*future internet*

MDPI

*Article*

# Multi-Agent Readiness Scoring Methodology in Bioinformatics Domain

**Blagojche Gjorgjioski [1,2,*], Djansel Bukovec [1], Ivana Vichentijevikj [3], Ivan Kitanovski [1], Kostadin Mishev [1] and Monika Simjanoska Misheva [1]**

[1] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rudjer Boshkovikj 16, P.O. Box 393, 1000 Skopje, North Macedonia; dzhansel.bukovec@students.finki.ukim.mk (D.B.); ivan.kitanovski@finki.ukim.mk (I.K.); kostadin.mishev@finki.ukim.mk (K.M.); monika.simjanoska@finki.ukim.mk (M.S.M.)
[2] Adagon LLC, Ibe Palikukja 25/1-8, 1000 Skopje, North Macedonia
[3] iReason LLC, 3rd Macedonian Brigade 37, 1000 Skopje, North Macedonia; ivana@ireason.mk
* Correspondence: blagojche@adagon.tech; Tel.: +389-78-418-838

**Abstract**

The emergence of Large Language Models (LLMs) has significantly advanced computational biology, yet their integration into autonomous, multi-agent systems (MASs) and clinical workflows remains challenging due to systemic architectural fragmentation. To quantify the operational readiness and regulatory compliance of bioinformatics LLMs, we developed the Multi-Agent Readiness Score (MARS), a standardized evaluation framework assessing models across four structural dimensions: Governance & Accessibility, Biological Competence, Technical Maturity, and Agentic Orchestration. The framework incorporates compliance criteria from the EU AI Act, HL7 FHIR, HL7 CDA, and MyHealth@EU standards. To empirically validate this domain-agnostic methodology, we applied it to a highly mature subset of the field: a diverse cohort of 43 prominent genomic LLMs. Our assessment revealed a severe, industry-wide readiness gap: the majority of models fell into "Not Suitable" or "Research Prototype" tiers, lacking essential technical interfaces, structured communication schemas, and provenance tracking. Furthermore, the data demonstrated a 'competence-readiness gap', where models scale in biological predictive competence without corresponding improvements in engineering utility. The primary barrier to scalable bioinformatics AI is no longer biological competence, but operational and architectural incompatibility. By quantifying integration friction, MARS provides a crucial, reproducible metric to audit model maturity, guide system architecture, and ensure future models are structurally prepared for the rigorous regulatory demands of precision medicine workflows.

**Keywords:** Large Language Models; multi-agent systems; interoperability standards; AI Governance; Clinical Bioinformatics; ChatMed



Academic Editor: Xueyi Li

Received: 27 June 2026
Revised: 17 July 2026
Accepted: 27 July 2026
Published: 2 August 2026



## 1. Introduction

The emergence of LLMs and foundation models (FMs) has fundamentally transformed computational biology. Transformer-based architectures have demonstrated a remarkable capacity to learn the latent "languages" of biological systems, capturing the grammar of genomic sequences, the syntax of protein structure, and the regulatory semantics of gene expression [1]. These advances have enabled highly capable models for variant

effect prediction, regulatory element identification, epigenetic modeling, and molecular generation, often exceeding the performance of traditional bioinformatics pipelines.

However, real-world biomedical reasoning extends beyond isolated prediction tasks. Biological interpretation requires the integration of heterogeneous data modalities—including genomic, epigenomic, transcriptomic, structural, and clinical information, alongside iterative hypothesis generation and validation. This complexity exceeds the scope of any single model and has motivated a shift toward Multi-Agent Systems (MASs), in which specialized models collaborate through orchestrated workflows.

Despite this progress, empirical observations suggest that the primary bottleneck in bioinformatics automation has shifted from biological predictive limits to operational deployability. A persistent "readiness gap" exists where state-of-the-art models, while scientifically valid in isolation, frequently fail to collaborate within multi-agent workflows due to systemic architectural fragmentation. This friction is particularly evident when attempting to integrate established epigenetic editing pipelines [2], where state-of-the-art models—including DNABERT-2 [3], Geneformer [4], EpiGePT [5], and MuLan-Methyl [6]—often encounter architectural collapse rather than predictive failure. These recurring failures, driven by technical incompatibility rather than scientific error, underscore the urgent need for standardized readiness frameworks.

These technical limitations are amplified by regulatory requirements. Under the European Union Artificial Intelligence Act (EU AI Act), traceability, technical transparency, and auditability are no longer optional engineering practices but mandatory compliance obligations [7]. Likewise, integration into clinical infrastructures requires outputs aligned with Health Level Seven (HL7) Clinical Document Architecture (CDA), HL7 Fast Healthcare Interoperability Resources (FHIR) and MyHealth@EU semantic standards [8–10]. Models lacking structured communication protocols and provenance tracking therefore present significant barriers to integration and pose substantial compliance risks for deployment in regulated biomedical systems.

Observations of the current ecosystem reveal four recurring structural deficits that prevent bioinformatics models from functioning as reliable agents:

1. **Lack of Accountability and Provenance.** Most models do not maintain traceable execution logs or structured audit trails, preventing attribution of outputs to specific model versions or reasoning steps, and obstructing regulatory compliance.
2. **Biological Granularity Mismatch.** Many models require input representations that do not align with available experimental or clinical data, limiting their applicability in real-world reasoning workflows.
3. **Integration Instability.** The absence of standardized deployment environments, documentation, and interface specifications leads to brittle pipelines dependent on ad-hoc adaptation and fragile glue code.
4. **Absence of Structured Communication Protocols.** Most models produce unstructured natural language or tensor outputs, rather than structured, machine-interpretable schemas compatible with interoperable systems such as HL7 FHIR.

These findings demonstrate that scientific performance alone is insufficient to ensure clinical or operational utility. A critical missing component in the current ecosystem is a formal mechanism to evaluate whether bioinformatics LLMs are structurally and operationally prepared to function within orchestrated, compliant MAS. This gap motivates the development of standardized readiness frameworks capable of assessing interoperability, governance, and deployment readiness alongside predictive performance.

To address this gap, we introduce the Multi-Agent Readiness Score (MARS), a structured evaluation methodology designed to quantify the operational readiness of bioinformatics LLMs for deployment within interoperable MAS. Unlike traditional benchmarks

that measure predictive accuracy in isolation, MARS evaluates models across critical dimensions required for real-world orchestration, including governance, interoperability, technical maturity, and agentic compatibility. This methodology provides a systematic mechanism for assessing whether models are not only scientifically capable, but structurally prepared to function as reliable components within compliant biomedical AI ecosystems.

While the MARS is designed to be domain-agnostic across bioinformatics, applicable to proteomics, transcriptomics, and structural biology, we selected genomics as the initial validation cohort for this study. Genomic AI currently represents the most densely populated and scientifically mature sub-field of bioinformatics. Consequently, evaluating genomic LLMs provides the most rigorous stress test for quantifying the gap between established biological predictive power and nascent agentic orchestrations.

## 2. Related Work

### 2.1. Multi-Agent Systems in Bioinformatics

Emerging frameworks such as BioAgents [11], PromptBio [12], and CellAgent [13] demonstrate the feasibility of coordinated agentic analysis, while systems such as Biomni [14] and MedAgents [15] show that collaborative reasoning can outperform monolithic models in complex diagnostic and analytical tasks. Despite this progress, the practical deployment of bioinformatics MAS remains constrained by systemic architectural fragmentation. Most bioinformatics LLMs are disseminated as standalone research artifacts optimized for narrow benchmarks, without standardized interfaces or structured communication protocols. These models expose heterogeneous input and output representations, lack formalized schemas, and frequently require bespoke adaptation to function within larger pipelines. As a result, multi-model workflows remain fragile, difficult to reproduce, and inherently unstable, limiting the transition from experimental capability to deployable biomedical systems.

### 2.2. Governance and Interoperability Standards

Existing standards address complementary aspects of this ecosystem but do not resolve the operational requirements of agentic interoperability. Data governance frameworks such as FAIR [16] and TRUST [17] ensure data accessibility and repository reliability, while reporting standards including CONSORT-AI [18], DECIDE-AI [19], and TRIPOD-LLM [20] improve transparency in model evaluation. However, these frameworks focus on retrospective reporting, data stewardship, or organizational governance rather than defining architectural requirements for models to function as interoperable agents. Consequently, the field lacks a systematic framework for evaluating whether bioinformatics LLMs are structurally prepared for integration into orchestrated, compliant multi-agent systems.

