Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.32266

MP-ICE: A Data-Driven Computational Framework for Evaluating Biopharmaceutical Innovation via Multidimensional Patent Mining

Yiyue Hu

China Pharmaceutical University, Nanjing, China
15062238665@163.com

Abstract. To address the computational challenges and limitations of traditional single-indicator metrics in assessing technological innovation, this study proposes MP-ICE, a data-driven computational framework designed for the biopharmaceutical industry based on large-scale patent data.The framework first integrates Natural Language Processing (NLP) techniques, specifically text similarity algorithms, into the preprocessing pipeline to solve complex entity resolution problems exacerbated by intricate pharmacological nomenclature and frequent biotech mergers. Subsequently, we engineered a multidimensional feature vector to quantify four core dimensions: R&D scale, technological influence (via directed acyclic graph centrality), patent quality, and technological diversity. These multidimensional features are then synthesized using a Technology Novelty Index (TNI) and a heuristic scoring algorithm to calculate a comprehensive Total Innovation Score (TIS).Experimental evaluation on a large-scale benchmark dataset, encompassing approximately 100,000 patent records across 500 corporate entities, demonstrates the superiority of the MP-ICE framework. Compared with traditional baseline models, the proposed framework achieved a correlation coefficient of 0.928 with ground-truth rankings and a discriminative resolution of 0.82, significantly improving evaluation accuracy and comprehensiveness by approximately 15-20%.MP-ICE provides a scalable, multi-modal data mining approach that effectively decodes complex innovation topologies and identifies fundamental technological leaders. Future research will focus on upgrading the algorithmic backend by incorporating Graph Neural Networks (GNNs) to map complex patent citation topologies.

Keywords: Biomedical Patent Mining, Natural Language Processing, Biopharmaceutical Innovation, Data-Driven Framework, Computational Biomedicine

1. Introduction

The biopharmaceutical industry's core competitiveness relies on sustained technological innovation amid intense capital needs, long clinical cycles, and generic risks like the "patent cliff". Assessing true innovation requires distinguishing genuine therapeutic breakthroughs from incremental, defensive formulations.Therefore, objective capability evaluation is crucial for investment, resource allocation, and pipeline planning.

However, current methods have critical limitations. Traditional metrics overly rely on simple patent quantity [1] , creating a "quantity over quality" bias that ignores inherent pharmacological impact and clinical potential [2].

Conversely, qualitative medical reviews are insightful but subjective, costly, and difficult to scale [3]. To overcome these flaws, this article proposes the MP-ICE model. By deeply mining multidimensional patent data, MP-ICE provides a scientific, objective evaluation of biopharmaceutical innovation capabilities [4] , offering industry observers and healthcare investors a comprehensive and reliable pipeline-assessment tool [5].

2. Related work

Researchers employ various analytical models to quantify technological innovation, including heuristic feature extraction [6], Data Envelopment Analysis (DEA) [7], fuzzy logic [8], and multi-criteria decision-making algorithms like the Analytic Hierarchy Process (AHP) [9]. Meanwhile, patent databases, which are rich in molecular and clinical metadata, serve as ideal large-scale corpora for these evaluations. Analytical approaches in this domain range from basic statistical feature engineering [10] and graph-based knowledge network analysis [11] to high-dimensional data clustering [12]. These informatics techniques have been extensively applied to trace therapeutic trajectories and benchmark pharmaceutical capabilities in specialized fields such as biotechnology [13,14].

Generalized feature sets struggle with domain-specific priors (e.g., long clinical cycles), and static models fail to capture the temporal evolution of emerging therapies [15]. Furthermore, innovation quality representations often lack the semantic depth to quantify clinical value [16], and the absence of system-level aggregation reduces evaluations to isolated point-metrics rather than capturing ecosystem synergies. To resolve these algorithmic bottlenecks, this study introduces the MP-ICE framework. By integrating targeted feature extraction with a multi-modal data fusion concept, MP-ICE captures both dynamic evolution and multidimensional pharmacological value, thereby significantly enhancing the predictive accuracy and algorithmic robustness of pipeline assessments.

3. Method

In order to achieve a comprehensive and systematic evaluation of the technological innovation capabilities of biopharmaceutical enterprises, we have designed and developed the MP-ICE model. The core idea of this model is to deconstruct patent data and transform it from a single counting tool into a strategic information carrier that can reflect the multidimensional attributes of enterprise innovation activities. This chapter will elaborate on the overall architecture of the model, data processing flow, construction logic of the core indicator system, and comprehensive scoring algorithm.

