers, and structural glycoproteins. The relative abundance, cross-linking patterns, and monosaccharide composition of these components vary across taxonomic groups, tissues, developmental stages, and processing conditions. Because these components are integral to biomass architecture, they offer a multidimensional signal that may remain detectable when softer chemical markers disappear. For medicinal plants, this creates an opportunity to classify species or populations based on cell wall component characteristics even after the raw material has been dried, milled, or partially extracted.

Conceptually, the approach can be understood as a high-dimensional pattern recognition problem in which biomass is converted into spectral, chromatographic, or imaging features and then mapped to taxonomic or quality labels. Deep learning has transformed such mapping tasks across domains by learning hierarchical representations from raw or minimally processed inputs [5]. In the context of biomass fingerprinting, deep models can potentially discover nonlinear combinations of wall-related features that are not visible through simple peak-based comparisons. The theoretical foundations of representation learning further suggest that model capacity must be matched with appropriate regularization, data diversity, and validation to avoid overfitting to narrow sample sets [6]. This is especially important when the number of authenticated reference samples is limited, as is frequently the case for rare or regionally restricted medicinal plants.

Visualization and interpretation tools are also central to system-level design. Dimensionality reduction methods enable analysts to explore clustering patterns among species, detect outliers, and communicate model behavior to regulators or botanical experts [7]. However, visualization is not evidence of robustness. A fingerprinting system must separate within-species variability from between-species differentiation, which requires careful sampling across geographic origins, harvest years, and processing batches. Without such separation, clustering may reflect batch effects rather than taxonomic identity. The systems framing therefore treats cell wall fingerprinting not as a static classifier but as a continuously updated infrastructure that must absorb environmental variability while maintaining decision boundaries.

A further conceptual challenge is that cell wall fingerprints are not inherently meaningful as chemical entities alone. Their value emerges when linked to reference metadata, provenance records, and quality outcomes. This relational property distinguishes a systems approach from a purely analytical chemistry perspective. A robust fingerprint is useful only within an institutional arrangement that defines what the fingerprint means for identity, purity, safety, or sustainability. Consequently, infrastructure design must address how reference data are generated, who maintains them, and how uncertainties are represented when classifying unknown samples.

3. Architectural Design for Cell Wall Fingerprinting Systems

A practical cell wall fingerprinting architecture typically spans multiple layers: sample acquisition, instrumental measurement, feature extraction, data harmonization, model training, inference, and reporting. Each layer introduces trade-offs between resolution, cost, scalability, and interpretability. Spectroscopic methods such as infrared or Raman spectroscopy can provide rapid, non-destructive signals related to polysaccharide and lignin composition, while chromatographic methods offer more detailed monosaccharide and linkage information but at higher cost and lower throughput. The choice between these modalities is not purely technical; it affects whether fingerprinting can be deployed near harvest sites, at border inspection points, or in centralized laboratories.

Computational support for such architectures increasingly relies on open machine learning libraries that provide standardized routines for preprocessing, feature selection, and cross-validation [8]. Deep learning frameworks allow more flexible model development but also demand greater computational resources and data engineering capacity [9]. The architectural question is therefore not simply which algorithm performs best on a curated benchmark, but which combination of sensing and computing can be maintained reliably under real-world constraints. For processed biomass, preprocessing is a decisive step because baseline shifts,

retention time drift, and particle size variation can introduce systematic noise. Advanced alignment and preprocessing methods must be integrated into the pipeline before modeling to reduce technical variance without removing biologically relevant signal [10].

Hyphenated analytical techniques can provide complementary structural information that strengthens cell wall fingerprinting systems [11]. By coupling separation with detection, these methods allow more precise assignment of sugar residues, phenolic domains, and other wall-related features. However, hyphenated instruments are often expensive and require specialized maintenance. In contrast, DNA barcoding has emerged as a powerful authentication tool for raw plant material, yet DNA may degrade in processed biomass or be absent in certain extracts [12]. Metabolomics, while highly informative, can be sensitive to postharvest handling and extraction conditions [13]. Cell wall fingerprinting occupies a middle space between these methods, offering stability under processing while still requiring complementary validation against genetic and metabolomic references.