This challenge is further intensified by the growing intersection of AI systems with regulated clinical environments. Under the EU AI Act [7], many biomedical AI systems are classified as “high-risk”, requiring traceability, auditability, and comprehensive technical documentation. Concurrently, participation in cross-border health infrastructures such as MyHealth@EU mandates semantic interoperability through structured standards including HL7 CDA and HL7 FHIR [8–10]. Yet most current bioinformatics models generate unstructured outputs and lack embedded provenance tracking, rendering them difficult to integrate into compliant clinical workflows.

## 3. Materials and Methods

The Multi-Agent Readiness Score (MARS) is a structured evaluation framework designed to assess the operational maturity of LLMs specifically for deployment in autonomous, multi-agent bioinformatics systems. Unlike general-purpose benchmarks that

focus solely on accuracy or reasoning capability, MARS evaluates a model's interoperability and its ability to function as a reliable, compliant, and technically sound component within a larger agentic workflows.

### 3.1. Validation Cohort: Model Selection and Scope

While the MARS methodology provides a generalized framework for the broader bioinformatics domain, its empirical validation requires a focused, domain-specific cohort. Therefore, to validate the framework, we curated a representative validation set of 43 Large Language Models (LLMs) specialized in genomics. Rather than an exhaustive systematic review, this cohort was assembled as a diverse test bed designed to capture both widely cited foundational models and niche, task-specific architectures published between 2021 and 2025.

Models were identified through literature repositories (e.g., PubMed, arXiv, bioRxiv) and model registries (e.g., Hugging Face, GitHub) based on the following strict inclusion criteria: (1) the architecture must utilize a Transformer or a modern foundational sequence-modeling paradigm (e.g., state-space models, advanced convolutional language models, or extended LSTMs); (2) the model must process, generate, or predict genomic, epigenomic, or directly related molecular data (including downstream multi-omic tasks such as protein-DNA binding, gene essentiality, or targeted oncology predictions); and (3) there must be a publicly accessible research paper, preprint, or formal technical report describing the system.

Models were explicitly excluded if they focused exclusively on 3D protein structure prediction (e.g., AlphaFold) or general clinical text processing (e.g., Electronic Health Record analysis). These domains represent distinctly different input granularities and operational workflows that fall outside the scope of this targeted molecular and genomic validation cohort.

### 3.2. Conceptual Foundation

The framework is built to bridge the gap between "academic proof-of-concepts" and "production-grade agents" by enforcing four core principles for translational biomedical AI: **Reproducibility**, **Interoperability**, **Accountability**, and **Specialization**.

- **Scope**: The framework assesses both Generative and Discriminative bioinformatics models, focusing on their readiness for integration into orchestrated pipelines (e.g., rigid workflows or autonomous agent systems).
- The primary intended users of MARS are:
  - **Bioinformatics Developers:** To select appropriate models for pipeline orchestration.
  - **System Architects:** To identify integration friction points before deployment.
  - **Regulatory and Compliance Teams:** To audit the provenance and safety of AI components in clinical or pre-clinical workflows.

### 3.3. Structural Organization

The MARS framework evaluates bioinformatics agents across four orthogonal dimensions, allowing for a decoupled assessment of scientific competence and engineering maturity. The assessment flows from foundational governance to high-level architectural capability:

1. **Dimension 1—Governance & Accessibility (Metadata):** This dimension functions as the "Gatekeeper" layer. It captures the immutable attributes of the model, such as licensing, source code availability, and versioning, that determine its legal and technical viability for pipelines. A model that fails here is structurally unavailable, regardless of its performance.

2. **Dimension 2—Biological Competence:** This dimension assesses the model's alignment with specific biological tasks. It evaluates whether the model operates on appropriate data granularities (e.g., single-cell vs. bulk) and validates its capability to handle domain-specific formats (e.g., FASTA, VCF) and reasoning types (e.g., perturbation analysis vs. static lookup).
3. **Dimension 3—Technical Maturity:** This dimension measures the "Friction of Deployment." It evaluates the engineering standards surrounding the model, ensuring that theoretically capable tools do not become "abandonware" due to dependencies, hardware constraints, or lack of containerization.
4. **Dimension 4—Agentic Orchestration:** This dimension evaluates "Interoperability." It focuses on features that allow the model to communicate autonomously with other agents, such as structured output schemas (JSON), API stability, and error handling, without human intervention.

### *3.4. Methodology and Question Design*

The framework utilizes a hybrid assessment model designed to capture both the qualitative context and quantitative readiness of a model. To operationalize the four dimensions described above, MARS departs from the uniform scoring often used in academic checklists. Instead, it employs a Weighted Readiness System where each question is assigned a point value (ranging from 1 to 5) proportionate to its criticality for autonomous agent workflows.

The complete rubric, detailing the specific evaluation criteria, answer options, and maximum point allocations across all dimensions, is provided in Table A1. The rubric is structured into two primary question categories:

1. **Governance & Metadata (Dimension 1–Non-Scored)**
   - **Purpose:** These questions (weighted at 0 points) serve as the "Gatekeeper" layer.
   - **Function:** They capture immutable attributes such as licensing status, source code availability, and versioning, that establish the model's legal and technical identity. While they do not contribute to the quantitative readiness score, they act as binary filters; a failure here (e.g., lack of open license) may disqualify a model from use regardless of its technical score.
2. **Weighted Performance Dimensions (Dimensions 2–4–Scored)**
   - **Purpose:** These questions evaluate specific capabilities across Biological Competence (Dim 2), Technical Maturity (Dim 3), and Agentic Orchestration (Dim 4).
   - **Scoring Logic:** Points are awarded based on the feature's impact on automation, with weights derived directly from regulatory blocking factors and architectural prerequisites:
     - **Critical Enablers (4–5 points):** Assigned to high-stakes capabilities derived from stringent interoperability and governance standards (e.g., EU AI Act, HL7 FHIR, HL7 CDA, MyHealth@EU) and essential multi-agent prerequisites (e.g., structured JSON I/O, Stability of Programmatic API ). As detailed in Section 3.5 and Table A2, their absence constitutes a fundamental compliance or deployment failure, preventing autonomous orchestration.
     - **Standard Features (2–3 points): Assigned to important operational standards and developer conveniences that streamline integration but do not strictly block autonomous execution (e.g., Ease of Installation, Documentation Quality).**
     - **Minor Enablers (1 point):** Assigned to supportive, highly specialized, or domain-specific features that enhance safety and usability but are not

universally mandated for baseline workflow integration (e.g., Sandboxed Testing Mode).

To mitigate evaluator subjectivity, the framework enforces an Evidence Rule. Scores are not awarded based on claimed capabilities or model popularity. A score $> 0$ is granted only when supported by substantiation, either through direct artifacts (e.g., documentation links, code repositories) or explicit evaluator commentary that details the functional verification of the feature. This ensures that capabilities are credited based on empirical observation rather than theoretical promise. If evidence is missing or the evaluator cannot explicitly justify the capability, the conservative option (lowest score) is selected by default.

### *3.5. Clinical Interoperability Compliance & Regulatory Alignment*

A defining characteristic of the MARS framework is its explicit alignment with emerging regulatory and clinical standards. Rather than treating compliance as a downstream concern, the framework incorporates criteria derived directly from the EU AI Act, specifically focusing on transparency, technical documentation, and human oversight.

Simultaneously, to ensure models are viable within modern electronic health ecosystems, the assessment questions are mapped to HL7 FHIR and CDA interoperability standards, as well as the MyHealth@EU infrastructure requirements. Consequently, a high MARS score indicates not only technical capability but also structural readiness for lawful clinical integration.

A detailed mapping of MARS questions to these specific regulatory frameworks is provided in Table A2.

### *3.6. The Scoring System: Four Tiers of Readiness*

The final output of the MARS framework is a unified "Readiness Score" that classifies the model into one of four categories. The tiered classification provides an immediate signal regarding the feasibility of integrating the model into an autonomous workflow. The four tiers of readiness and their descriptions are presented in Table 1.

