

Intelligent Identification of Plant-Derived Biomass Components for Pharmaceutical and Biorefinery Applications

Neikhil Methur

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

nikhilmathur76@unr.edu

Deminak Backer

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.

dominik971@uc.edu

Abstract

The identification of plant-derived biomass components has traditionally depended on targeted analytical chemistry and empirical botanical expertise. As pharmaceutical discovery and biorefinery operations become more data-intensive and distributed, there is a growing need for intelligent identification systems that can integrate multimodal spectral, chemical, and genomic information while operating across heterogeneous feedstocks, scales, and regulatory environments. This paper examines the systems-level design of such identification infrastructures. It addresses the structural trade-offs among centralized high-resolution laboratory models, portable edge-deployable classifiers, and distributed learning frameworks. The discussion covers data acquisition architectures, deep representation learning, domain adaptation, chemometric preprocessing, and model interpretability. It further analyzes governance challenges related to data provenance, fairness, transparency, and interoperability across public repositories, industrial consortia, and regulatory agencies. The pharmaceutical and biorefinery sectors are compared to highlight divergent performance requirements: pharmaceutical applications prioritize specificity, reproducibility, and regulatory traceability, whereas biorefinery applications emphasize throughput, robustness to feedstock variability, and lifecycle sustainability. The analysis shows that sustainable intelligent identification depends less on any single algorithmic advance than on the deliberate alignment of sensing infrastructure, model architecture, data governance, and institutional policy. Future systems should therefore be designed as adaptive socio-technical infrastructures rather than isolated classification tools.

Keywords

intelligent identification; plant biomass; pharmaceutical analysis; biorefinery; machine learning; multimodal spectroscopy; data governance; sustainability.

1. Introduction

The transition toward sustainable biobased economies has intensified interest in plant-derived biomass as both a pharmaceutical resource and a feedstock for biorefinery processes. In biorefinery systems, lignocellulosic streams can be converted into fuels, chemicals, and materials, with feedstock composition directly influencing conversion yield and downstream processing [1]. In pharmaceutical research, plant extracts remain an important source of

bioactive secondary metabolites, and the identification of these compounds is a prerequisite for dereplication, standardization, and eventual clinical development [2]. These two application domains differ in their analytical traditions, yet both require reliable, interpretable, and scalable methods for identifying the molecular components of complex plant matrices.

Analytical methods for identifying plant biomass components have historically combined chromatographic separation, mass spectrometry, and vibrational spectroscopy. Near-infrared spectroscopy has become established as a rapid, nondestructive technique capable of capturing bulk chemical information from complex plant matrices [3]. Hyperspectral imaging extends this capacity by resolving spatial heterogeneity across plant tissues, enabling spatially resolved chemical mapping in agricultural and processing environments [4]. However, these instrument platforms generate high-dimensional signals whose interpretation is complicated by overlapping absorption bands, scattering effects, moisture variability, and the chemical diversity of plant metabolism.

Advances in machine learning have opened new possibilities for extracting compositional information from such signals. Deep learning models have been applied to complex spectral and image data, learning hierarchical representations that can outperform traditional linear chemometric methods under nonlinear conditions [5][6]. Similarly, machine learning is increasingly used in molecular and materials science to relate chemical structure to function, suggesting pathways for identifying plant-derived metabolites from multimodal chemical data [7]. Nevertheless, the integration of these capabilities into reliable systems for biomass identification requires attention to architecture, infrastructure, and governance rather than algorithmic development alone.

2. Plant-Derived Biomass as a Systems Challenge

Plant biomass is not a stable analytical object. Its chemical composition varies with genotype, tissue type, developmental stage, harvest timing, soil chemistry, climate, and postharvest handling. A leaf, root, bark, or seed sample may contain thousands of metabolites distributed across multiple chemical classes, including phenolics, alkaloids, terpenoids, polysaccharides, lignin, lipids, and specialized signaling molecules. In pharmaceutical applications, the target may be a single active compound that must be distinguished from numerous structural analogues. In biorefinery applications, the relevant output is often a bulk compositional profile that must be estimated rapidly to optimize pretreatment, fermentation, or catalytic conversion. Thus, intelligent identification is not merely a classification task; it is a systems problem involving heterogeneous measurements, uncertain labels, evolving biological variation, and multiple operational objectives.

