

AI-Enabled Metabolite Marker Discovery for Authentication and Quality Control of Herbal Medicines

Anlong Lin

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

anlongwork@uc.edu

Bjorn Park

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

bjornwork@ucf.edu

Mason R. Oliver

Department of Computer Science, University of Houston, Houston, TX, USA.

contactmason@uh.edu

Abstract

The authentication and quality control of herbal medicines present persistent analytical and regulatory challenges because herbal materials exhibit complex chemical profiles, taxonomic ambiguity, and substantial variation arising from geographic origin, cultivation practice, harvest timing, and post-harvest processing. Metabolomics has emerged as a powerful strategy for capturing this chemical complexity, yet the transformation of high-dimensional spectral data into reliable, operationally usable marker panels remains difficult. Artificial intelligence offers considerable promise for metabolite marker discovery by learning latent patterns from large analytical datasets, but its deployment in herbal medicine quality systems requires attention to far more than predictive accuracy. This paper examines the system-level architecture of AI-enabled marker discovery, emphasizing data infrastructure, standardization, modeling trade-offs, interpretability, integration into authentication workflows, robustness, fairness, governance, and sustainable deployment. It argues that AI-based marker discovery should be understood as a socio-technical infrastructure rather than as an isolated computational exercise. The discussion connects analytical chemistry, machine learning, regulatory science, and supply chain governance to identify structural conditions under which AI can produce defensible and durable markers for herbal authentication. Particular attention is given to the need for metadata completeness, transparent model explanations, staged validation across biological and operational variability, and alignment with international regulatory expectations. The paper concludes by outlining forward-looking strategies for federated data governance, lightweight model deployment, and continuous post-market surveillance in distributed herbal supply chains.

Keywords

artificial intelligence; metabolomics; herbal medicine authentication; quality control; data infrastructure; regulatory governance.

1. Introduction

The global utilization of herbal medicines has expanded substantially, driven by consumer demand for complementary therapeutic options and the incorporation of traditional botanical preparations into national health systems. Authentication and quality control nevertheless remain difficult because herbal materials are chemically complex, taxonomically heterogeneous, and strongly influenced by environmental and processing factors. Metabolomics, which involves the comprehensive measurement of small-molecule metabolites in botanical and biological matrices, has become a central analytical approach for characterizing this complexity [1]. Advances in chromatographic separation, high-resolution mass spectrometry, and nuclear magnetic resonance spectroscopy now allow investigators to acquire highly detailed spectral profiles from herbal extracts [2]. These profiles, however, pose substantial interpretive difficulties because many spectral signals remain unidentified, and natural product metabolite annotation is frequently incomplete [3]. The central opportunity for artificial intelligence is to convert large, noisy, and high-dimensional analytical datasets into robust marker panels that can support authentication decisions and routine quality control operations.

Chemical information resources and reporting frameworks provide an essential foundation for this work. Ontology resources such as ChEBI support the normalization of metabolite identity across heterogeneous studies, reducing ambiguity in chemical names and structural annotations [4]. Minimum reporting standards for chemical analysis and for data analysis have been proposed to improve reproducibility and comparability across laboratories [5][6]. Open analytical platforms such as MetaboAnalyst have lowered barriers for exploratory metabolomics, but they also reveal the limitations of purely retrospective analysis when deployment requires stable predictive markers [7]. Herbal materials often contain hundreds to thousands of detectable chemical features, and conventional univariate marker selection is rarely sufficient to capture the multivariate signatures needed for reliable authentication. AI-enabled discovery has therefore emerged as a system-level response that integrates data processing, feature extraction, model training, validation, and deployment within a single pipeline.

Deep learning has transformed pattern recognition across many scientific and clinical domains [8]. In medicine, artificial intelligence systems increasingly support diagnostic workflows, biomarker interpretation, and clinical decision making [9][10]. These advances demonstrate that AI can learn useful latent representations from complex biomedical data, but they also highlight the need for careful validation and governance before deployment. In the context of herbal medicine authentication, AI methods must operate under stricter conditions than many exploratory research applications. They must generalize across batches and geographic origins, resist accidental or deliberate data shifts, and generate evidence that regulators can evaluate. This paper examines the system-level design of AI-enabled metabolite marker discovery for herbal medicines, with emphasis on architectural choices, data governance, integration into quality systems, robustness, fairness, policy implications, and long-term sustainability.