**Table 1.** MARS Readiness Tiers and Operational Implications.

| Score Range | Readiness Tier | Operational Description |
|---|---|---|
| 0–40 pts | Level 0: Not suitable for orchestration | Early-stage research artifacts or not suitable for orchestration. These models often lack code availability, clear licensing, or documentation. While potentially novel, they are functionally "black boxes" suitable only for manual inspection, not automated integration. |
| 40–59 pts | Level 1: Research prototype, limited orchestration | Models that are functional in isolation but disjointed. They may be technically sound but biologically narrow, or biologically potent but lacking technical access (no APIs). Integrating them requires significant custom engineering or "glue code" to bridge these gaps. |
| 60–84 pts | Level 2: Partially ready, needs wrappers or adapters | Robust standalone tools that require adaptation for orchestration. While they likely possess decent documentation or technical foundations, they typically lack standardized interfaces or features required for clinical safety. Usage in a multi-agent system requires developing custom wrappers, adapters, or fine-tuning to align inputs and outputs. |
| 85–100 pts | Level 3: Plug & Play, fully multi-agent ready | The aspirational standard for bioinformatics AI. These models are designed with interoperability in mind, prioritizing standardized interfaces and transparent logging. They are fundamentally ready to participate in a multi-agent environment from day one with minimal configuration. |

### *3.7. Scoring Protocol and Reliability*

To ensure a rigorous and reproducible evaluation, all 43 models were assessed using a standardized protocol based on publicly available artifacts (e.g., published papers, code

repositories, and model registries). The evaluation of the full cohort was conducted by a primary bioinformatics engineer. To mitigate subjective bias, the MARS framework strictly enforced the Evidence Rule (detailed in Section 3.4): no model was awarded points for a capability unless explicit, verifiable proof (e.g., repository links, exact documentation quotes) was located. To validate the reliability of the primary evaluator and formally assess inter-rater agreement, a secondary independent reviewer blindly evaluated a randomly selected 20% subset of the models ($N = 9$). Discrepancies between the two evaluators were minimal and primarily stemmed from ambiguous deployment documentation. These edge cases were resolved through joint consensus discussion, ensuring strict calibration to the Evidence Rule across the entire dataset. Consequently, the scores represent the objective presence of documented technical artifacts rather than subjective quality judgments. The final, verified dataset including the exact evidence quotes and URLs justifying every assigned score is openly available in the Zenodo repository [21] to enable full third-party auditing.

## 4. Results

### *4.1. Overview of Model Readiness*

To validate the Multi-Agent Readiness Scoring (MARS) framework, we conducted a comprehensive evaluation of 43 LLMs specialized in genomics and bioinformatics [3,5,6,22–61]. This cohort spans widely cited foundational models as well as niche, task-specific architectures. Figure 1 summarizes the aggregated readiness scores across the full set, providing a high-level view of field maturity.

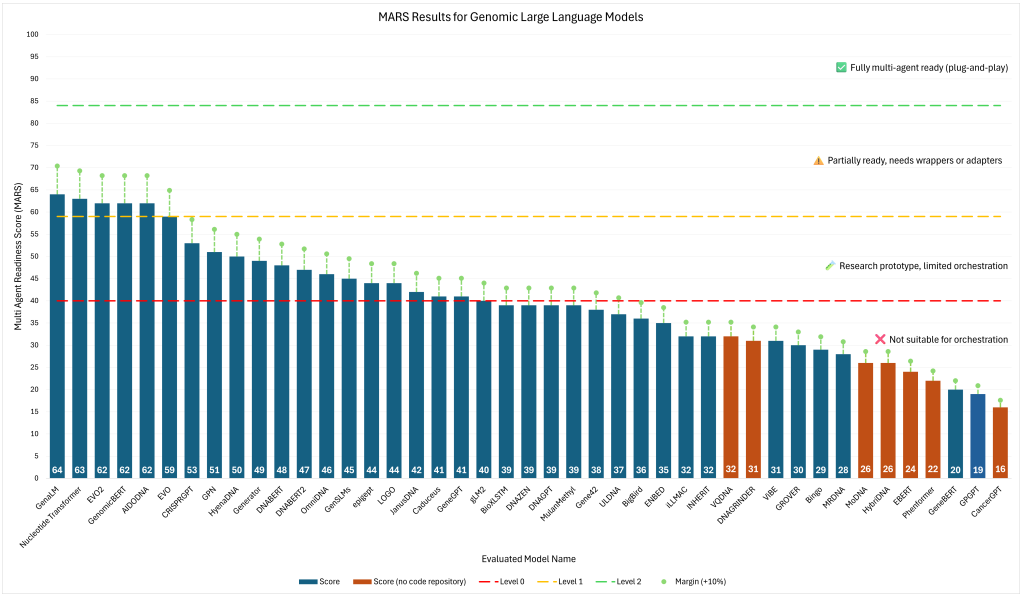


**Figure 1.** Evaluation of Large Language Model (LLM) Readiness for Bioinformatics Agents Total MARS framework scores for the cohort of 43 genomic LLMs, stratified by readiness tiers (Not Suitable, Research Prototype, Partially Ready, Plug & Play). The majority of models cluster in the lower quartiles, with zero models achieving the "Plug & Play" threshold (>84%). Error bars indicate a positive 10% tolerance margin to account for potential evaluation variability. Red bars denote models failing the fundamental "Availability" criteria (Dim1), representing systems with no publicly accessible source code or model weights.

The framework stratifies results into four Readiness Tiers defined by the framework: Not Suitable, Research Prototype, Partially Ready, and Plug & Play. As shown in Figure 1, the distribution is strongly skewed toward the lower tiers. Most models fall into Not

Suitable or Research Prototype, suggesting that despite potential academic value, they lack the structural robustness required for deployment.

Only a small subset (GenaLM [22], Nucleotide Transformer (NT) [23], EVO 2 [24], GenomicBERT [25], and AIDO.DNA [26]) reached the Partially Ready tier. These models can be integrated into pipelines but are not autonomous, consistently requiring external wrappers or middleware for input/output handling and state management. Notably, no model achieved Plug & Play status, underscoring a systemic gap that there is currently no off-the-shelf genomic LLM that can operate as a fully autonomous agent without substantial engineering intervention.

To ensure that these tier classifications accurately reflect structural readiness rather than artifacts of the rubric's specific point allocations, we conducted a sensitivity analysis. We recalculated the total scores using an unweighted, equal-item scoring model, where the raw score for each question was normalized by its maximum possible value, and all questions contributed equally to the final percentage. The model rankings demonstrated robust stability: 88.4% of the cohort (38 out of 43 models) retained their identical readiness tier classifications under the unweighted framework. This stability confirms that the observed readiness gap represents an intrinsic architectural limitation of the models rather than a byproduct of the evaluation weights.

Furthermore, academic software often suffers from incomplete documentation, raising the possibility that models possess capabilities unverified by our empirical evaluation. To systematically account for this potential documentation deficit and mitigate evaluator bias, Figure 1 incorporates a 10% positive error margin (green error bars). On the 100-point MARS scale, a 10-point buffer is mathematically equivalent to a model lacking documentation for three to four standard engineering features (e.g., structured error handling or deterministic execution). Even under this conservative, 'benefit of the doubt' adjustment, tier assignments remain largely static. Research Prototype models fail to cross the threshold into the Partially Ready tier, underscoring that the low readiness scores reflect systemic structural deficiencies rather than marginal evaluation variances or hidden codebase features.

Finally, the red visual indicators in Figure 1 highlight a transparency deficit. Although Dimension 1 (Governance & Metadata) does not contribute to the quantitative readiness score, it shows that 7 models (16.3%) do not meet basic reproducibility criteria, such as public release of source code or model weights. While many are described in peer-reviewed publications reporting state-of-the-art biological performance, the absence of accessible artifacts makes these systems unverifiable. From the perspective of open science and practical deployment, such closed systems impede progress by consuming research attention without contributing reusable infrastructure for agentic workflows.

### *4.2. The Competence-Readiness Gap: Domain Competence vs. Deployment Readiness*

To examine the relationship between biological competence and engineering utility, we mapped Biological Competence (Dimension 2) against Agentic Orchestration (Dimension 4). Figure 2 reveals a pronounced disconnect in the current genomic AI landscape.