A further complication arises from the distributed nature of biomass supply chains. Feedstock may be sourced from smallholder farms, managed forests, agroforestry systems, or botanical gardens, each with different collection and documentation practices. The resulting data are rarely uniform. Some samples are characterized by high-resolution mass spectrometry in centralized laboratories, others by portable spectrometers at collection sites, and still others by imaging platforms mounted on vehicles or drones. These modalities differ in sensitivity, resolution, cost, and interpretability. An intelligent identification system must therefore function as an integrating layer that can align data of varying quality and provenance while preserving enough traceability for downstream decision-making.

The pharmaceutical and biorefinery domains also impose distinct regulatory and economic constraints. Pharmaceutical identification systems are embedded in quality-by-design

frameworks, good manufacturing practice, and pharmacopoeial standards. Their outputs may inform clinical development, quality control, or forensic authentication. Biorefinery systems, by contrast, operate under tighter margins and higher throughput requirements, where fast compositional screening can determine whether a batch is economically viable. A single architectural design is unlikely to satisfy both domains without substantial contextual adaptation. This tension motivates a system-level discussion of sensing infrastructure, model architecture, and governance.

3. Data Acquisition and Multimodal Sensing Infrastructure

Data acquisition for intelligent biomass identification increasingly depends on distributed sensing networks. In forest and agricultural settings, Internet of Things platforms support continuous monitoring of biomass quality, environmental conditions, and logistics parameters [8]. Large-scale digital agriculture and food systems similarly rely on heterogeneous sensor arrays, cloud storage, and real-time analytics to convert field observations into operational decisions [9]. Plant phenotyping infrastructures illustrate how imaging, spectroscopy, and environmental sensors can be organized around common data models and shared experimental protocols [10]. These sensing infrastructures provide the raw material for identification models, but they also introduce challenges related to calibration, metadata standardization, and data drift.

At the laboratory scale, high-resolution methods such as liquid chromatography coupled to tandem mass spectrometry offer detailed structural information but are expensive and low-throughput. Portable vibrational spectroscopy instruments can be deployed near field sites or biorefinery intake points, offering rapid measurements with lower chemical specificity. Hyperspectral cameras can map spatial variation across plant tissues or processed biomass streams, but they generate extremely large data volumes that require substantial preprocessing. The design of an intelligent identification system therefore involves a trade-off among analytical depth, spatial coverage, temporal frequency, and operational cost. No single sensor modality is sufficient for all pharmaceutical and biorefinery tasks, and the most robust systems are typically multimodal.

However, the value of these sensing systems depends on the availability of well curated reference data. The FAIR principles for scientific data management call for data to be findable, accessible, interoperable, and reusable, which is particularly challenging for plant biomass because of its biological variability and the diversity of analytical protocols [11]. Cell wall component characteristics have been explored as stable identification markers for closely related medicinal species, illustrating how biochemical signatures can support robust classification when combined with appropriate data structures [12]. Such marker-based approaches can complement spectral and metabolomic data by providing a relatively stable chemical fingerprint that persists across environmental variation.

4. Intelligent Identification Architectures

Once multimodal data are acquired, identification systems may adopt several architectural patterns. End-to-end deep learning approaches have been developed for quantitative spectral analysis, allowing models to learn feature extraction and regression jointly from raw or minimally processed spectra [13]. Geometric deep learning generalizes neural networks to non-Euclidean structures such as molecular graphs and point clouds, which is relevant when biomass components are represented as chemical graphs or three-dimensional structural data [14]. Tensor-based neural networks have further shown that quantum-chemical and structural

information can be encoded in learned representations for molecular property prediction [15]. These model families share a capacity to capture nonlinear interactions but differ substantially in their data requirements, interpretability, and computational cost [16].

Centralized architectures often achieve high accuracy because they can be trained on large, carefully curated datasets with substantial computational resources. They are suitable for pharmacopoeial standards development, retrospective analysis of voucher specimens, and regulatory reference libraries. However, centralized models may be impractical for field deployment where connectivity is limited, latency must be low, and privacy or proprietary concerns restrict raw data sharing. In such settings, edge-deployable classifiers can operate directly on portable instruments, providing immediate screening results without transmitting sensitive spectra to a central server. The trade-off is that edge models may sacrifice accuracy, particularly on rare species or unusual chemical profiles.