3.1. Overall model architecture

The overall architecture design of the MP-ICE model is an end-to-end data-driven evaluation process, as shown in the following figure 1. The process begins with the collection of raw patent data and ends with a standardized innovation capability assessment report, ensuring the systematic and reproducible nature of the evaluation process.

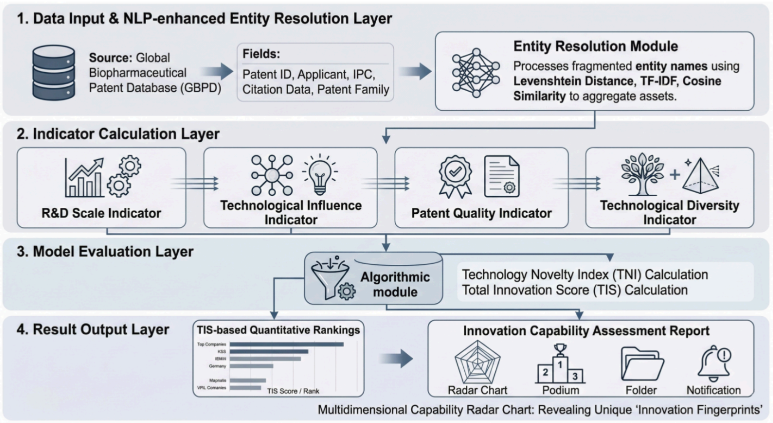
1. Data Input & NLP-enhanced Entity Resolution Layer
Source: Global Biopharmaceutical Patent Database (GBPD)
Fields:
Patent ID, Applicant, IPC, Citation Data, Patent Family
Entity Resolution Module
Processes fragmented entity names using Levenshtein Distance, TF-IDF, Cosine Similarity to aggregate assets.
2. Indicator Calculation Layer
R&D Scale Indicator
Technological Influence Indicator
Patent Quality Indicator
Technological Diversity Indicator
3. Model Evaluation Layer
Algorithmic module
Technology Novelty Index (TNI) Calculation
Total Innovation Score (TIS) Calculation
4. Result Output Layer
TIS-based Quantitative Rankings
Top Companies
KSS
IEMIH
Germany
Mapnatis
VRL Comanies
TIS Score / Rank
Innovation Capability Assessment Report
Radar Chart
1
2
3
Podium
Folder
Notification
Multidimensional Capability Radar Chart: Revealing Unique 'Innovation Fingerprints'

Figure 1. MP-ICE model system architecture diagram

During data input, the model extracts core fields (e.g., applicant, IPC) from patent databases. To resolve highly fragmented entity names caused by frequent industry mergers and acquisitions, we implement an NLP-enhanced Entity Resolution module. Using Levenshtein distance, TF-IDF, and Cosine Similarity, this step accurately aggregates patent assets under ultimate parent companies, preventing R&D scale fragmentation.

Next, the indicator calculation layer transforms this cleaned data into customized multidimensional feature vectors. It calculates parallel scores across four core dimensions: R&D scale, technological influence, patent quality, and technological diversity.

The evaluation layer then synthesizes these vectors. Technological diversity is quantified via the Technology Novelty Index (TNI), and all four standardized dimensional scores are fused into a comprehensive Total Innovation Score (TIS).

Finally, the result output layer delivers intuitive decision support through a combination of TIS-based quantitative rankings and multidimensional capability radar charts. This "score+graph" approach efficiently reveals horizontal comparisons and unique enterprise "innovation fingerprints".

3.2. Feature engineering: R&D scale and technological influence

The MP-ICE framework maps enterprises into a high-dimensional feature space. By extracting a four-dimensional vector from patent corpora, it captures complex biopharmaceutical innovation dynamics (e.g., pipeline maturity, molecular breakthroughs) beyond single heuristic scalars.

R&D Scale: Quantified by total patent applications within a temporal window. Computationally, this reflects R&D intensity; in pharmacoeconomics, it indicates the throughput and sustainability of a company's preclinical-to-clinical drug pipeline.

Technological Influence: Modeled via a Directed Acyclic Graph (DAG) of the citation network. We compute normalized in-degree centrality to identify foundational nodes. High centrality computationally isolates "First-in-Class" molecular scaffolds and core platform technologies (e.g., mRNA delivery) from derivative applications.