Models span a wide range of biological competence along the X-axis, yet most remain compressed near the bottom of the Y-axis (Agentic Orchestration). This pattern reflects an a pronounced competence-readiness gap: genomic LLMs are scaling in biological knowledge but not in agency or interoperability. Even high-biological-scoring models often remain in the low-readiness region, consistent with designs optimized as standalone academic proofs-of-concept rather than interoperable system components.

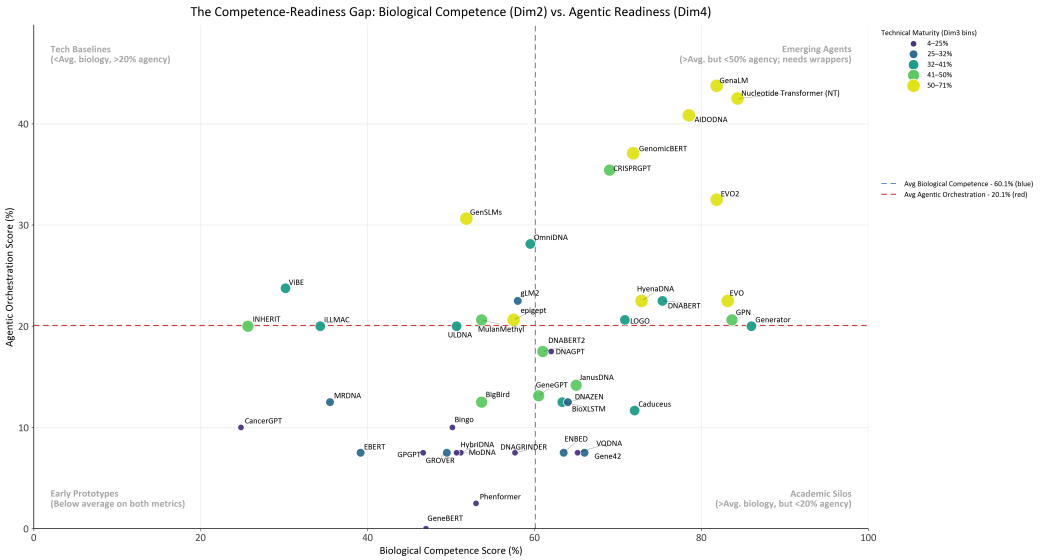


**Figure 2.** The Competence-Readiness Gap: Biological Competence vs. Agentic Readiness Scatter plot comparing normalized scores for Biological Competence (Dim2, x-axis) against Agentic Orchestration (Dim4, y-axis). The quadrant dividers represent the true cohort averages (60.1% and 20.1%, respectively) rather than absolute midpoints, categorizing models into four descriptive archetypes. The overall distribution reveals a non-linear relationship (Spearman's $\rho = 0.48$), where models scale in biological knowledge without corresponding increases in engineering utility. This structural disconnect is visually emphasized by the heavy clustering in the "Academic Silos" quadrant. Ultimately, this highlights that high genomic competency does not predict the ability to function within autonomous multi-agent workflows, as even the leading "Emerging Agents" remain below 50% agentic readiness and require manual integration wrappers.

This separation is also evident in cohort averages: mean Biological Competence (Dim2) is ∼60% (21/35), whereas mean Agentic orchestration (Dim4) remains below 20% (7/35), a three-fold gap. In addition, no model exceeds 50% on the Agentic orchestration axis, indicating a "glass ceiling" that constrains even best-in-class systems to largely passive question-answering rather than active workflow participation.

To quantify these relationships while rigorously avoiding part-whole correlation bias, a statistical artifact where a component variable artificially inflates the correlation when tested against a total score that includes it, we computed Spearman' s rank correlations $(r_s)$. Specifically, when evaluating the predictive power of individual technical features on overall readiness, we correlated the target feature's score against a modified total score (calculated as the Total MARS score minus the target feature's score). Agentic orchestration correlates strongly with technical maturity (Dim3) $(r_s = 0.81,\ p < 0.001)$ but only moderately with biological competence (Dim2) $(r_s = 0.48,\ p = 0.001)$, indicating substantial rank divergence between biological and agentic hierarchies. Crucially, the presence of a Stable Programmatic API (Dim4:1) remains a robust, highly significant predictor of overall model readiness even after removing its own weight from the dependent variable $(r_s = 0.70,\ p < 0.001)$. Similarly, Ease of Pipeline Integration (Dim3:3) maintains a strong predictive association $(r_s = 0.68,\ p < 0.001)$. This methodology confirms that overall multi-agent readiness is fundamentally driven by deliberate software engineering architectures rather than emerging automatically from stronger genomic pretraining.

### *4.3. Distribution of Readiness and Critical Deficiencies*

To further characterize the readiness gap, we examined the score distributions across the three primary rubric sections. The raincloud plot (Figure 3), which overlays bounded

density estimates, boxplots, and individual model beeswarm scatters, contrasts the maturity of biological competence with the immaturity of agentic engineering.

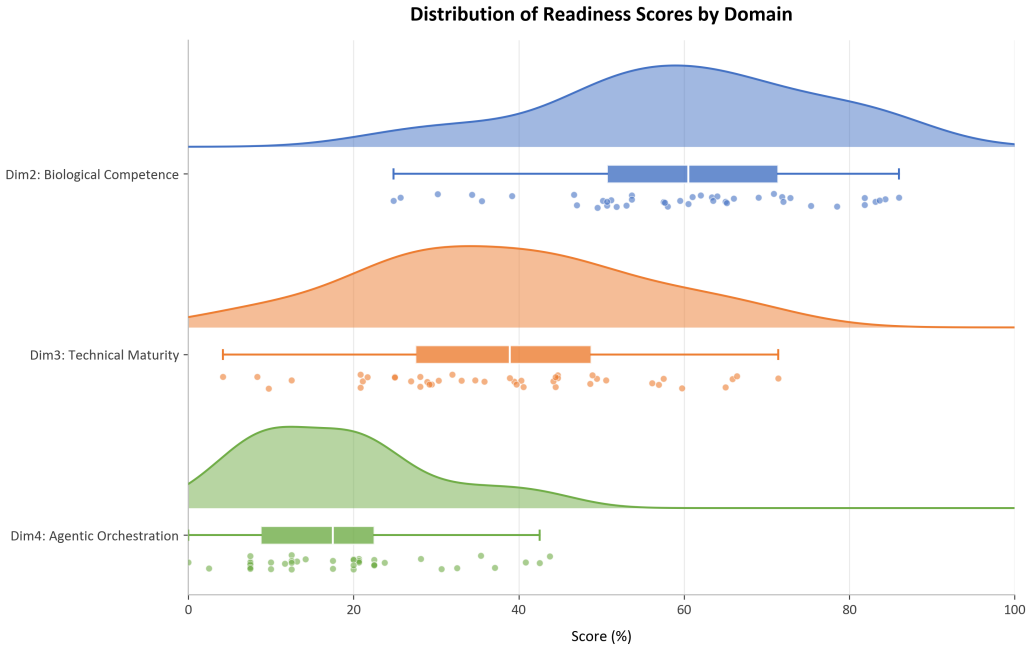


**Figure 3.** Raincloud Plot of Readiness Distributions: Score distributions across Biological Competence (blue), Technical Maturity (orange), and Agentic Orchestration (green). The visualization combines bounded density estimation (top), boxplots indicating medians and interquartile ranges (middle), and beeswarm scatter plots of individual model scores (bottom). The biological distribution features a wide spread and broad interquartile range, indicating a mature field with diverse model capabilities. In contrast, the agentic distribution is heavily skewed, with individual models clustering densely near zero. This visualization mathematically respects the 0–100% boundaries of the rubric and statistically confirms ($\chi_F^2 = 74.36$, $p < 0.001$) that the lack of agentic readiness is a systemic, industry-wide bottleneck rather than model-specific variance.

The Biological competence distribution (blue) exhibits a broad interquartile range and a wide spread of individual model scores, indicating a mature field with substantial variation from low to high performance. In contrast, the Agentic orchestration distribution (green) clusters densely near zero, with a highly compressed interquartile range and individual data points revealing a severe left-skew. This strict clustering at the lower bound demonstrates that agency is not merely weaker but a near-universal bottleneck — most models entirely lack the core engineering features required for autonomy.

