Work Package	**WP4 - GenDAI Discovery - Biomarker prospection in the microbiome using AI**
Task(s)	**T4.1 Microbiome Feature selector using Genetic Algorithm and ML** **T4.2 NLP for Metagenomics Analysis**
Deliverable	**D4.1 GenDAI Discovery Classification, Configuration & Methodology**: This initial delivery will focus on providing a simple classification model based on the improved BiGaMi algorithm to support early piloting and user feedback. It includes data and methodologies used in implementing the AI models as well configuration and technical requirements for model deployment.
Author(s)	Haithem Afli, Bruno Andrade, Mike Leske, Thomas Krause, Philippe Tamla
Version	1.0
Dissemination Level	PU

Table of Contents

Executive summary

This deliverable establishes the core technical methodology for the GenDAI Discovery component by validating a shift from traditional alignment-based metagenomic classification to a novel, alignment-free approach using Genomic Language Models (gLMs). The key results are: 1) the achievement of 98.0% classification accuracy and a ~2800x speed increase over conventional tools like BLAST+, 2) the provision of a detailed operational blueprint for model deployment and integration, and 3) the design of a secure, privacy-preserving Training UI based on split-fed learning. The value added by these results are: a) they provide the foundational, high-performance AI engine required for the high-throughput, automated diagnostic pipeline in WP2, and b) the validated methodology and open-access publication position this work as a state-of-the-art contribution to the field of computational metagenomics, with clear potential for exploitation in other scientific and clinical diagnostics platforms.

Outputs and Project Progress

This deliverable directly contributes to the achievement of the project's Research and Innovation Objectives (RIOs) and associated milestones as stated in the Description of Action (DoA). It provides the validated classification methodology and configuration specifications that are essential inputs for downstream tasks. Specifically:

- It supports O4.1 (Develop an improved feature selection model) by establishing the embedding-based AI framework for biomarker prospection.
- It delivers on O4.2 (Develop NLP for Microbiome Data Analysis) through the development and validation of the alignment-free 16S rRNA classification method using gLMs, a milestone formally published as a preprint.
- It enables O4.3 (Provision of the GenDAI Discovery Training UI) by specifying the technical requirements, including the split-fed learning architecture, for the UI's development in T4.4.

The operational blueprint ensures compliance with key DoA requirements for scalability, explainability (XAI), and clinical integration, feeding directly into the WP2 Diagnostics Workflow (D2.1/D2.2) and preparing the system for WP6 piloting (T6.2).

Contributions to Impacts

The results presented herein create the foundational conditions for future impacts as stated in the DoA, primarily by enabling a fast, accurate, and explainable AI-driven diagnostics platform that accelerates the translation of metagenomic data into actionable clinical insights for personalized medicine, starting with conditions like IBS.

1. Introduction

The foundational mission of the GenDAI project is the creation of a sophisticated medical diagnostics platform that leverages AI and metagenomic potential to provide powerful diagnostic results. This initiative is designed to accelerate the conversion of innovative ideas into breakthroughs in medical analysis services, moving the field towards Personalized Medicine and Work Package 4 (WP4), designated as 'GenDAI Discovery,' is central to this mission.

This Work package's defined research and innovation objectives (RIOs) include O4.1 (Develop an improved feature selection model to identify candidate biomarkers), O4.2 (Develop Natural Language Processing for Microbiome Data Analysis), and O4.3 (Provision of the GenDAI Discovery Training UI). The ultimate responsibility of WP4 is the development and improvement of AI methods capable of identifying relevant biomarkers and classifying corresponding metagenomic sequences to characterize a patient's microbiome profile and state of health. The output of this methodological phase is documented across two core deliverables: D4.1 (Classification, Configuration & Methodology, M15) and D4.2 (Improved Classification and Training UI, M30).

1.1 Purpose of this Document

D4.1 serves as the foundational technical report, formalizing the AI methodology that underpins biomarker prospection within the GenDAI platform. The primary purpose of this document is to ensure that the strategic technical choices are scientifically validated and operationally prepared for deployment. The results documented herein establish the necessary technical blueprint, which acts as direct and essential input for subsequent development, particularly the GenDAI Diagnostics Workflow (WP2) and the data management services (WP3).