The architecture should therefore be multimodal rather than dependent on a single analytical signal. A modular design can allow rapid spectroscopic screening to trigger more detailed chromatographic or mass spectrometric confirmation when uncertainty exceeds a defined threshold. This tiered approach reduces cost while preserving diagnostic confidence. It also enables different actors, such as smallholder cooperatives, commercial testing laboratories, and regulatory agencies, to participate according to their technical capacity. The underlying information system must preserve traceability across these tiers, ensuring that screening results and confirmatory analyses remain linked to the same physical sample and provenance record.

4. Data Infrastructure and Machine Learning Integration

Data infrastructure is the most overlooked constraint in fingerprinting systems. A model trained on one instrument type, one geographic region, or one processing tradition may fail when transferred to another context. Algorithmic fairness must therefore be integrated at the data curation stage rather than treated as a post hoc correction [14]. This means ensuring that reference libraries include wild relatives, locally used species, and economically marginal taxa, not only commercially dominant medicinal plants. A fingerprinting system that is highly accurate for a small number of export crops but unreliable for regionally important species may reinforce market concentration and erode local knowledge systems.

Cell wall component profiling has demonstrated that polysaccharide-associated characteristics can serve as stable discrimination inputs for closely related medicinal plant species [15]. Such findings are encouraging, but they also highlight the need for representative training data. If the cell wall features that distinguish a particular species are only characterized in samples from one province or one cultivation system, the resulting model may mistake regional environmental variation for taxonomic difference. Workflow environments that support visual programming and reproducible data mining can help domain scientists maintain transparency and auditability across complex pipelines [16]. These environments allow analytical steps to be documented and rerun without requiring advanced programming expertise.

Integration with broader biomedical knowledge is also relevant for valorization. Bioinformatics platforms developed for traditional Chinese medicine have shown how network-oriented analysis can connect phytochemical composition to putative molecular mechanisms [17]. Linking cell wall fingerprints to such platforms could strengthen the functional interpretation of quality markers, although the connection between structural wall

components and pharmacological activity remains indirect. The primary role of cell wall fingerprinting in valorization is therefore not to prove bioactivity but to provide reliable identity and quality anchors that downstream pharmacological studies can reference.

Supply chain integration introduces additional data governance demands. Distributed ledger technologies have been proposed to improve transparency and reduce fraud in complex supply chains [18]. In medicinal plant systems, blockchain-based records could theoretically link a cell wall fingerprint to a specific harvest batch, processor, and certification event. However, blockchain does not solve the oracle problem: the recorded fingerprint is only as trustworthy as the sensor, reference library, and governance process that generated it. A hybrid approach combining analytical verification with tamper-resistant logging may be more realistic than fully automated provenance claims. The information architecture must therefore distinguish between data integrity and analytical correctness, because both are necessary for trustworthy authentication.

5. Governance, Fairness, and Sociotechnical Implications

Governance arrangements determine who can define reference standards, who bears the cost of generating authenticated samples, and who benefits from the resulting verification systems. In many medicinal plant value chains, wild harvesters and smallholder cultivators possess extensive taxonomic knowledge but may have little influence over digital standardization. If fingerprinting reference libraries are built solely by distant laboratories or large commercial actors, local knowledge can be marginalized even while local materials are being classified. This raises procedural fairness concerns that go beyond model accuracy. Governance mechanisms should include benefit-sharing provisions, participatory reference sample collection, and transparent procedures for updating taxonomic reference data.

Ethical frameworks for artificial intelligence increasingly emphasize accountability, transparency, and societal impact [19]. These frameworks are relevant to medicinal plant fingerprinting because the technology can influence market access. A false negative classification may exclude a legitimate batch from trade, while a false positive may allow adulterated material to enter pharmaceutical supply chains. The distribution of these errors has direct economic and public health consequences. Regulatory regimes must therefore specify acceptable error rates, confirmatory testing requirements, and liability pathways rather than assuming that algorithmic outputs are self-evidently authoritative. Continuous auditing, including blind testing and interlaboratory comparisons, is necessary to maintain trust.