To confirm that this divergence is structural rather than random, we performed a Friedman Test ( $\chi_F^2 = 74.36$, $p < 0.001$). The magnitude of the test statistic (74.36) and the extremely low $p$-value ($7.11 \times 10^{-17}$) provide strong statistical evidence that the performance distributions are significantly different. This result confirms that the field operates in two distinct performance regimes: a mature, variable distribution for biological competence, and a nascent, statistically inferior distribution for agentic readiness.

This skew is driven by systematic omissions in engineering practice. Three rubric dimensions received near-zero scores across the cohort, indicating field-wide failures:

1. Structured error handling (Dim4:7)—mean 0.02/3: models rarely emit structured error messages (e.g., JSON codes), instead producing verbose natural-language failures that hinder autonomous recovery loops.

2. Security standards (Dim3:6 & Dim4:8)—combined mean 0.15/4: on-premise deployment mechanisms and audit controls are largely absent, limiting compliance for clinical settings where privacy and traceability are critical.
3. Safe testing environments (Dim3:12)—mean 0.23/1: dedicated sandbox modes are uncommon, increasing risk when agents interact with real experimental infrastructure without continuous human oversight.

## 5. Discussion

Overall, these findings imply that the field is optimizing for discovery (biological performance) rather than utility (operational reliability). The primary barrier to agentic AI in bioinformatics is no longer biological knowledge but software engineering enabling LLMs to move from passive tools to workflow agents which will require prioritizing interface readiness (e.g., structured I/O and state management) over marginal gains on biological benchmarks.

### *5.1. Regulatory Implications: The Compliance Gap*

Critically, this deficit in software engineering has consequences that extend beyond technical inefficiency; it creates a fundamental incompatibility with the emerging legal landscape. As detailed in Table A2, every dimension of the MARS framework is explicitly mapped to requirements from the EU AI Act and HL7 FHIR standards. When the cohort's performance is analyzed through this regulatory lens, the "readiness gap" reveals itself to be a systemic compliance failure.

- **Transparency Violations (EU AI Act Art. 11 & 13):** Although Dimension 1 (Governance & Accessibility) is non-scored, the fact that 16.3% of models fail to provide basic artifacts like source code or weight hashes is critical. In a high-risk clinical setting, this opacity makes it impossible to construct the "Technical Documentation" required by the EU AI Act, making it highly challenging to legally deploy these models in clinical settings, regardless of their predictive accuracy.
- **Interoperability Dead Ends (HL7 FHIR & MyHealth@EU):** The competence-readiness gap observed in Figure 2— where models excel at biology but fail at agentic orchestration—has direct consequences for cross-border healthcare. Without the ability to enforce structured outputs (Dim4:7, mean score 0.02/3), these models cannot generate valid HL7 FHIR resources. Consequently, they are structurally incompatible with the MyHealth@EU infrastructure, operating as isolated silos rather than connected nodes in a European health data space.
- **Accountability Deficits (EU AI Act Art. 12):** A particularly notable finding is the near-total absence of immutable logging and transaction tracking (Dim4:8, mean 0.05/5). Under the EU AI Act, high-risk AI systems must provide automatic recording of events ("logs") to trace errors. The current cohort's inability to generate unique transaction IDs breaks the "chain of custody" required for post-market monitoring, meaning that tracing the provenance of clinical errors caused by these agents would be highly problematic.

In summary, while many current models exhibit high scientific capability, they lack the structural features necessary for regulatory compliance. Without the adoption of the readiness standards highlighted by the MARS framework to enforce these missing standards from day one, bioinformatics LLMs will remain "academic artifacts", therefore impressive in the lab, but difficult to operationalize within the regulated landscape of precision medicine.

### 5.2. Future Directions

To scale the impact of this framework, the evaluation paradigm must transition from a static rubric to an automated ecosystem. The critical next step is the development of MARS-CI, an open-source testing suite able to be integrated into CI/CD pipelines to programmatically verify compliance criteria such as API callability, schema validity, reproducibility and traceability. This automation should provide the technical foundation for integrating live readiness scores into major model repositories and bio-agent hubs (e.g., Hugging Face, BioContainers) and allowing them to be displayed as real-time "readiness badges". By embedding these metrics at the point of discovery, a self-regulating ecosystem can emerge where users filter models based on verified operational maturity, effectively shifting the community standard from post-hoc troubleshooting to pre-deployment quality assurance. Furthermore, as regulatory frameworks like the EU AI Act evolve, the MARS weighting logic must undergo continuous updates to reflect new compliance mandates, ensuring the framework remains a "living standard" for the bio-agent community.

The broader significance of this direction lies in the fact that these challenges are not unique to bioinformatics. Closely related technical failures are also emerging across other biomedical AI domains, especially in AI-enabled drug discovery and development, where heterogeneous components such as molecular generation models, docking tools, ADMET predictors, and therapeutic reasoning agents must function together within reproducible pipelines. Based on our experience in this area, agentic systems like Prompt-to-Pill [62] encounter identical limitations: fragile interfaces, unstructured outputs, and insufficient traceability. Therefore, this suggests that MARS should evolve toward a generalized biomedical readiness protocol. By preserving core governance and interoperability principles while adapting domain-specific criteria such as validating Simplified Molecular Input Line Entry System (SMILES) or Structure Data File (SDF) molecular outputs or Protein Data Bank(PDB) or PDBQT (PDB with partial charges and atom types) structural inputs we can establish a universal standard for interoperable, auditable, deployment-ready biomedical AI systems.

### 5.3. Limitations

While the MARS framework provides a necessary standardized metric for agentic readiness, this study has several limitations. First, MARS is a checklist-based evaluation framework that assesses documented technical and structural readiness; it does not measure real-world deployment performance, execution speed, or live clinical efficacy. A model scoring highly on the MARS rubric is structurally prepared for integration, but its actual utility still depends on the specific multi-agent pipeline environment. Second, despite the strict enforcement of an 'Evidence Rule' and independent auditing to minimize bias, the evaluation inherently relies on the completeness of publicly available documentation. Models may possess capabilities (such as deterministic execution or secure deployment modes) that were not captured in this assessment simply because they were undocumented at the time of evaluation. Future work will focus on transitioning MARS from a manual rubric to an automated CI/CD testing suite to continuously verify these capabilities programmatically.

## 6. Conclusions

The future of precision medicine relies not merely on the existence of powerful AI models, but on their ability to reason collaboratively within compliant, automated workflows. However, this study establishes that the primary barrier to such scalability is no longer the scientific competence of individual models, but their operational incompatibility. Through the application of the Multi-Agent Readiness Score (MARS) to a cohort of over 40 diverse LLMs, we have empirically quantified the "integration debt" currently accumulating in the field. Our findings reveal a landscape rich in specialized biological reasoners yet exhibiting a significant deficit of architecturally compliant agents. The majority of tools operate as scientifically valid but technically isolated artifacts, requiring prohibitive custom engineering to participate in collaborative workflows.

The MARS framework addresses this fragmentation by shifting the evaluation paradigm from generative quality to systemic interoperability. By explicitly linking technical specifications, such as structured I/O and containerization, with rigorous provenance and regulatory requirements (EU AI Act, HL7 FHIR, HL7 CDA and MyHealth@EU), the framework serves as a necessary governance layer for the bio-agent ecosystem. We argue that the path toward autonomous discovery cannot rely on the post-hoc adaptation of incompatible tools. Instead, it requires that future models are developed with the connectivity, transparency, and auditability features necessary for high-stakes clinical environments from day one. Ultimately, standardized readiness assessment is the prerequisite for transforming a collection of isolated algorithms into a cohesive, trustworthy, and interoperable multi-agent ecosystem.

**Author Contributions:** Conceptualization, B.G. and M.S.M.; Methodology, B.G. and I.K.; Formal Analysis, B.G. and D.B.; Investigation, D.B. and I.V.; Data Curation, D.B.; Writing—Original Draft Preparation, B.G.; Writing—Review & Editing, I.V., I.K., and K.M.; Validation, K.M.; Resources, D.B.; Supervision, B.G. and M.S.M.; Project Administration, B.G. and M.S.M.; Funding Acquisition, M.S.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the European Union under Horizon Europe, grant number 101159214.