The scope of this deliverable is exhaustive, covering the scientific justification for the strategic methodology, detailed performance benchmarks that validate the new alignment-free methodology, and the detailed configuration requirements essential for the operational Tasks T4.3 (Model Configuration) and T4.4 (Training UI).

1.2 Document Organization

This document is structured to logically present the scientific justification, technical methodology, and operational plan for the GenDAI Discovery component. It begins with an introduction to the project and the purpose of this deliverable. Section 2 provides a state-of-the-art analysis, outlining the limitations of traditional metagenomic methods and how the chosen gLM approach addresses core project requirements.

The technical core is detailed in Section 3, which explains the conceptual model of genomic embeddings, the operational optimization achieved through dimensionality reduction, and comprehensive performance validation against established benchmarks. Section 4 discusses advanced methodological needs for enhancing model robustness.

The document then transitions to implementation, with Section 5 providing a deployment blueprint and specifying configuration requirements for the subsequent operational tasks, T4.3 and T4.4. Section 6 addresses critical aspects of clinical integration, including workflow compatibility, explainability, and bias mitigation. Finally, Section 7 summarizes the conclusions and outlines the immediate next steps for the work package, ensuring a clear path forward for development and piloting.

1.3 Methodology

The technical scope of WP4 encompasses AI approaches for metagenomic analysis, particularly focusing on robust **feature selection** methodologies for biomarker identification. To ensure the project's methodology remains at the cutting edge, WP4 is actively following the rapid, global scientific progression in genomic research and incorporating the domain shift towards Genomic Large Language Models (LLMs), also known as Genomic Foundation Models (GFMs), as part of a pipeline for sequence classification, biomarker discovery and metagenomic analysis. This forward-looking approach ensures the project's output aligns with state-of-the-art computational genomics, directly addressing the project's Research and Innovation Objectives (RIOs).

Genomic LLMs, such as DNABERT-2 [1] and Evo [2], apply the highly scalable Transformer architecture—originally used in NLP—directly to DNA sequences. This allows these models to effectively capture complex contextual relationships between nucleotides that extend far beyond the capabilities of traditional feature engineering or alignment methods [3]. The enhanced performance in both accuracy and computational efficiency provided by gLMs is the core technical requirement necessary to implement the high-throughput, comprehensive diagnostic platform envisioned in the project’s RIO, ensuring competitive advantage and enabling rapid clinical deployment.

2. Scientific State of the Art and Requirements Mapping

The wide adoption of metagenomic diagnostics in clinical practice is hampered by inherent technical barriers. Foremost among these is the challenge of **high-dimensionality**. Metagenomic datasets are uniquely complex, featuring hundreds of thousands of sequences, millions of genes, and often sparse patient metadata, creating data sets where the number of features significantly outstrips the number of samples. Traditional statistical and machine learning methods are typically insufficient to process this volume without either losing critical information or becoming computationally prohibitive.

A second, persistent constraint is the reliance on **alignment-based methods** such as BLAST+ and VSEARCH. These tools utilize complex sequence alignment to determine taxonomic identity, a process that demands extensive computational resources, large memory footprints, and prolonged processing times. In addition, these conventional methods often fail to classify **novel or uncharacterized microorganisms** not yet cataloged in reference databases. Since these overlooked microbial components may hold significant diagnostic potential, traditional pipelines inherently restrict the discovery of novel biomarkers. The necessary response to these limitations dictated the adoption of computationally efficient, **alignment-free solutions** for metagenomic classification.

2.1 Requirements Integration: Addressing Explainability and, Robustness and High dimensional data

The use of gLMs was a way to align the platform with several key non-functional requirements defined in D1.1.

Regarding the **Requirement for Handling High-Dimensional Data** (D1.1, Section 2.5.2.A), the method described by our group included an improved feature selection process (T4.1) designed to integrate diverse data types, such as continuous clinical measurements and categorical co-variables (e.g., ethnicity and sex). This fusion is essential to identify microbiome components that are biologically relevant for a specific set of metadata.