Fairness also includes geographic and ecological dimensions. Fingerprinting systems may be optimized for species that are commercially significant in North America, Europe, or East Asia, while species important to South Asia, Africa, or Latin America remain underrepresented. This is not necessarily intentional, but it emerges from funding streams, publication priorities, and access to analytical instrumentation. A fairness-oriented design would actively include regional herbaria, ethnobotanical institutes, and community-based conservation organizations in the creation of reference collections. Such inclusion improves taxonomic coverage and simultaneously strengthens the legitimacy of the system.

Regulatory harmonization remains challenging because medicinal plant products are governed under different frameworks across countries. Some jurisdictions treat them as dietary supplements, others as medicines, and others as agricultural commodities. These differences affect which quality attributes must be verified and which analytical methods are accepted. Cell wall fingerprinting can serve as a bridging technology if its outputs are mapped

to multiple regulatory categories rather than encoded into a single certification standard. This requires careful metadata modeling that preserves the distinction between identity, purity, geographic origin, and processing method.

6. Deployment, Robustness, and Sustainability Trade-Offs

Deployment of cell wall fingerprinting systems involves trade-offs between precision, accessibility, and long-term maintenance. High-resolution laboratory equipment produces detailed fingerprints but cannot be widely distributed. Portable spectrometers can be deployed in field settings, but their spectral resolution may be insufficient for closely related species. Hybrid models can address this tension by using field-level screening to identify ambiguous cases and centralized laboratory confirmation to resolve uncertainty. Such architectures require reliable communication channels, sample tracking, and feedback loops that update the screening model as new confirmed cases accumulate.

Robustness across instruments and environmental conditions remains a central concern. Preprocessing algorithms can reduce variability, but they cannot fully eliminate instrumental drift or sample matrix effects. Dynamic metabolomic data analysis has emphasized the importance of modeling time-dependent and batch-dependent variation rather than treating it as simple noise [20]. Similarly, cell wall fingerprinting systems should incorporate batch correction and domain adaptation methods that allow models to transfer across instruments or seasons without catastrophic performance loss. Validation protocols should include external datasets collected under conditions that differ from the original training environment.

Sustainability involves more than environmental conservation; it also includes institutional and financial durability. A fingerprinting system that depends on expensive proprietary software, frequent recalibration by external specialists, or continuous cloud connectivity may not be sustainable for long-term use in remote producing regions. Open data standards, modular software, and offline-capable inference can reduce ongoing costs. At the same time, open systems are not automatically fair or secure. They require governance to manage data quality, prevent gaming of the fingerprinting process, and ensure that local participants retain access to the value generated by verification.

The circular economy potential of medicinal plant biomass also intersects with fingerprinting. If cell wall-derived features can guide the separation of high-value medicinal fractions from residual biomass suitable for biopolymers, energy, or soil amendments, then authentication and valorization become more tightly coupled. This coupling can improve resource efficiency and create new revenue streams for harvesters. However, it also creates risks if valorization incentives drive overharvesting. Fingerprinting systems should therefore be embedded within conservation frameworks that monitor harvest intensity and ecological impact, not only product identity.

7. Policy and Future Integration

Policy integration is needed to convert cell wall fingerprinting from a promising analytical approach into a widely accepted component of medicinal plant governance. Standards organizations, pharmacopeial committees, and trade authorities must reach consensus on how cell wall fingerprints are generated, validated, and interpreted. This does not require imposing a single instrument or algorithm, but it does require common performance metrics, reference materials, and reporting templates. Interlaboratory studies can establish reproducibility limits and identify sources of disagreement. Regulatory guidance should distinguish between

screening results and confirmatory identification, ensuring that rapid field methods are not used in isolation for high-stakes decisions.

Future integration should also consider the role of public research infrastructure. National botanical repositories, university herbaria, and international databases can serve as neutral custodians of reference fingerprints. Their participation reduces dependence on private data silos and increases the likelihood that reference libraries cover a broad range of species. Funding models should support long-term curation rather than only initial data generation. Without sustained curation, reference databases decay in usefulness even as analytical instruments improve.

Open science practices are valuable but insufficient on their own. Sharing data without metadata, version control, or uncertainty estimates can produce misleading reanalyses. The fingerprinting community should adopt minimum reporting standards that include sample provenance, preparation history, instrument configuration, and model validation procedures. Such standards make it easier to compare studies, identify biases, and reproduce findings across laboratories. They also enable regulators to assess whether a proposed fingerprinting method is fit for a specific legal purpose.