A third architectural pattern involves distributed or federated learning, in which model parameters are shared across institutions while raw data remain local. This approach can improve generalization across geographic regions and supply chains without requiring the construction of a monolithic central database. Yet federated systems introduce additional complexity in model aggregation, heterogeneity handling, and incentive alignment. Participating organizations must agree on common data schemas, label definitions, and performance metrics. These institutional agreements are often more difficult to design than the machine learning algorithms themselves.

5. Representation Learning and Biomolecular Feature Extraction

Before high-level representations are learned, careful preprocessing of spectral data is necessary to reduce physical artifacts such as baseline drift and multiplicative scattering. Reviews of preprocessing methods for near-infrared spectra show that decisions about smoothing, derivative transformation, and normalization can determine model generalization to new instruments and sample batches [17]. Chemometric preprocessing is thus not a preliminary detail but a structural component of the identification pipeline. The same principle applies to chromatographic and mass spectrometric data, where alignment, peak picking, and noise filtering shape the molecular features available to downstream models.

Representation learning can then transform processed spectra or molecular descriptors into task-relevant feature spaces. In pharmaceutical identification, representations may be optimized to distinguish structurally similar metabolites, requiring supervision from high-quality reference standards. In biorefinery applications, representations may be optimized for compositional regression, such as estimating cellulose, hemicellulose, lignin, or ash content. These different objectives imply different loss functions, architectural choices, and validation strategies. Transfer learning can reduce the need for large labeled datasets by adapting representations learned on related chemical tasks to new plant species or analytical instruments.

Interpretability remains a central concern. A model that identifies a medicinal species from near-infrared spectra may be accurate but may also rely on subtle, nonintuitive spectral features that cannot be easily linked to known cell wall components or metabolites. In pharmaceutical quality control, such black-box behavior conflicts with regulatory expectations for mechanistic understanding and auditability. Saliency maps, spectral feature attribution, and prototype-based explanations can help, but these techniques require careful validation. An interpretable identification system must demonstrate not only which spectral

regions contribute to a decision but also whether those regions correspond to chemically meaningful signals.

6. Structural Trade-offs in Deployment and Scaling

Deployment of intelligent identification systems across diverse regions also raises concerns about algorithmic bias and fairness. Machine learning models trained on narrow geographic or species datasets may systematically underperform on underrepresented feedstocks, leading to inequitable outcomes in agricultural supply chains [18]. For example, a model trained primarily on temperate plant species may fail when applied to tropical medicinal plants, not because the underlying chemistry is fundamentally different but because the training distribution did not include relevant spectral variability. This form of distributional bias can reinforce existing inequalities in agricultural research and biobased value chains.

Scaling intelligent identification from a laboratory demonstration to operational infrastructure requires more than improved accuracy. It requires reliable instrument calibration across devices, robust handling of missing or noisy measurements, version control for models and reference data, and clear protocols for model updates. A model deployed in a biorefinery may need to adapt to seasonal variation in feedstock composition, while a model used in pharmaceutical authentication may need to remain stable under strict regulatory oversight. These requirements can conflict, because frequent model updates improve adaptability but complicate validation and traceability.

System designers must also consider the division of labor between human expertise and automated inference. In many pharmaceutical laboratories, chemists and botanists still make final identification decisions. An intelligent system should therefore be designed as a decision-support instrument rather than a replacement for expert judgment. This implies careful attention to confidence calibration, uncertainty reporting, and the design of user interfaces that allow experts to inspect and override model outputs. In biorefineries, where decisions occur more rapidly, automated alerts and control actions may be acceptable, but only if the system includes safeguards against sensor failure and adversarial or anomalous inputs.

7. Governance, Fairness, and Interoperability

Governance of AI-based identification systems extends beyond technical correctness to include ethical principles such as beneficence, non-maleficence, autonomy, justice, and explicability [19]. In the context of plant biomass, beneficence might involve improving access to high-quality pharmaceutical raw materials or enabling sustainable bioenergy. Non-maleficence requires avoiding harm from misidentification, such as the release of adulterated medicinal products or the misrouting of biomass destined for regulated markets. Justice concerns arise when the benefits of intelligent identification are concentrated among well-funded laboratories and industrial actors while smallholder producers or traditional knowledge holders remain excluded.