The need for **Explainability (XAI)** in clinical settings (D1.1, Section 2.5.2.C) is structurally supported by the gLM methodology. Clinicians require AI outputs that not only classify taxa but also **directly suggest biomarkers** along with corresponding confidence scores. The classification relies on Cosine Similarity and the Transformer's attention mechanism, providing an accessible foundation for post-hoc analysis for methods, such as the one described by our group [4]. This architectural suitability facilitates the seamless integration of XAI techniques, such as SHAP or LIME, which can highlight feature importance and quantify the rationale behind the AI's prediction. This capability ensures transparency, which is vital for building clinical trust and adoption.

The **Requirement for Clinical Integration and Fine-tuning** (D1.1, Section 2.5.2.F) mandates that the system allow domain experts to refine models for local use cases (e.g., regional patient cohorts) or updated sample collections. This necessity is directly addressed by confirming the design of the subsequent T4.4 (Training UI) around a secure **split-fed learning framework**, which enables localized model updates without compromising patient data privacy.

2.2 Discussion

Genomic Language Models utilize Transformer architectures to view genomic sequences as a form of language. These models learn the contextual relationships between nucleotides, converting the raw genetic data into expressive fixed-length numerical representations known as embeddings [5].

Comparative evaluation focused on established models, including DNABERT-2 [1], DNABERT-S [6], Evo [2], and NT2 [7]. The analysis confirmed that **DNABERT-2 and DNABERT-S** were the most effective choices for 16S classification tasks, demonstrating a superior capability to cluster and distinguish phylogenetic groups in the embedding space. This rigorous scientific validation underpins the final choice of model architecture. The core deliverable of Task T4.2 culminated in the development and validation of the alignment-free bacterial classification method for 16S rRNA sequences, a milestone formally published via a bioRxiv preprint.

3. Core Classification Methodology: Alignment-Free Genomic Embeddings

The foundation of the GenDAI Discovery methodology is the transformation of sequence analysis into efficient vector operations. The sequence encoding process converts raw DNA sequences (including full 16S genes and V3-V4 regions) into fixed-length numerical vectors, or embeddings, created by different gLMs, such as DNABERT-2.

3.1 Conceptual Model: Genomic Sequences as Language and Representation Learning

Classification relies on storing these precomputed embeddings in a high-speed vector store (e.g., FAISS). When a new sequence requires classification, its embedding is computed, and its similarity to the stored embeddings is rapidly determined using **cosine similarity**. This efficient, algebra-based metric effectively replaces the need for time-consuming sequence alignment, confirming the viability of the alignment-free approach.

The DNABERT-S model, specifically, was trained using a contrastive loss function designed to maximize the distance between dissimilar sequences. This training approach yields standardized and predictable ranges of cosine similarity for various taxonomic ranks (e.g., an average similarity of 0.91 for sequences of the same species versus 0.62 for the same phylum). This property of explicit clustering is crucial for establishing quantitative, clinically reliable classification thresholds.

3.2 Operational Optimization: The Role of UMAP Dimensionality Reduction

To optimize the models for high-throughput clinical deployment, the Uniform Manifold Approximation and Projection (UMAP) algorithm was applied. This technique was used to reduce the dimensionality of the default 768-dimensional embeddings to targeted subsets ranging from 8 to 256 dimensions.

This step, intended primarily for computational efficiency, unexpectedly resulted in a measurable **improvement in classification accuracy**. This phenomenon suggests that the initial 768-dimensional embedding contained features that constituted noise or irrelevant details for the specific 16S classification task. UMAP acts as a filtering mechanism, generating a more concentrated and biologically meaningful vector representation that surpasses the performance of the original, higher-dimensional embedding. This confirms that UMAP is not merely an efficiency measure but an essential step for reaching optimal performance.