**Institutional Review Board Statement:** Ethics approval was not required as this research evaluates computational models and involves no human or animal subjects.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The original data presented in the study are openly available in Zenodo at https://doi.org/10.5281/zenodo.21404473 (accessed on 17 July 2026) reference number [21]. The detailed scoring results, evaluation matrices, and full dataset for the MARS (Multi-Agent Readiness Score) assessment of the 43 genomic Large Language Models analyzed in this study are openly available for review and reproducibility. These results provide the granular data supporting the findings illustrated in Figures 1–3.

**Acknowledgments:** During the preparation of this manuscript, the authors used Google Gemini 3.1 Pro for the purposes of improving English grammar, readability, and flow. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** Author Mr. Gjorgjioski was employed by the company Adagon LLC and author Ms. Vichentijevikj was employed by the company iReason LLC. These employments did not influence the design, execution, or interpretation of this research. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Appendix A. MARS Framework for Evaluating Biomedical Large Language Models

**Table A1.** The MARS Framework for Evaluating Biomedical Large Language Models.

| # | Question | Answer Type | Answer Options | Max Points |
|---|---|---|---|---|
| **DIMENSION 1: Governance & Accessibility (Metadata)** | | | | |
| Dim1:1 | Model identifier (name, version, release date) | Free text | Open | 0 |
| Dim1:2 | Developer/Institution/ Organization | Free text | Open | 0 |
| Dim1:3 | Model architecture | Free text | Open | 0 |
| Dim1:4 | License(s) type (paper, source code, model) | Free text (structured) | Examples: MIT/Apache 2.0/CC BY 4.0/CC BY-NC 4.0/(Proprietary/custom)/Mixed (specify: paper/code/weights) | 0 |
| Dim1:5 | Access type | Multi-select choice | Hosted API/Source code/Pretrained model weights/Container/Binary executable/Demo -webUI only/No public access | 0 |
| Dim1:6 | Documentation available? | Links | Open | 0 |
| Dim1:7 | Source code available? | Links | Open | 0 |
| Dim1:8 | Training data sources (high-level) | Free text | Open | 0 |
| **DIMENSION 2: Biological Competence** | | | | |
| Dim2:1 | Which biological domains are empirically demonstrated by experiments, benchmarks, or case studies in the model's documentation or publication? | Choice | 0-No biological domain demonstrated<br>1-One biological domain demonstrated<br>2-Two distinct biological domains demonstrated<br>3-Three biological domains demonstrated<br>4-Four biological domains demonstrated<br>5-Five or more biological domains demonstrated | 5 |
| Dim2:2 | What task types are supported? | Choice | 0-No biological task demonstrated<br>1-One task type demonstrated<br>2-Two distinct task types demonstrated<br>3-Three or more distinct task types demonstrated | 3 |
| Dim2:3 | Can the model interpret or produce common bioinformatics formats? | Choice | 0-No standard bioinformatics formats used<br>1-Accepts or produces one standard format (e.g., FASTA, GFF)<br>2-Accepts and produces standard formats<br>3-Supports multiple standard formats across inputs and outputs | 3 |
| Dim2:4 | Training or fine-tuning on domain-specific datasets? | Choice | 0-No domain datasets<br>1-One dataset<br>2-Multiple datasets from one domain<br>3-Multiple datasets across domains<br>4-Multiple curated benchmark datasets | 4 |
| Dim2:5 | Is the model evaluated against established, named baseline methods using standard quantitative metrics? | Choice | 0-No benchmark comparison<br>1-Compared only to internal baselines or ablations<br>2-Compared to established, named methods using standard metrics | 2 |
| Dim2:6 | Is the model validated on real-world experimental or clinical data? | Choice | 0-Synthetic or simulated only<br>1-Retrospective real-world data<br>3-Curated real-world datasets (ClinVar, EHR, GWAS)<br>5-Prospective or operational deployment | 5 |

**Table A1.** *Cont.*

| # | Question | Answer Type | Answer Options | Max Points |
|---|---|---|---|---|
| Dim2:7 | Outputs reference biological evidence (e.g., motifs, conservation)? | Choice | 0-No biological evidence or interpretation provided<br>1-Basic biological rationale provided<br>2-Uses known biological signals (e.g., motifs, conservation)<br>3-Demonstrates mechanistic biological interpretation<br>4-Multiple lines of biological evidence consistently validated | 4 |
| Dim2:8 | How broad is the biological scope of the model's pretraining and downstream tasks, based on species coverage and biological problem diversity? | Choice | 0-Single task on a single biological problem<br>1-Multiple tasks within a single biological problem<br>2-Multiple tasks within a single biological domain<br>3-Multiple biological domains within one species<br>4-Multiple biological domains across multiple species<br>5-Broad biological scope across species, domains, and tasks | 5 |
| Dim2:9 | Does the model explicitly cite or link to biomedical databases (e.g., PubMed, ENCODE, UniProt, ClinVar)? | Choice | 0-No citations or links to biomedical databases<br>1-References biomedical databases without direct links<br>2-Explicit citations and links to biomedical databases | 2 |
| Dim2:10 | Is there evidence the model is updated with new biomedical knowledge? | Choice | 0-No post-release updates, releases, or retraining events documented<br>1-Minor post-release update (e.g., small patch, doc update, or maintenance commit)<br>2-Major post-release update (new model version, retraining, new datasets) | 2 |
| **DIMENSION 3: Technical Maturity & Integration** | | | | |
| Dim3:1 | Structured I/O supported (e.g., JSON in/out)? | Choice | 0-No documented structured external I/O<br>1-Structured outputs only (no input schema)<br>2-Structured inputs and outputs, but no published schema<br>3-Published, reusable schema (e.g., JSON, OpenAPI, protobuf) | 3 |
| Dim3:2 | How complex is it to install and run the model in a fresh environment, based on dependencies, hardware requirements, and setup steps? | Choice | 0-Not installable with provided instructions<br>1-Requires specialized hardware and legacy stack (e.g., multi-GPU, pinned old frameworks)<br>2-Requires GPU and non-trivial environment setup<br>3-Requires GPU but uses modern, well-supported stack<br>4-CPU/GPU optional; standard dependency management<br>5-Single-command install (e.g., pip install, Docker etc.) with minimal setup | 5 |
| Dim3:3 | How easily can the model be integrated into an existing software or data pipeline after installation? | Choice | 0-Integration requires major code rewriting<br>1-Script-based only; tightly coupled to data preprocessing<br>2-Modular scripts but no reusable interfaces<br>3-Importable modules or stable CLI<br>4-Well-documented library-style integration<br>5-Clean API with examples and backward compatibility | 5 |

**Table A1.** *Cont.*

| # | Question | Answer Type | Answer Options | Max Points |
|---|---|---|---|---|
| Dim3:4 | Have examples for reproducability? | Choice | 0-No reproducibility guidance or examples<br>1-Partial reproducibility instructions or examples<br>2-Clear, end-to-end reproducibility examples provided | 2 |
| Dim3:5 | How usable is the model without additional fine-tuning or domain adaptation? | Choice | 0-Requires fine-tuning; no defaults/baseline provided<br>1-Fine-tuning required but baseline pipeline or checkpoints provided<br>2-Usable out-of-the-box; fine-tuning improves performance<br>3-Out-of-the-box + configurable without retraining (prompting/config only) | 3 |
| Dim3:6 | Supports secure deployment (on-prem, access control)? | Choice | 0-No secure deployment considerations described<br>1-Secure or isolated deployment considerations explicitly described | 1 |
| Dim3:7 | Outputs are consistent and reproducible? | Choice | 0-Reproducibility not addressed<br>1-Partial controls (e.g., some fixed seeds or version pinning)<br>2-Explicit deterministic inference guarantees | 2 |
| Dim3:8 | Is hallucination (generation of biologically implausible sequences) reported or likely? | Choice | N/A → auto-assign 2-Non-generative system<br>0-Generative system, no mitigation<br>1-Partial mitigation<br>2-Explicit mitigation strategy | 2 |
| Dim3:9 | Does the model support or enhance human-in-the-loop workflows | Choice | 0-No HITL workflow described<br>1-Human review occurs but is external to the system<br>2-Human feedback is operationally integrated into the system loop | 2 |
| Dim3:10 | Are quantitative latency or throughput metrics reported, and are they clearly labeled as training or inference? | Choice | 0-No latency or throughput metrics reported<br>1-Metrics reported but not clearly scoped (training vs inference)<br>2-Clearly reported and labeled latency or throughput metrics | 2 |
| Dim3:11 | Is cost or resource efficiency demonstrated relative to established baselines? | Choice | 0-No efficiency analysis<br>1-Efficiency claimed without baseline comparison<br>2-Efficiency demonstrated relative to named baselines | 2 |
| Dim3:12 | Is a dedicated sandbox or safe testing mode explicitly provided? | Choice | 0-No sandbox or safe test mode<br>1-Explicit sandbox, demo environment, or isolated test mode described | 1 |
| **DIMENSION 4: Agentic Orchestration** | | | | |
| Dim4:1 | Does the model expose a stable programmatic API for invocation by external agents or systems? | Choice | 0-No programmatic API<br>1-Ad-hoc scripts only (manual modification required)<br>2-Unstable or undocumented API (CLI, Python, etc.)<br>3-Documented API (CLI or library)<br>4-Stable API with versioning<br>5-Versioned, backward-compatible API designed for orchestration | 5 |