Ultimately, the most important future direction is the co-design of technical and institutional systems. Cell wall-derived biomass fingerprinting should be developed with the participation of botanists, chemists, data scientists, supply chain actors, and local communities. This participatory approach is slower and more complex than a purely technical deployment, but it is more likely to produce systems that remain accurate, fair, and institutionally legitimate over time. The goal is not only to classify biomass but to support medicinal plant value chains that are ecologically sustainable and socially accountable.

8. Conclusion

Cell wall-derived biomass fingerprinting offers a distinctive systems pathway for sustainable identification and valorization of medicinal plants. By targeting structurally stable wall components rather than labile metabolites or degradable DNA, it provides a complementary signal that can persist through drying, storage, and moderate processing. However, the technical promise of this approach cannot be realized through isolated analytical improvements. It demands a coherent architecture that integrates sample handling, instrumental measurement, preprocessing, machine learning, reference databases, and governance mechanisms. The structural trade-offs between speed and resolution, central and edge deployment, proprietary and open infrastructure, and local and global reference coverage must be managed deliberately.

A robust fingerprinting system must be designed for fairness as well as accuracy. Training data should include wild relatives, regionally significant species, and materials from diverse production systems. Governance arrangements should ensure that the communities who steward medicinal plant biodiversity participate in the systems that classify and valorize their resources. Policy frameworks should support transparency, auditing, and continuous validation rather than treating algorithmic outputs as fixed authorities. When these technical, institutional, and ethical dimensions are jointly addressed, cell wall-derived biomass fingerprinting can become more than a classification tool. It can function as part of a larger infrastructure for sustainable medicinal plant economies, supporting reliable authentication, responsible resource use, and equitable participation in global value chains.

References

1. Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H., & Bertrand, S. (2019). Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. *Analytical Chemistry*, 91(1), 704–742.
2. Allard, P.-M., Genta-Jouve, G., & Wolfender, J.-L. (2017). Deep metabolome annotation in natural products research: Towards a generalizable approach. *Current Opinion in Chemical Biology*, 36, 40–49.
3. Heinrich, M. (2015). Quality and safety of herbal medicinal products: The need for integrating ethnopharmacology, phytochemistry, and analytical science. *Journal of Pharmacy and Pharmacology*, 67(3), 303–308.
4. 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.
5. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
7. van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
9. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates.
10. Christin, C., Smilde, A. K., Hoefsloot, H. C. J., Suits, F., Bischoff, R., & Horvatovich, P. L. (2013). Optimized time alignment algorithm for liquid chromatography-mass spectrometry data: Correlation optimized warping using component detection algorithm-selected mass chromatograms. *Analytical Chemistry*, 85(6), 3383–3390.
11. Sarker, S. D., & Nahar, L. (2012). Hyphenated techniques and their applications in natural products analysis. In *Natural Products Isolation* (pp. 171–198). Humana Press.
12. Ganie, S. H., Upadhyay, P., Das, S., & Prasad Sharma, M. (2015). Authentication of medicinal plants by DNA markers. *Plant Gene*, 4, 83–99.
13. Aliferis, K. A., & Jabaji, S. (2011). Metabolomics – a robust bioanalytical approach for the discovery of the modes of action of pesticides: A review. *Pesticide Biochemistry and Physiology*, 100(2), 105–117.

14. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). Association for Computing Machinery.
15. 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.
16. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications* (pp. 319–326). Springer.
17. Liu, Z., Guo, F., Wang, Y., Li, C., Zhang, X., Li, H., Diao, L., Gu, J., Wang, W., Li, D., & He, F. (2016). BATMAN-TCM: A bioinformatics analysis tool for molecular mechanism of traditional Chinese medicine. *Scientific Reports*, 6, 21146.
18. Kshetri, N. (2018). Blockchain's roles in meeting key supply chain management objectives. *International Journal of Information Management*, 39, 80–89.
19. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
20. Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., Jellema, R. H., Pijl, H., Roelfsema, F., & van der Greef, J. (2010). Dynamic metabolomic data analysis: A tutorial review. *Metabolomics*, 6(1), 3–17.