Datasheets for datasets have been proposed as an institutional mechanism for documenting data provenance, biases, and recommended uses, making algorithmic behavior more auditable across organizations [20]. In plant biomass identification, such documentation might include geographic origin, species identity verification, voucher specimen references, collection date, drying and storage conditions, analytical instrument settings, and known chemical interferences. Without this metadata, it is difficult to know whether a model can safely be transferred from one region or instrument to another. Standardized documentation also

supports regulatory acceptance, because it enables inspectors to reconstruct the lineage of training data and model decisions.

Biodiversity and ecological observation networks similarly depend on essential variables and shared data standards to ensure that remotely sensed and field measurements can be integrated across jurisdictions [21]. The same logic applies to plant-derived biomass. Harmonized standards for metabolite annotation, biomass composition, and provenance reporting would allow pharmaceutical, agricultural, and biorefinery data collections to be linked. Such interoperability does not require a single centralized database; it can be achieved through common metadata schemas, persistent identifiers, and application programming interfaces that allow distributed repositories to query one another.

8. Sustainability and Lifecycle Implications

The deployment of intelligent biomass identification systems should also be evaluated in relation to material and energy flows. The scale of global human-made mass now exceeds total living biomass, signaling the importance of biobased alternatives and efficient biomass use [22]. Intelligent identification can contribute to sustainability by reducing waste, enabling the use of underutilized feedstocks, and improving the traceability of plant-derived products. However, the computational infrastructure required for large-scale model training and continuous monitoring also consumes energy and materials. A systems perspective must account for these lifecycle impacts rather than assuming that digitalization is automatically sustainable.

Transitions toward biobased industry are not only technological but involve changes in institutions, user practices, and policy regimes. Sustainability transitions scholarship emphasizes that technological innovations become viable only when embedded in supportive regulatory and market structures [23]. For intelligent biomass identification, this means that advances in spectral analysis and machine learning must be accompanied by standards, incentives, and procurement rules that reward robust and transparent systems. Without such institutional support, promising technical capabilities may remain fragmented and underused.

Responsible innovation frameworks further demand anticipation, inclusion, reflexivity, and responsiveness in the development of emerging technologies [24]. Anticipation requires examining how intelligent identification might reshape labor, supply chains, and access to biological resources. Inclusion requires involving stakeholders from pharmaceutical regulation, biorefinery operations, biodiversity conservation, and smallholder agriculture. Reflexivity demands that system designers question their own assumptions about accuracy, efficiency, and progress. Responsiveness means that governance arrangements should be capable of adapting as new evidence emerges about model performance, environmental impacts, or social consequences.

9. Cross-Domain Comparisons and Forward-Looking Perspectives

Comparing pharmaceutical and biorefinery applications reveals how the same underlying identification technology can require fundamentally different system configurations. Pharmaceutical systems tend to favor centralized, high-resolution, and tightly validated models that operate within strict regulatory frameworks. Their primary risks are false positives that might allow adulterated or misidentified plant material into clinical or commercial products. Biorefinery systems tend to favor distributed, rapid, and adaptive models that can tolerate higher uncertainty because subsequent processing stages can buffer compositional variation. Their primary risks are economic inefficiency and process instability

rather than direct harm to patients. These differences should guide architecture selection, data governance, and performance validation.

Future systems are likely to integrate digital twins of biomass supply chains, in which compositional models are continuously updated with real-time sensor data. Such digital twins could support predictive control of biorefineries, dynamic routing of feedstock, and verification of sustainable sourcing claims. In pharmaceutical discovery, similar infrastructures could accelerate the screening of plant extracts while preserving the evidentiary standards required for regulatory submission. Achieving this vision will require not only better algorithms but also stronger institutional frameworks for data sharing, model auditing, and cross-sector collaboration.

A forward-looking perspective must also recognize the value of traditional botanical and ecological knowledge. Intelligent identification systems trained exclusively on instrumental data may fail to capture the practical distinctions that local communities use to identify useful or hazardous plants. Incorporating this knowledge into formal data structures and validation processes is difficult, but it can improve model robustness and social legitimacy. It also raises questions about data sovereignty, benefit sharing, and the governance of digital sequence information, all of which extend well beyond the technical design of classifiers.