2. Data Infrastructure and Standardization for Herbal Metabolomics

The reliability of AI-discovered markers depends heavily on the quality and governance of the underlying data infrastructure. Herbal metabolomics data are generated by diverse instruments, laboratories, extraction protocols, and spectral processing pipelines. The resulting datasets are high-dimensional, sparse with respect to fully identified compounds, and strongly conditioned by analytical parameters. Curated compound databases provide

useful structural and spectral references for cross-domain metabolite annotation [1]. Chemical ontology resources additionally support the normalization of synonyms, structural relationships, and metadata, enabling integration across otherwise incompatible datasets [4]. These resources are necessary but not sufficient. They must be embedded in a governance framework that enforces metadata completeness, traceability, and provenance documentation.

Reporting standards play a critical role in converting raw analytical output into reusable scientific evidence. Minimum reporting standards for chemical analysis and for data analysis define the information needed to reproduce metabolomics experiments and to evaluate statistical models [5][6]. Such standards typically include descriptions of sample preparation, chromatographic conditions, spectral acquisition parameters, data processing steps, and model validation procedures. Without this metadata, AI systems may learn laboratory-specific artifacts rather than chemically meaningful signals. Batch effects, column aging, mobile phase variability, and instrument drift can all become correlated with sample class in unintentional ways. Data governance must therefore treat spectral acquisition and processing metadata as first-class inputs to the machine learning lifecycle rather than as peripheral documentation.

Data infrastructure must also account for botanical identity and provenance. Herbal medicines are frequently sourced from multiple regions, processed by different methods, and distributed through complex international supply chains. The same species can exhibit substantial phytochemical variation because of environmental stress, soil composition, and harvest timing. A marker discovery system built solely on samples from one region may fail when confronted with legitimate material from another region. Reference datasets should therefore include authenticated voucher specimens, geographic metadata, and detailed processing histories. International regulatory frameworks for botanical quality emphasize the need for consistent specifications and documented analytical procedures. These expectations reinforce the importance of data governance, because a marker discovered without defensible provenance cannot support a credible regulatory decision.

3. AI Architectures for Metabolite Marker Discovery

Several artificial intelligence architectures can support metabolite marker discovery, but their selection involves structural trade-offs rather than purely technical preferences. Classical machine learning approaches, particularly random forests, have been widely adopted because they tolerate high-dimensional and correlated input spaces and provide heuristic feature importance scores [11]. Such models are useful for initial screening when the number of candidate spectral features is very large. However, tree-based methods can be sensitive to irrelevant variation when the training set is small or improperly curated, and their importance measures may not reflect chemical interpretability. Deep neural networks, by contrast, learn hierarchical representations that can integrate adjacent spectral channels and detect subtle patterns across retention time and mass-to-charge dimensions [8]. The choice between these architectures should reflect the downstream need for a small set of interpretable markers that can be measured in routine quality control, not simply the maximization of classification accuracy. A parsimonious model with a small number of chemically identified markers is often preferable in regulatory and operational settings to a highly accurate but opaque model.

Transparency is essential because marker discovery feeds directly into regulatory documentation. Post hoc explanation algorithms such as SHAP and LIME can assign importance to features in complex models, allowing analysts to inspect whether a model relies on plausible spectral regions or known metabolites [12][13]. These explanation methods are approximate and can vary across training runs or data perturbations. International regulatory

frameworks for herbal medicinal products therefore emphasize documented analytical procedures, acceptance criteria, and scientific justification for selected markers [14][15][16]. The broader explainable artificial intelligence agenda adds a further expectation: explanations should be faithful to model behavior, understandable to domain experts, and actionable in deployment [17]. Architectural choices for marker discovery should therefore incorporate interpretability constraints from the outset rather than attempting to explain an unconstrained model after training.

A robust AI-enabled marker discovery architecture typically includes several coupled modules: spectral preprocessing, chemical annotation, feature extraction, model training, marker candidate ranking, and external validation. Each module introduces assumptions that affect downstream performance. Spectral alignment and normalization decisions can determine whether a model learns true chemical variation or batch-specific artifacts. Annotation workflows that rely on chemical ontologies and reference databases reduce false identity assignments but may discard novel signals that are not yet present in public libraries [3][4]. The system should therefore treat marker discovery as an iterative process in which model outputs are reviewed by analytical chemists and botanical taxonomists. This human-in-the-loop design is not a weakness; it is a necessary mechanism for aligning statistical signal with chemical and biological plausibility.

