# Development of an Agentic AI for Protein Candidate Screening in Superbug Vaccine Development

Jaemin Lee [a, c], Yonggyun Yu [a], Min-Kyu Kim [b], Ho Seong Seo [b], and Yujong Kim [a]*
[a]*Applied Artificial Intelligence Section, KAERI, Daejeon 34057, Republic of Korea*
[b]*Cyclotron Application Research Section, KAERI, Jeongeup 56212, Republic of Korea*
[c]*Hanyang University ERICA, Ansan 15588, Republic of Korea*
**Corresponding author: yjkim@kaeri.re.kr*

***Keywords :*** Agentic AI, AI Scientist, Vaccine candidate, BioBERT, Model Context Protocol (MCP)

## 1. Introduction

Antimicrobial resistance (AMR) occurs when bacteria evolve to resist antibiotics that once worked effectively. Among them, carbapenem-resistant Enterobacteriaceae (CRE) has become one of the most alarming "superbugs," causing infections that are extremely difficult to treat. The World Health Organization (WHO [1]) has listed AMR as one of the "Top 10 Global Public Health Threats," and projections suggest millions of deaths and enormous economic costs if new solutions are not found. Vaccines offer one of the most promising strategies to combat AMR, but identifying effective vaccine candidates from the rapidly growing scientific literature remains a formidable challenge.

Traditionally, researchers have relied on manual reviews of papers or simple keyword searches to find potential vaccine targets. While useful, these methods are time-consuming, error-prone, and often miss important findings hidden in the text. Recent advances in artificial intelligence (AI) have introduced powerful language models that can read and understand scientific text. Yet most existing approaches are limited to document retrieval and do not provide structured datasets that connect key elements such as vaccine type, animal model, immune response, and protection rate. This leaves a major gap in efforts to accelerate vaccine discovery.

To address this challenge, we propose an Agentic AI framework. Unlike conventional systems that only search or summarize, Agentic AI actively plans and executes multi-step research tasks—much like a junior scientist who can scan papers, extract key results, and organize them into usable data. In this work, the system integrates BioBERT [2], a biomedical language model, with rule-based reasoning to automatically detect vaccine candidates across diverse publications. It is deployed in the Model Context Protocol (MCP) [3] environment, which allows natural language queries to be translated into structured datasets.

The resulting pipeline substantially reduces manual workload while improving the accuracy of immune marker and animal model detection. Although this study focuses on vaccine candidate screening for superbugs, the same approach can be extended to other fields such as nuclear safety reports, materials science, and environmental monitoring. By automating the transformation of unstructured scientific text into structured insights,

this work demonstrates how Agentic AI can enhance research efficiency and reproducibility across domains.

## 2. Methods and Results

Our pipeline comprises data collection, section-aware preprocessing, candidate detection, CSV export, and validation, as shown in Fig. 1.

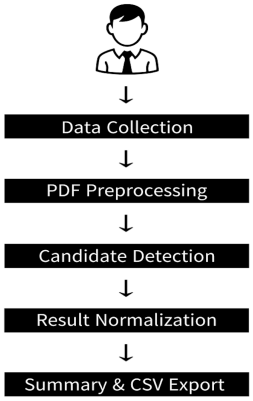


Fig. 1. Overall workflow of the proposed Agentic AI pipeline, consisting of five stages: data collection, PDF preprocessing, candidate detection, result normalization, and CSV export.

### *2.1. Data Collection and Normalization*

A total of 112 vaccine- and immunity-related papers were collected from PubMed Central, arXiv, and medRxiv. Identifiers (PMID/DOI) were normalized to remove duplicates, and the results were stored in a standardized CSV schema. As summarized in Table I, the schema includes vaccine name, animal model, protection rate, immune observations, and metadata. This structure makes the dataset easy to reuse and compare across studies, supporting reproducibility and further analysis.

### *2.2. PDF Text Preprocessing*

PDFs were converted to text using PyMuPDF [4], and the core sections—Abstract, Introduction, Results, and Conclusion—were extracted. The Results section was filtered using protection-related keywords such as

protection, efficacy, survival, neutralizing, and mortality to prioritize experimental data. To reduce noise, a maximum length cap defined by the parameter *results_budget_chars* was set to 1500.

Table I: Standardized CSV Schema for Extracted Vaccine Data

| Column Name | Description |
|---|---|
| *vaccine_name* | Standardized vaccine name |
| *animal_model* | Standardized animal model name |
| *protection_rate* | Protection rate (%) |
| *protective_immunity_observed* | Representative observation sentence |
| *corresponding_author_country* | Country of the corresponding author |
| *author_count* | Number of authors |
| *pmid* | Identifier (PMID/DOI, etc.) |

### *2.3. Candidate Detection*

Candidate detection combined BioBERT embeddings with keyword matching to identify vaccine names, animal models, protection rates, and immune responses. The process included similarity-based sentence detection with a threshold of $\tau = 0.68$, rule-based completion for missing values, and normalization of animal models and vaccine names.

### *2.4. CSV Extraction Module*

A CSV extraction module was developed to automatically store results in a standardized schema, enabling reproducible datasets. The module runs in the MCP environment and can be seamlessly integrated with other MCP-based servers. In this schema, *protection_rate* denotes the final protection value.

### *2.5. Filtering and Validation*

Irrelevant or incomplete cases were excluded to ensure dataset quality. Papers unrelated to vaccine protection were removed. If explicit efficacy values were absent, rule-based contextual estimates were used as a supplement, while negative results such as *not protected* were excluded. This approach ensured that only reliable and consistent records were retained for analysis.

For example, statements such as '70% survival after viral challenge' were normalized to a 70% protection rate entry, ensuring consistency across studies.