The clinical relevance of this optimization is underscored by the results for short sequences. Since the V3-V4 region (~400bp) is the current clinical standard for high-volume diagnostics, its reliable classification is paramount. The down-projected model, DNABERT-2@256, achieved performance enhancements of up to 12.5 percentage points over BLAST+ for V3-V4 classification. This validates the optimized methodology's immediate suitability for high-volume clinical and **direct market adoption** within the GenDAI platform.

3.3 Performance Validation: Speed, Accuracy, and Efficiency Metrics

The alignment-free methodology based on gLMs exhibits marked superiority over legacy tools across all critical metrics, ensuring operational viability in clinical environments. The accuracy benchmark for the optimal model, DNABERT-2@128, achieved an average classification accuracy of **98.0%** for the Greengenes2/Full16S task. This is a higher performance metric than both BLAST+ (97.6%) and VSEARCH (97.0%).

The crucial advantage, however, lies in computational speed. The vector similarity search is achieved up to **four orders of magnitude faster** than traditional alignment methods. Specifically, the runtime for the optimized DNABERT-2@128 was recorded at 0.35 seconds, compared to 1010 seconds for BLAST+ on the same task. This speed increase, approaching 2886 times, is the central technical enabler required to achieve the rapid turnaround times mandated for the automated clinical data processing pipeline (WP2 objective).

Table 1 is presented to compare these performance differentials formally.

Table 1: Comparative Performance of Alignment-Free Classification (DNABERT-2@128) vs. Traditional Methods (Greengenes2/Full16S Dataset)

Methodology	Classification Accuracy (Average)	Runtime (s)	Runtime Reduction Factor (vs. BLAST+)	Primary Classification Mechanism
BLAST+ (Alignment-Based)	97.6%	1010s	1x (Baseline)	Heuristic Alignment Search
VSEARCH (Alignment-Based)	97.0%	9.12s	110 times	Sequence Clustering/Alignment
DNABERT-2 @128 (Embedding/ UMAP)	**98.0%**	**0.35s**	**2886 times**	Cosine Similarity of Embeddings

3.4 Quality Assurance and Data Fidelity: Detection of Mislabeled Reference Sequences

A substantial, inherent advantage of the gLM methodology is its capacity for quality control through biological signal detection in the embedding space. This technique successfully identified potentially mislabeled sequences within widely used public repositories, specifically noting numerous misannotations in complex databases such as GTDB R220.

By analyzing pairwise cosine similarity scores, the system flagged mislabeled entries in critical genera (e.g., *Wolbachia*, *Escherichia*, and *Staphylococcus*) based on their geometric distance from expected phylogenetic clusters. This built-in quality control capability is essential for addressing the high pre-analytical/analytical variability outlined in the D1.1 requirements. This capacity to validate the integrity of the **reference standards** used for diagnosis represents a powerful mechanism for enhancing data fidelity in clinical and regulatory reviews.

Furthermore, the analysis of these mislabeled sequences revealed a critical trend: over 96% of the identified mislabeled sequences in GTDB R220 were traceable to metagenome-assembled genomes (MAGs) in NCBI. This finding highlights the pronounced risks associated with integrating non-verified data into diagnostic pipelines, functionally justifying GenDAI's stringent data quality protocols.

4. Advanced Methodology: Configuration & Augmentation for Robustness

Despite their high accuracy and speed, gLM embeddings exhibit a major operational limitation: they show **sensitivity to sequence length variations**. Since the Transformer architecture computes attention across the entire input sequence, minor differences in length, such as truncations of 50 base pairs, can significantly reduce the resulting cosine similarity. This length dependency poses a substantial theoretical risk, particularly when classifying sequences of variable length, which is common in shotgun metagenomics (random reads).

To ensure the clinical utility and reproducibility required by D1.1 (Section 2.5.2.B), the embeddings must be robustified against both natural biological variability and technical noise.

5. Operational Configuration and Deployment Blueprint

This section provides the implementation blueprint for T4.3 (Model Configuration and Operationalisation, Lead OKK) and T4.4 (Training UI, Lead ICT), detailing the integration necessary to bring the WP4 methodology into the functional GenDAI technical architecture.