4. Integration into Authentication and Quality Control Systems

AI-discovered metabolite markers must be integrated into authentication workflows that combine multiple lines of evidence. A marker panel that distinguishes a target species from common adulterants is operationally useful only if the markers can be measured with acceptable specificity, sensitivity, and robustness under routine laboratory conditions. Metabolomics is not the only source of diagnostic information. Complementary marker classes, including cell wall component characteristics, have been proposed for distinguishing *Salvia miltiorrhiza* and its close relatives [18]. Such complementary markers illustrate the value of integrating metabolomic evidence with broader botanical and chemical profiling strategies. The system design challenge is to determine how AI-derived spectral markers, identified metabolite features, and structural or macromolecular markers should be weighted in a final authentication decision.

Translational metabolomics has demonstrated that moving from a statistically significant marker to a reliable deployed assay involves substantial additional validation [19]. Many candidate markers fail in field settings because of high biological variability, low analyte concentration, or instability during sample storage and preparation. Quality control systems therefore require staged validation: analytical validation of the measurement method, biological validation of marker specificity across geographic and seasonal variation, and operational validation under real supply chain conditions. Artificial intelligence can support each stage by modeling sources of variation, but it cannot replace the need for independently collected test sets and blinded reproducibility studies.

The broader conceptual framework of metabolomics has moved beyond simple biomarker identification toward mechanistic understanding of dynamic metabolic states [20]. For herbal medicines, this shift suggests that marker discovery should not treat chemical features as static labels. Markers may reflect biosynthetic responses to environmental stress, developmental stage, or post-harvest processing. AI systems that model the joint distribution of metabolite features across conditions are better able to distinguish informative variation from incidental noise. A marker that is mechanistically linked to a biosynthetic pathway may

be more stable across environments than a marker selected solely because it is predictive in a single dataset.

Historical work on chemical markers for herbal quality control also highlights the tension between marker specificity and practical measurability [21]. Many proposed chemical markers are not abundant, require sophisticated analytical instrumentation, or are not commercially available as reference standards. AI-enabled discovery can address this by explicitly optimizing for markers that are both discriminative and operationally feasible. This requires the model objective to include criteria such as detectability, chemical stability, and availability of reference compounds, rather than maximizing classification accuracy alone. Because such criteria are often available as structured metadata, they can be incorporated into the training and ranking process as constraints or utility functions.

5. Robustness, Fairness, and Risk Management

Robustness is a central concern for AI-enabled authentication systems because herbal materials vary across legitimate sources and can be subject to deliberate adulteration. A model that performs well on data from a single region or laboratory may generalize poorly when deployed across multiple supply chains. Distribution shifts can arise from changes in cultivation practice, drying methods, storage conditions, or analytical instrumentation. Robustness therefore requires representative training data, explicit monitoring for data drift, and periodic recalibration of deployed models. Without these mechanisms, an authentication system may produce false alarms for legitimate material or fail to detect sophisticated adulteration.

Fairness is closely related to robustness in this context. AI marker models may perform differently across geographic origins, leading to false adulteration flags for legitimate material from underrepresented regions. This is not merely a technical issue; it can have economic consequences for smallholder producers and traditional medicine suppliers. Fairness requires deliberate attention to the composition of training datasets and to the distribution of model performance across relevant subgroups. Governance structures should ensure that no region or production system is systematically disadvantaged by the deployment of AI-based quality control.

Risk management must also account for the asymmetric consequences of authentication errors. False positives can disrupt supply chains and damage producer livelihoods, while false negatives can allow adulterated products to reach consumers. Decision thresholds should be informed by the relative costs of these errors in specific regulatory and commercial contexts. Model calibration, audit trails, and independent review are necessary to support accountable decision making. In high-stakes settings, AI outputs should be treated as decision support rather than as automated determinations, particularly when marker evidence is ambiguous or conflicting.