**Table A1.** *Cont.*

| # | Question | Answer Type | Answer Options | Max Points |
|---|---|---|---|---|
| Dim4:2 | Does the model produce task-specific metadata/tags in output? | Choice | 0-No standardized metadata<br>1-Informal or ad-hoc metadata<br>2-Structured metadata, not standardized<br>3-Standardized metadata schema<br>4-Metadata explicitly designed for agent interoperability | 4 |
| Dim4:3 | Is the system's functional role explicitly declared in a way that an external agent or orchestrator can rely on? | Choice | 0-No role information<br>1-Role implied by usage or examples<br>2-Role inferred from structured artifacts (tasks/modules/IO patterns), but not stated<br>3-Role explicitly stated in documentation (human-readable)<br>4-Role explicitly stated and constrained (clear inputs/outputs, scope)<br>5-Role declared in a system/agent manifest (machine-readable) | 5 |
| Dim4:4 | Supports role conditioning and/or dynamic role switching? (e.g., prompt-based, context-adaptive) | Choice | 0-No role conditioning or switching<br>1-Static role configuration only<br>2-Limited role selection via configuration<br>3-Role switching supported through retraining or reconfiguration<br>4-Dynamic role switching supported programmatically<br>5-Dynamic role switching integrated into an agent framework | 5 |
| Dim4:5 | Can an external system automatically infer the model's capabilities from machine-readable metadata? | Choice | 0-No metadata<br>1-Informal human-readable metadata only<br>2-Structured metadata, not standardized<br>3-Standardized metadata (e.g., model card schema)<br>4-Metadata sufficient for automated capability inference<br>5-Machine-actionable metadata designed for agent orchestration | 5 |
| Dim4:6 | Is a standardized agent integration interface provided? (e.g., function calling, schema definitions, orchestration libraries) | Choice | 0-No standardized integration interface<br>1-Informal or ad-hoc integration patterns<br>2-Partially standardized interface<br>3-Standardized interface documented<br>4-Standardized interface with examples<br>5-Standardized interface designed for agent orchestration | 5 |
| Dim4:7 | Does the system expose structured error handling or fallback mechanisms for programmatic use? | Choice | 0-No error handling described<br>1-Errors logged or printed only<br>2-Structured error codes or exceptions<br>3-Explicit fallback or recovery mechanisms | 3 |
| Dim4:8 | Does the system record provenance or trace information sufficient for auditing or reproducibility? | Choice | 0-No provenance or trace logging<br>1-Basic runtime logs only<br>2-Input/output identifiers, hashes, or version tags recorded<br>3-End-to-end traceability (inputs, parameters, versions) | 3 |

**Note:** Bold text indicates the primary dimensions of the framework. The # symbol denotes the unique identifier for each question (e.g., Dim1:1).

## Appendix B. Alignment of the MARS Framework with Regulatory and Interoperability Standards

**Table A2.** Alignment of the MARS Framework with Regulatory and Interoperability Standards.

| ID | Question | Principles & Standards |
|---|---|---|
| Dim1:1 | Model identifier (name, version, release date) | **EU AI Act:** Transparency (Art. 13)/Technical Documentation (Art. 11)<br>**HL7 FHIR:** Device Resource (DeviceName, Version)<br>**HL7 CDA:** Header (SoftwareName/Version) |
| Dim1:2 | Developer/Institution/Organization | **EU AI Act:** Provider Obligations (Art. 16)<br>**HL7 FHIR:** Organization Resource<br>**HL7 CDA:** Custodian/Author |
| Dim1:3 | Model architecture | **EU AI Act:** Technical Documentation (Art. 11) |
| Dim1:4 | License(s) type (paper, source code, model) | **EU AI Act:** Transparency/IP Rights |
| Dim1:5 | Access type | **MyHealth@EU:** Access Control/Security<br>**EU AI Act:** Accessibility requirements |
| Dim1:6 | Documentation available? | **EU AI Act:** Instructions for Use (Art. 13)<br>**MyHealth@EU:** Deployment documentation |
| Dim1:7 | Source code available? | **EU AI Act:** Auditability/Transparency |
| Dim1:8 | Training data sources (high-level) | **EU AI Act:** Data Governance (Art. 10)/Transparency (Art. 13)<br>**HL7 FHIR:** Provenance (Entity source) |
| Dim2:1 | Which biological domains are empirically demonstrated by experiments, benchmarks, or case studies in the model's documentation or publication? | **EU AI Act:** Intended Purpose (Art. 7)<br>**HL7 FHIR:** Clinical Terminology (SNOMED CT, LOINC)<br>**MyHealth@EU:** Master Value Sets |
| Dim2:2 | What task types are supported? | **EU AI Act:** Intended Purpose (Art. 7)<br>**HL7 FHIR:** Task Resource/Workflow Module |
| Dim2:3 | Can the model interpret or produce common bioinformatics formats? | **HL7 FHIR:** Genomics Reporting/MolecularSequence Resource<br>**HL7 CDA:** Structured Body (Level 2/3 CDA)<br>**MyHealth@EU:** Semantic Interoperability |
| Dim2:4 | Training or fine-tuning on domain-specific datasets? | **EU AI Act:** Data Governance (Art. 10) |
| Dim2:5 | Is the model evaluated against established, named baseline methods using standard quantitative metrics? | **EU AI Act:** Accuracy & Robustness (Art. 15) |
| Dim2:6 | Validated on real-world experimental or clinical data? | **EU AI Act:** Post-market monitoring/Accuracy (Art. 15)<br>**MyHealth@EU:** Quality assurance |
| Dim2:7 | Outputs reference biological evidence (e.g., motifs, conservation)? | **HL7 FHIR:** Provenance/DiagnosticReport (evidence)<br>**HL7 CDA:** Reference/External Evidence<br>**EU AI Act:** Explainability |
| Dim2:8 | How broad is the biological scope of the model's pretraining and downstream tasks, based on species coverage and biological problem diversity? | **EU AI Act:** Data Governance (Art. 10) |