5.1 High-Level Component Configuration for GenDAI Discovery

The GenDAI Discovery output is configured as a core service layer component, integrating model logic (WP4), persistence (WP3), and execution (WP2).

Table 2: GenDAI Discovery Component Mapping and Operational Requirements

GenDAI Discovery Component	**WP/Task Owner**	**D1.1 Requirement Addressed**	**Core Configuration Detail (M15 Status)**
Alignment-Free Classifier Engine	MTU (T4.1/T4.2)	Scalability, High-D Data Handling	DNABERT-2/S Embeddings, UMAP Projection, FAISS Vector Index (Query) [1]
Model Registry	OKK (T4.3 Lead)	Reproducibility,	PIDs (OKK ENS)

		Version Control	assigned to model weights and configurations; Metadata Logging [1]
Training UI	ICT (T4.4 Lead)	Clinical Integration, Fine-tuning	Split-Fed Learning Architecture, Explainability Interface [1]

5.2 Model Operationalization Requirements (T4.3 Input)

Operationalizing the gLMs mandates specialized infrastructure. **GPU Compute** is essential for efficiently generating embeddings and must be procured and configured via the WP3 Cloud Infrastructure (Google Cloud, leveraging Compute Engine or Vertex AI services).

Model Management and Provenance are critical for regulatory compliance. And as such, the **Model Registry** (T4.3, OKK) must utilize Persistent Unique Identifiers (PUIs), supplied by OKK's ENS technology, to track model versions. This ensures compliance with AI Explainability mandates and directives like IVDR 2017/746: every diagnostic finding (UC4) must be traceable back to the exact version of the gLM used, including its training history and hyperparameters. These PUIs will link the model weights and configurations directly to the GenDAI Safe Long-Term Archiving System (WP3).

The **Deployment Architecture** relies on container-based technologies (Docker/Kubernetes) within the WP3 cloud. This configuration maximizes portability, ensures scalability to handle large, variable loads, and guarantees integration resilience with the WP2 Workflow Management System (WMS).

5.3 Training UI Specification (T4.4 Input)

The Training UI (T4.4, ICT Lead) is designed to facilitate clinical engagement and localized innovation. Its architecture must incorporate a **Split-Fed Learning Framework**. This architecture enables authorized users to refine the final layer of the globally trained model using their own local, pre-processed data, supporting personalized diagnostics and local validation efforts without violating GDPR protocols concerning the sharing of sensitive raw data.

To support **Human-in-the-Loop Validation**, the UI must transparently present model results. This includes displaying **probabilistic scores and confidence intervals** derived from the gLM cosine similarity search, and presenting biomarker candidates as clear **ranked lists** for clinician review and actionability.

Finally, the UI must integrate a dedicated **Explainability Interface** to fulfill the transparency requirement. This interface must visually integrate XAI outputs (e.g., visual representations of SHAP or LIME outputs) to allow users to efficiently investigate the feature attribution driving a classification decision, thereby enhancing clinical trust and utility.

6. Integration and Compliance: Alignment with Clinical Workflow

This section details the critical integration of the WP4 methodology into the GenDAI platform's operational and clinical framework. It outlines how the alignment-free classification engine connects to the core diagnostics workflow, addresses the essential need for explainability to ensure clinical trust, and incorporates ethical safeguards for bias mitigation.

6.1 Integrating Classification into the WP2 GenDAI Diagnostics Workflow

The WP4 Alignment-Free Classifier Engine is engineered as a core computational service by the WP2 Diagnostics Workflow, which operates as an event-based Workflow Management System (WMS). The integration occurs via a Sample Analysis API. The classifier service provides rapid taxonomic assignment and biomarker identification, injecting the results directly into downstream modules, including WP5 (Interactive Reporting), facilitating the automated generation of personalized clinical reports. D4.1 confirms the technical prerequisite established for both the D2.1 Updated Integrated Diagnostics Pipeline (M24) and the D2.2 Final Diagnostics Pipeline (M30).