6. Governance, Policy, and Regulatory Implications

Regulatory frameworks for herbal medicines have traditionally relied on morphological authentication, chromatographic fingerprints, and selected marker compounds. Artificial intelligence is not yet fully integrated into many pharmacopeial standards. The introduction of AI-discovered markers therefore raises questions about evidentiary expectations, validation requirements, and the role of computational evidence in regulatory decisions. Regulatory authorities may require that marker discovery be accompanied by detailed documentation of training data provenance, model architecture, validation procedures, and independent test

results. These expectations are consistent with the broader movement toward transparent and reproducible machine learning in regulated domains.

Policy development should also address data sharing and the governance of traditional knowledge. Herbal medicine data often originate from multiple countries and may include knowledge held by Indigenous and local communities. Data sharing arrangements must respect ownership, cultural significance, and benefit-sharing obligations. At the same time, overly restrictive data governance can impede the independent validation and cross-border collaboration that are necessary for building durable scientific consensus. A recurring policy challenge is that herbal authentication data are often fragmented across private quality control laboratories, academic metabolomics centers, and regulatory agencies. Standardized data-sharing agreements can support the construction of reference datasets that are sufficiently diverse to capture legitimate phytochemical variation while protecting commercially sensitive information. Federated learning architectures, in which models are trained across distributed data holdings without centralizing raw spectra, offer a promising governance mechanism. Such architectures, however, require careful technical coordination to ensure that local data preprocessing does not introduce divergent distributions and that model updates do not leak information about individual samples.

The evidentiary status of AI-derived markers also depends on their alignment with existing pharmacopeial frameworks. Many pharmacopeias define quality through specified marker compounds, chromatographic fingerprints, and physical or microscopic tests. AI-discovered metabolomic markers can complement these established methods, but their acceptance will require a clear demonstration of analytical validity, biological relevance, and robustness under the conditions described in regulatory monographs. Regulators may reasonably ask whether the markers reflect genuine chemical identity or merely laboratory-specific artifacts. To answer this, discovery workflows should include independent analytical methods, such as targeted tandem mass spectrometry or nuclear magnetic resonance confirmation, for a subset of proposed markers. The selection of a marker panel should also be accompanied by documentation of the model's decision boundaries and the expected uncertainty across sources of variation. Such documentation strengthens the legal and scientific credibility of AI-based authentication evidence.

National and regional policies differ in how they regulate herbal products, ranging from conventional drug frameworks to food supplement categories. These differences create complexities for AI-enabled authentication because the same marker panel may need to satisfy different evidentiary standards. A harmonized approach across international regulatory bodies would reduce duplication and enable the development of shared reference materials and spectral libraries. Initiatives that promote open exchange of spectral data, while respecting traditional knowledge and commercial confidentiality, can accelerate the transition from research-grade metabolomics to operational quality systems. The policy emphasis should be on transparent methods, auditable data provenance, and staged validation rather than on prescribing a single computational technique.

7. Sustainability, Deployment, and Future Perspectives

Sustainability considerations influence the design and long-term operation of AI-enabled marker discovery systems. Many metabolomics platforms in academic settings rely on short-term project funding and specialized expertise that may not be available in routine quality control laboratories. A marker panel that requires expensive high-resolution mass

spectrometry and complex data processing workflows may be technically valid but operationally unsustainable for decentralized supply chain monitoring. Sustainable deployment therefore calls for a tiered architecture in which initial comprehensive profiling is used to discover and validate markers, while routine monitoring uses more accessible targeted assays. Artificial intelligence can support this transition by identifying the smallest set of markers that retains discrimination while remaining measurable with lower-cost instruments.

The deployment of AI models across distributed supply chains raises additional issues of maintenance and update. Herbal materials are living products, and their chemical profiles may shift with climate variability, cultivation practices, and new adulteration strategies. A static model will degrade over time unless mechanisms exist for periodic performance review and recalibration. This suggests a lifecycle approach in which models are registered, versioned, and monitored much like analytical instruments. Post-market surveillance data can be used to refine models, but such updates must be governed by documented change management procedures to avoid introducing unanticipated biases or drift. The integration of model monitoring with laboratory information management systems can provide the audit trails needed for regulatory accountability.