10. Conclusion

Intelligent identification of plant-derived biomass components sits at the intersection of analytical chemistry, machine learning, industrial process design, and environmental governance. The central challenge is not the creation of a more accurate classifier in isolation, but the construction of a coherent socio-technical infrastructure that can support identification across pharmaceutical and biorefinery domains. This requires aligning multimodal sensing, data standards, representation learning, model deployment, and institutional oversight. Different structural choices produce different trade-offs among accuracy, speed, interpretability, fairness, and resource consumption. No single architecture is universally optimal. Instead, systems should be designed with explicit attention to the contexts in which decisions are made, the risks associated with misidentification, and the distribution of benefits across actors. By approaching intelligent identification as an infrastructural problem, academic researchers, industrial practitioners, and policymakers can jointly develop systems that are not only technically capable but also robust, transparent, and sustainable.

References

1. Naik, S. N., Goud, V. V., Rout, P. K., & Dalai, A. K. (2010). Production of first and second generation biofuels: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 14(2), 578-597. <https://doi.org/10.1016/j.rser.2009.10.003>
2. Wolfender, J.-L., Marti, G., Thomas, A., & Bertrand, S. (2015). Current approaches and challenges for the metabolite profiling of complex natural extracts. *Journal of Chromatography A*, 1382, 136-164. <https://doi.org/10.1016/j.chroma.2014.10.091>
3. Workman, J., & Weyer, L. (2008). *Practical guide to interpretive near-infrared spectroscopy*. CRC Press.
4. Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, 18(12), 590-598. <https://doi.org/10.1016/j.tifs.2007.06.001>

5. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195-204. <https://doi.org/10.1038/s41586-019-0912-1>
6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
7. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547-555. <https://doi.org/10.1038/s41586-018-0337-2>
8. Singh, R., Gehlot, A., Akram, S. V., Thakur, A. K., Buddhi, D., & Das, P. K. (2022). Forest 4.0: Digitalization of forest using the Internet of Things (IoT). *Journal of King Saud University - Computer and Information Sciences*, 34(8), 5587-5601. <https://doi.org/10.1016/j.jksuci.2021.02.009>
9. Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2022). IoT, big data, and artificial intelligence in agriculture and food industry. *IEEE Internet of Things Journal*, 9(9), 6305-6324. <https://doi.org/10.1109/JIOT.2020.2998584>
10. Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., & Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Current Biology*, 27(15), R770-R783. <https://doi.org/10.1016/j.cub.2017.05.055>
11. 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. <https://doi.org/10.1038/sdata.2016.18>
12. 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.
13. Yang, J., Xu, J., Zhang, X., Wu, C., Lin, T., & Ying, Y. (2019). Deep learning for vibrational spectral analysis: Recent progress and a practical guide. *Analytica Chimica Acta*, 1081, 6-17. <https://doi.org/10.1016/j.aca.2019.06.012>
14. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4), 18-42. <https://doi.org/10.1109/MSP.2017.2693418>
15. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8, 13890. <https://doi.org/10.1038/ncomms13890>
16. Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16), 1291-1307. <https://doi.org/10.1002/jcc.24764>
17. Rinnan, Å., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10), 1201-1222. <https://doi.org/10.1016/j.trac.2009.07.007>

18. 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), 1-35.
<https://doi.org/10.1145/3457607>
19. Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.
<https://doi.org/10.1007/s11023-018-9482-5>
20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://doi.org/10.1145/3458723>
21. Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., Bruford, M. W., Brummitt, N., Butchart, S. H. M., Cardoso, A. C., Coops, N. C., Dulloo, E., Faith, D. P., Freyhof, J., Gregory, R. D., Heip, C., Höft, R., Hurtt, G., Jetz, W., ... Wegmann, M. (2013). Essential biodiversity variables. *Science*, 339(6117), 277-278.
<https://doi.org/10.1126/science.1229931>
22. Elhacham, E., Ben-Uri, L., Grozovski, J., Bar-On, Y. M., & Milo, R. (2020). Global human-made mass exceeds all living biomass. *Nature*, 588(7838), 442-444.
<https://doi.org/10.1038/s41586-020-3010-5>
23. Geels, F. W. (2011). The multi-level perspective on sustainability transitions: Responses to seven criticisms. *Environmental Innovation and Societal Transitions*, 1(1), 24-40.
<https://doi.org/10.1016/j.eist.2011.02.002>
24. Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568-1580.
<https://doi.org/10.1016/j.respol.2013.05.008>