6.2 Addressing Explainability (XAI): SHAP/LIME Integration Requirements

The necessity for robust XAI is driven by the complexity of gLMs and the mandate for clinical trust. The implementation plan confirms the use of SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) techniques during the inference phase (UC8.2). These tools attribute the prediction to specific microbial features, quantifying their importance. To ensure continuity, D4.1 confirms that the model's data output structure is specifically designed to accommodate and transmit these XAI attribution metrics, ensuring WP5 can successfully parse and visualize them for the end-user interface.

6.3 Risk Mitigation: Addressing Bias

Regarding ethics and compliance, **Bias Mitigation** is integrated through the consistent handling of co-variables, fulfilling D1.1 ethics mandates. The protocol requires mandatory co-variables (age, sex, ethnicity) during model training and refinement to computationally control for and remove non-microbial phenotypic variation, ensuring that the model's predictions are focused on true microbial biomarkers and promote equitable diagnostic outcomes across all patient demographics.

7. Conclusions and Next Steps

Deliverable D4.1 successfully validates and documents the critical methodological approach of WP4 toward an Alignment-Free Classification methodology utilizing optimized Genomic Language Models. The methodology achieves a scientifically superior accuracy (98.0%) and delivers an unprecedented operational speed increase (2800 times) compared to conventional methods. This validation confirms the operational feasibility of high-throughput clinical processing and ensures WP4 is prepared to deliver on its RIOs. The report also formalizes the rigorous technical blueprint necessary for model deployment, specifically detailing the regulatory requirement for PUI-based model provenance and the split-fed learning architecture.

7.1 Immediate Forward Work Plan (M16 onwards for T4.3 and T4.4)

With the methodological definition complete, the focus immediately shifts to operational execution and advanced methodological research:

- **T4.3 Operationalization (OKK Lead):** Immediate focus is required on implementing the Model Registry, configuring PUI assignment (using OKK ENS) for comprehensive model provenance tracking, and establishing the Sample Analysis API gateway for WP2 integration.
- **T4.4 Training UI Development (ICT Lead):** Development of the Training UI must commence immediately, prioritizing the implementation of the split-fed learning architecture and the initial XAI visualization interface to prepare for early user testing.
- **WP6 Piloting Preparation:** The initial, validated gLM classification models must be prepared for seamless integration into the WP2 Diagnostics Workflow for deployment in the first piloting phase (T6.2, scheduled between M12-M36), ensuring end-users can evaluate real-world performance and provide crucial feedback.

References

[1] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, “DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome,” 2023, *arXiv*. doi: 10.48550/ARXIV.2306.15006.

[2] A. T. Merchant, S. H. King, E. Nguyen, and B. L. Hie, “Semantic design of functional de novo genes from a genomic language model,” *Nature*, Nov. 2025, doi: 10.1038/s41586-025-09749-7.

[3] H. Feng *et al.*, “Benchmarking DNA Foundation Models for Genomic and Genetic Tasks,” Aug. 18, 2024. doi: 10.1101/2024.08.16.608288.

[4] M. Leske, F. Bottacini, H. Afli, and B. G. N. Andrade, “BiGAMi: Bi-Objective Genetic Algorithm Fitness Function for Feature Selection on Microbiome Datasets,” *Methods Protoc.*, vol. 5, no. 3, p. 42, May 2022, doi: 10.3390/mps5030042.

[5] G. Benegas, C. Ye, C. Albors, J. C. Li, and Y. S. Song, “Genomic language models: opportunities and challenges,” *Trends Genet.*, vol. 41, no. 4, pp. 286–302, Apr. 2025, doi: 10.1016/j.tig.2024.11.013.

[6] Z. Zhou *et al.*, “DNABERT-S: pioneering species differentiation with species-aware DNA embeddings,” *Bioinformatics*, vol. 41, no. Supplement_1, pp. i255–i264, Jul. 2025, doi: 10.1093/bioinformatics/btaf188.

[7] H. Dalla-Torre *et al.*, “Nucleotide Transformer: building and evaluating robust foundation models for human genomics,” *Nat. Methods*, vol. 22, no. 2, pp. 287–297, Feb. 2025, doi: 10.1038/s41592-024-02523-z.