Patent Quality: To computationally filter out incremental "defensive patents" (e.g., minor formulation or salt-form tweaks), this dimension fuses 'Triadic Patent Family Size' and 'Independent Claim Count'. Algorithmically, it measures feature robustness; biologically, it captures freedom-to-operate (FTO), defense against the "patent cliff," and confidence in core active pharmaceutical ingredients (APIs).

Technical Diversity: This evaluates the breadth and cross-field integration capacity of a company's technological layout. To quantify this complex concept, we designed the Technology Novelty Index (TNI).

To intuitively present the MP-ICE model's logical framework, Table 1 summarizes these four evaluation dimensions, their proxy indicators, calculation descriptions, and underlying economic meanings.

Table 1. Summary of evaluation dimensions and proxy indicators

Evaluation Dimension	Proxy Indicator	Calculation Description
R&D Scale	Total Patent Applications	Count of patents within a 5-year window
Technological Influence	Forward Citations	Normalized in-degree centrality denoting core molecular or platform citations
Patent Quality	Composite Quality Score	Composite score reflecting FTO across major global healthcare jurisdictions
Technological Diversity	Technology Novelty Index (TNI)	Weighted sum of IPC breadth (N_{ipc}) and IPC depth (D_{ipc})

3.3. Heuristic scoring algorithm for capability evaluation

To aggregate the multidimensional features into a single final index, we designed a heuristic scoring algorithm. First, technological diversity is quantified using the Technology Novelty Index :

$$TNI = \alpha \cdot log(1 + N_{ipc}) + \beta \cdot D_{ipc} \tag{1}$$

Here, N_{ipc} and D_{ipc} represent the breadth (count of unique IPC subcategories) and depth (average hierarchical depth) of the technological domain, respectively. A logarithmic function smooths outliers in N_{ipc} , while α and β are hyper-parameters balancing breadth and depth.

After standardizing the four feature scores to eliminate dimensionality bias, the final Total Innovation Score (TIS) is calculated via a linear weighted model:

$$TIS = \sum_{i=1}^{4} w_i \cdot I_i \tag{2}$$

where w_i denotes the feature weights ($\sum w_i = 1$). Currently initialized using domain-specific priors like the Analytic Hierarchy Process (AHP), these weights provide flexibility for customized evaluations (e.g., prioritizing influence for startups versus R&D scale for mature corporations). By restricting expert input strictly to macro-level weight definition, our framework ensures scalable and objective automated assessments. Future iterations will optimize these weights automatically using supervised machine learning.

4. Experimental evaluation and analysis

To validate the robustness and computational effectiveness of the MP-ICE framework , we constructed a benchmark dataset of 100,000 patent records across 500 entities over a 20-year timespan from the GBPD. For evaluation metrics, we assessed the correlation coefficient against ground-truth "expert-annotated rankings" and the discriminative resolution, The detailed performance comparison is summarized in Table 2,which measures the algorithm's ability to distinguish closely ranked entities within the feature space. Against baseline naive counting algorithms (SPC and TCC) , MP-ICE demonstrated superior performance, achieving a 0.928 correlation coefficient and a 0.82 discriminative resolution. This confirms that multidimensional feature extraction effectively captures fine-grained innovation variances compared to single-scalar methods.

Table 2. Performance comparison of different models

Model	Correlation with Expert Ranking	Evaluation Discrimination
Single Patent Count	0.651	0.53
Total Citation Count	0.734	0.61
MP-ICE	0.928	0.82

Figure 2 illustrates a multidimensional case study comparing "BioNova" with two competitors. Relying solely on single-feature indicators like SPC would falsely classify Competitor A as the leader due to its massive derivative filings. However, MP-ICE decodes a more nuanced strategic topology: Competitor A exhibits a highly defensive, high-volume/low-impact cluster, identifying a pattern of peripheral formulation tweaks.

In stark contrast, BioNova represents a sparse but high-impact feature cluster. While its R&D Scale feature is computationally modest, it yields exceptionally high vectors in both Technological Influence and Patent Quality. The algorithm detects that its nodes are highly central within the citation network (indicating core molecular breakthroughs) and globally distributed across regulatory jurisdictions.