### *2.6. Implementation Details*

The pipeline was implemented in Python, using PyMuPDF for PDF text extraction and BioBERT sentence embeddings with cosine similarity provided by PyTorch.

### *2.7. Query-Based Dataset Retrieval*

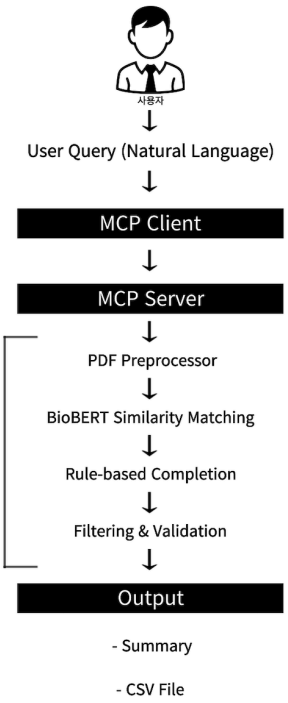


Fig. 2. Internal MCP workflow for query-based dataset Retrieval.

Figure 2 illustrates the internal workflow of the MCP module for query-based dataset retrieval. A user's natural language query is first converted into structured conditions such as vaccine name, animal model, or protection rate. In this study, the MCP tools (*search_and_extract_local_vaccine_info* and *export_vaccine_info_to_csv*) handle the structured input directly, while the conversion from natural language to structured conditions is assumed to be performed by an upstream agent module. The system then performs PDF preprocessing, BioBERT-based similarity matching, rule-based completion, and filtering, finally producing standardized summaries and CSV files.

This workflow substantially reduces manual review effort by extracting only the papers relevant to user-specified criteria.

### *2.8. Candidate Detection Performance*

In the subset of mRNA vaccines tested in mouse models, two high-confidence vaccine candidates were identified with protective efficacy equal to or exceeding 70%, as summarized in Table II. The top-performing candidate, All polyvalent MPXV mRNA, achieved a 90% protection rate and induced robust immune responses to both extracellular enveloped virus, abbreviated as EEV, and intracellular mature virus, abbreviated as IMV. The second candidate, SPIKE

mRNA-1273, demonstrated a 70% protection rate, accompanied by robust immune signatures including strong neutralizing antibodies and multiple T-cell responses, while also providing protective effects in an acute lung injury model.

These results show that the pipeline can recover both well-known candidates like SPIKE mRNA-1273 and newer designs such as polyvalent MPXV vaccines, pointing to its ability to identify promising multi-antigen strategies with strong protection in animal studies.

### *2.9. Cross-Domain Applicability*

While optimized for vaccine candidate detection, the pipeline can also be applied to extract key performance indicators from nuclear safety reports, such as reactor output trends, coolant flow rates, and recorded safety events, enabling integration with real-time monitoring dashboards. Beyond nuclear applications, the framework is adaptable to other domains, including materials science and environmental monitoring datasets.

### *2.10. Limitations and Future Work*

This pipeline has several limitations: (I) the dataset size is restricted to 112 papers, (II) it supports only English literature, and (III) it is limited to detecting locally stored PDFs, which restricts scalability and generalizability.

Future work will extend the framework to multilingual and structured data sources, with particular focus on nuclear applications.

Additionally, performance metrics such as precision and recall have not yet been benchmarked; a formal evaluation is planned as future work to validate the expected advantage of BioBERT-based semantic matching over keyword-only baselines.

## 3. Conclusions

This study presents an Agentic AI–based pipeline that automates the retrieval and normalization of protection-related information from vaccine and immunity literature. The approach enhances efficiency and reproducibility in selecting candidate proteins for superbug vaccine development by combining semantic analysis with rule-based filtering.

Validated through vaccine candidate detection, the methodology demonstrates improved accuracy in immune marker detection and reliability in animal model identification, while reducing manual review effort. Beyond this case study, the pipeline is adaptable to diverse scientific and engineering domains, supporting broader research productivity.

This work shows how Agentic AI can turn unstructured scientific text into structured, actionable insights, supporting faster and more reliable discovery across domains.

## REFERENCES

[1] World Health Organization, “Antimicrobial resistance,” Fact sheet, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance
[2] P. Deka, BioBERT-mnli-snli-scinli-scitail-mednli-stsb, Hugging Face, 2021. URL: https://huggingface.co/pritamdeka/BioBERT-mnli-snli-scinli-scitail-mednli-stsb
[3] Model Context Protocol (MCP) Documentation, URL: https://modelcontextprotocol.io (accessed: Aug. 2025)
[4] PyMuPDF Documentation,
URL: https://pymupdf.readthedocs.io

Table II: Summary of high-protection mRNA vaccines tested in mouse models.

| Vaccine Name | Animal Model | Protection Rate (%) | Protective Immunity Observed | Corresponding Author Country | Author Count | PMID |
|---|---|---|---|---|---|---|
| All polyvalent MPXV mRNA | mouse | 90 | Multiple MPXV antigens are expressed in one mRNA molecule by a 2A peptide; All polyvalent MPXV mRNA vaccines induce robust immune responses to EEV and IMV antigens The tetravalent MPXV mRNA vaccin.... | China | 1 | 10.1016/j.celrep.2024.114269 |
| SPIKE mRNA-1273 | mouse | 70 | Elicited neutralizing antibodies, spike-binding germinal center B cells, and spike- specific CD4+ Th1/ Tfh and CD8+ T cells; demonstrated protection in acute lung injury model. | USA | 2 | 10.1016/j.immuni.2021.06.018 |

*EEV: Extracellular enveloped virus; IMV: Intracellular mature virus; HA: Hemagglutinin; Th1: T helper type 1; Tfh: Follicular helper T cell*