There is also a growing opportunity to link AI-enabled metabolomics with broader digital infrastructure for botanical supply chains. Traceability systems, blockchain-based documentation, and mobile spectral acquisition tools can generate rich contextual data about provenance and handling. These data sources, when integrated with metabolomic profiles, allow authentication systems to reason not only about chemical identity but also about the conditions that produced a particular chemical state. Machine learning models that jointly learn from spectral features and supply chain metadata may improve robustness by conditioning predictions on known sources of variation. However, such integrations require careful attention to data quality, privacy, and the risk of reinforcing existing power asymmetries in international botanical trade.

Future research should prioritize open benchmark datasets that represent authentic herbal materials, common adulterants, and legitimate variants across multiple geographic origins and seasons. Benchmarks should include metadata sufficient to test model generalization under distribution shift. Without such shared evaluation resources, the field will continue to produce models with impressive internal validation that fail under realistic deployment conditions. Standardized reporting of negative results and model failure modes is equally important, because the absence of documented failures can create a false sense of security. Scientific journals and funding agencies can support this by incentivizing replication studies and the publication of validation datasets alongside methodological innovations.

Ethical and societal dimensions should not be overlooked. The introduction of AI-based authentication may alter the balance of power among producers, processors, regulators, and consumers. If deployed in ways that disproportionately burden small producers or that favor centralized control over traditional knowledge, the technology could undermine the very communities that maintain botanical diversity. Participatory design processes, benefit-sharing arrangements, and transparent governance can help ensure that AI-enabled quality systems serve public health and equity, not only commercial efficiency. The long-term contribution of AI to herbal medicine authentication will depend as much on these institutional arrangements as on algorithmic advances.

8. Conclusion

This paper has examined AI-enabled metabolite marker discovery for herbal medicine authentication and quality control from a system-level perspective. The central argument is that marker discovery should not be treated as a purely computational task, but rather as the design of a socio-technical infrastructure that connects analytical chemistry, machine learning, regulatory science, and supply chain governance. High-dimensional metabolomic data present opportunities for learning subtle chemical signatures, yet their value depends on the quality of data provenance, the interpretability of models, and the validation of markers under real-world variation. Architectural choices should optimize for operational feasibility and regulatory defensibility, not solely for predictive accuracy on retrospective datasets. The integration of AI-derived markers with complementary authentication methods, including cell wall component characteristics, strengthens the evidence base for robust quality decisions. Robustness, fairness, and risk management require attention to distribution shift, geographic representativeness, and the asymmetric consequences of authentication errors. Policy frameworks must evolve to address data governance, traditional knowledge, transparency, and the evidentiary requirements for computational marker evidence. Sustainable deployment calls for tiered analytical workflows, lifecycle model management, and participatory governance. Looking ahead, the field would benefit from open benchmark datasets, shared reference resources, and harmonized regulatory expectations. Under these conditions, AI-enabled metabolomics can make a meaningful and durable contribution to the authentication and quality control of herbal medicines.

References

1. Fiehn, O. (2002). Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171.
2. Gika, H. G., Theodoridis, G. A., Plumb, R. S., & Wilson, I. D. (2014). Current practice of liquid chromatography–mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and Biomedical Analysis*, 87, 12–25.
3. da Silva, R. R., Dorrestein, P. C., & Quinn, R. A. (2015). Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences*, 112(41), 12549–12550.
4. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, 44(D1), D1214–D1219.
5. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., ... Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3), 211–221.
6. Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D. B., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J., & Wulfert, F. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, 3(3), 231–241.
7. Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.-É., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*, 49(W1), W388–W396.

8. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
9. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56.
10. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
11. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates.
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
14. World Health Organization. (2007). WHO guidelines for assessing quality of herbal medicines with reference to contaminants and residues. World Health Organization.
15. European Medicines Agency. (2011). Guideline on quality of herbal medicinal products/traditional herbal medicinal products (EMA/HMPC/201116/2005 Rev. 2). European Medicines Agency.
16. Food and Drug Administration. (2016). Botanical drug development: Guidance for industry. U.S. Department of Health and Human Services.
17. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42.
18. 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.
19. Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 15(7), 473–484.
20. Nicholson, J. K., & Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature*, 455(7216), 1054–1056.
21. Li, S., Han, Q., Qiao, C., Song, J., Cheng, C. L., & Xu, H. (2008). Chemical markers for the quality control of herbal medicines: an overview. *Chinese Medicine*, 3, 7.