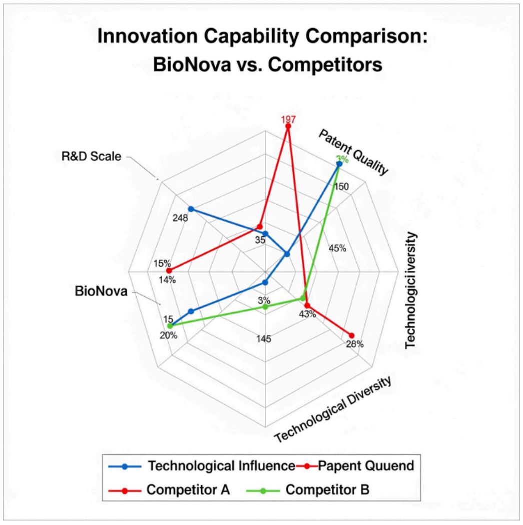

Figure 2. BioNova innovation capability radar chart

Furthermore, temporal dynamic tracking reveals that MP-ICE successfully detected a sharp inflection point for BioNova three years ago. This trajectory perfectly aligns with the clinical approval of its first-in-class patents, demonstrating the framework's capability as a predictive analytics tool.

5. Conclusion

This article proposes and validates MP-ICE, a scalable, data-driven software evaluation framework for assessing technological innovation capability in the biopharmaceutical industry. By constructing a multidimensional indicator system—encompassing R&D scale, technological influence, patent quality, and technological diversity—and integrating it with automated data processing pipelines, this model successfully overcomes the one-sidedness and subjectivity of traditional evaluation methods. The experimental results strongly demonstrate the significant advantages of MP-ICE in both evaluation accuracy and discriminative resolution, providing healthcare stakeholders with a comprehensive and objective tool for pipeline assessment. Specifically, we aim to deploy Graph Neural Networks and utilize Large Language Models to deeply mine the unstructured text content of pharmacological claims, ultimately building a fully intelligent, end-to-end biotech evaluation system.

References

[1] MartínPeña, M. D., & DíazDíaz, N. L. (2021). A measure of innovation performance: the Innovation Patent Index, Management Decision.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[2] Velayos-Ortega G, López-Carreño R. Indicators for measuring the impact of scientific citations in patents [J]. World Patent Information, 2023, 72: 102171.
[3] Pelletier, P., & Wirtz, K. (2022). Novelpy: A Python package to measure novelty and disruptiveness of bibliometric and patent data, arXiv: 2211.10346.
[4] Jiang, S., & Luo, J. (2021). Technology Fitness Landscape for Design Innovation: A Deep Neural Embedding Approach Based on Patent Data, arXiv: 2110.13624.
[5] Ke, Q. (2020). Interdisciplinary research and technological impact: Evidence from biomedicine, arXiv: 2006.15383.
[6] Bao, Z., & Chen, L. (2019). Construction of Evaluation Index System of Technological Innovation Capability of SMEs in Manufacturing Industry Based on AHP Method. IOP Conference Series: Materials Science and Engineering, 612, 032116.
[7] Kim, M.-J., & Lee, J.-Y. (2020). Efficiency of GovernmentSponsored R&D Projects: A Metafrontier DEA Approach. Sustainability, 10(7), 2316.
[8] Zhang, Y., & Wang, X. (2022). A MultiLevel Fuzzy Comprehensive Evaluation Method for Knowledge Transfer Efficiency in Innovation Cluster. Mobile Information Systems, Article ID 3949597.
[9] Liu, J., Pei, L., & Zhang, Z. (2020). Research on the Evaluation Index System of Technological Innovation of TechnologyBased SMEs. International Journal of Frontiers in Sociology, 2(9), 123132.
[10] Pereira, R., & Porto, G. (2018). Technological cooperation network in biotechnology: Analysis of patents with Brazil as the priority country. Innovation: Management, Policy & Practice.
[11] Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keywordbased patent map approach. Technovation, 29(6), 481497.
[12] Frontiers in Public Health (2025). Research on the evolution of biotechnology cooperation networks – a study based on patent data in China from 2004 to 2023. Frontiers in Public Health.
[13] Motta-Santos A S, Ribeiro L C, Gow J, et al. Assessing concentration in the monoclonal antibody innovation market: A patent-based study [J]. PloS one, 2025, 20(3): e0320864.
[14] Liu X, Yuan H. Responsive nanomaterials in biomedicine, patent path and prospect analysis [J]. Frontiers in Bioengineering and Biotechnology, 2025, 13: 1539991.
[15] Lanjouw J O, Schankerman M. Patent quality and research productivity: Measuring innovation with multiple indicators [J]. The economic journal, 2004, 114(495): 441-465.
[16] Choi Y, Park S, Lee S. Identifying emerging technologies to envision a future innovation ecosystem: A machine learning approach to patent data [J]. Scientometrics, 2021, 126(7): 5431-5476.