**Table A2.** *Cont.*

| ID | Question | Principles & Standards |
|---|---|---|
| Dim2:9 | Does the model explicitly cite or link to biomedical databases (e.g., PubMed, ENCODE, UniProt, ClinVar)? | **HL7 FHIR:** RelatedArtifact/Citation Resource<br>**HL7 CDA:** References to external documents<br>**EU AI Act:** Transparency/Record Keeping |
| Dim2:10 | Is there evidence the model is updated with new biomedical knowledge? | **EU AI Act:** Post-market monitoring (Art. 61)<br>**MyHealth@EU:** Maintenance & Lifecycle Management |
| Dim3:1 | Structured I/O supported (e.g., JSON in/out)? | **HL7 FHIR:** Core Specification (JSON/XML support)<br>**HL7 CDA:** XML Structure<br>**MyHealth@EU:** Structured Patient Summary (EPSOS) |
| Dim3:2 | How complex is it to install and run the model in a fresh environment, based on dependencies, hardware requirements, and setup steps? | **MyHealth@EU:** National Contact Point (NCPeH) deployment feasibility |
| Dim3:3 | How easily can the model be integrated into an existing software or data pipeline after installation? | **HL7 FHIR:** Implementation Guides (IGs)<br>**MyHealth@EU:** Technical Interoperability |
| Dim3:4 | Have examples for reproducability? | **EU AI Act:** Technical Documentation (Art. 11)<br>**HL7 FHIR:** Example scenarios in IGs |
| Dim3:5 | How usable is the model without additional fine-tuning or domain adaptation? | **N/A:** Specific to model lifecycle, not standard compliance |
| Dim3:6 | Supports secure deployment (on-prem, access control)? | **EU AI Act:** Cybersecurity (Art. 15)<br>**MyHealth@EU:** Security & Trust Services (Circle of Trust)<br>**HL7 FHIR:** Security Module (OAuth2/SMART on FHIR) |
| Dim3:7 | Outputs are consistent and reproducible? | **EU AI Act:** Accuracy, Robustness (Art. 15)<br>**HL7 FHIR:** AuditEvent (for tracing consistency) |
| Dim3:8 | Is hallucination (generation of biologically implausible sequences) reported or likely? | **EU AI Act:** Risk Management System (Art. 9)/Accuracy (Art. 15)<br>**MyHealth@EU:** Patient Safety risks |
| Dim3:9 | Does the model support or enhance human-in-the-loop workflows | **EU AI Act:** Human Oversight (Art. 14)<br>**HL7 FHIR:** Task/Workflow (Human actor assignment) |
| Dim3:10 | Are quantitative latency or throughput metrics reported, and are they clearly labeled as training or inference? | **EU AI Act:** Technical Documentation (Computational resources)<br>**MyHealth@EU:** Service Level Agreements (SLAs) |
| Dim3:11 | Is cost or resource efficiency demonstrated relative to established baselines? | **EU AI Act:** Environmental reporting (for GPAI) |
| Dim3:12 | Is a dedicated sandbox or safe testing mode explicitly provided? | **EU AI Act:** AI Regulatory Sandboxes (Art. 53)<br>**HL7 FHIR:** Test Servers/Connectathons |
| Dim4:1 | Does the model expose a stable programmatic API for invocation by external agents or systems? | **HL7 FHIR:** RESTful API (GET/POST interactions)<br>**MyHealth@EU:** Cross-border gateway services<br>**EU AI Act:** Interoperability (specifically for high-risk AI) |

**Table A2.** *Cont.*

| ID | Question | Principles & Standards |
| --- | --- | --- |
| Dim4:2 | Does the model produce task-specific metadata/tags in output? | **HL7 FHIR:** Meta element/Coding/Tags<br>**HL7 CDA:** Header metadata<br>**EU AI Act:** Transparency (marking AI output) |
| Dim4:3 | Is the system's functional role explicitly declared in a way that an external agent or orchestrator can rely on? | **HL7 FHIR:** Device (Type/Classification)<br>**HL7 CDA:** Participant roles |
| Dim4:4 | Supports role conditioning and/or dynamic role switching? (e.g., prompt-based, context-adaptive) | **N/A:** Specific to Agentic AI architecture, not currently standardized |
| Dim4:5 | Can an external system automatically infer the model's capabilities from machine-readable metadata? | **HL7 FHIR:** CapabilityStatement (auto-discovery of capabilities)<br>**EU AI Act:** Transparency |
| Dim4:6 | Is a standardized agent integration interface provided? (e.g., function calling, schema definitions, orchestration libraries) | **HL7 FHIR:** OperationDefinition/StructureDefinition<br>**MyHealth@EU:** Interface specifications |
| Dim4:7 | Does the system expose structured error handling or fallback mechanisms for programmatic use? | **HL7 FHIR:** OperationOutcome (Standardized error handling)<br>**EU AI Act:** Robustness (Fail-safe modes) |
| Dim4:8 | Does the system record provenance or trace information sufficient for auditing or reproducibility? | **EU AI Act:** Automatic Recording of Events (Logs) (Art. 12)<br>**HL7 FHIR:** Provenance/AuditEvent<br>**HL7 CDA:** Legal Authenticator/Data Enterer |

**Note:** Bold text is used to denote the primary column headers and emphasize key regulatory frameworks.

## References

1. Li, Q.; Hu, Z.; Wang, Y.; Li, L.; Fan, Y.; King, I.; Jia, G.; Wang, S.; Song, L.; Li, Y. Progress and opportunities of foundation models in bioinformatics. *Brief. Bioinform.* **2024**, *25*, bbae548. [CrossRef] [PubMed]
2. Broche, J.; Kungulovski, G.; Bashtrykov, P.; Rathert, P.; Jeltsch, A. Genome-wide investigation of the dynamic changes of epigenome modifications after global DNA methylation editing. *Nucleic Acids Res.* **2020**, *49*, 158–176. [CrossRef] [PubMed]
3. Zhou, Z.; Ji, Y.; Li, W.; Dutta, P.; Davuluri, R.; Liu, H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome. *arXiv* **2023**. [CrossRef]
4. Theodoris, C.V.; Xiao, L.; Chopra, A.; Chaffin, M.D.; Al Sayed, Z.R.; Hill, M.C.; Mantineo, H.; Brydon, E.M.; Zeng, Z.; Liu, X.S.; et al. Transfer learning enables predictions in network biology. *Nature* **2023**, *618*, 616–624. [CrossRef] [PubMed]
5. Gao, Z.; Liu, Q.; Zeng, W.; Jiang, R.; Wong, W.H. EpiGePT: A Pretrained Transformer model for epigenomics. *bioRxiv* **2024**. [CrossRef] [PubMed]
6. Zeng, W.; Gautam, A.; Huson, D. MuLan-Methyl—Multiple transformer-based language models for accurate DNA methylation prediction. *GigaScience* **2023**, *12*, giad054. [CrossRef] [PubMed]
7. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance). EUR-Lex. 2024. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 15 May 2026).
8. HL7 International. Health Level Seven Clinical Document Architecture, Release 2.0 (CDA R2). 2005. ANSI/HL7 OID: 2.16.840.1.113883.1.11.20456. Available online: https://www.hl7.org/implement/standards/product_brief.cfm?product_id=7 (accessed on 10 May 2026).
9. Health Level Seven International. Health Level Seven (HL7) Standard. 2024. Available online: https://www.hl7.org/fhir/ (accessed on 5 May 2026).

10. European Commission. MyHealth@EU Test Framework. European Commission Online Document. 2023. Available online: https://health.ec.europa.eu/ehealth-digital-health-and-care/electronic-cross-border-health-services_en (accessed on 5 May 2026).
11. Mehandru, N.; Hall, A.K.; Melnichenko, O.; Dubinina, Y.; Tsirulnikov, D.; Bamman, D.; Alaa, A.; Saponas, S.; Malladi, V.S. BioAgents: Bridging the gap in bioinformatics analysis with multi-agent systems. *Sci. Rep.* **2025**, *15*, 39036. [CrossRef] [PubMed]
12. Zhang, M.; Gu, W.; Han, B.; Guo, V.; Addoni, C.; Chen, J.; Ma, Y.; Leng, Y.; Li, K.; Lin, X.; et al. PromptBio: A Multi-Agent AI Platform for Bioinformatics Data Analysis. *bioRxiv* **2025**. [CrossRef]
13. Xiao, Y.; Liu, J.; Zheng, Y.; Xie, X.; Hao, J.; Li, M.; Wang, R.; Ni, F.; Li, Y.; Luo, J.; et al. CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis. *arXiv* **2024**. [CrossRef]
14. Huang, K.; Zhang, S.; Wang, H.; Qu, Y.; Lu, Y.; Roohani, Y.; Li, R.; Qiu, L.; Li, G.; Zhang, J.; et al. Biomni: A General-Purpose Biomedical AI Agent. *bioRxiv* **2025**. [CrossRef] [PubMed]
15. Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; Gerstein, M. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv* **2024**. [CrossRef]
16. 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.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
17. Lin, D.; Crabtree, J.; Dillo, I.; Downs, R.R.; Edmunds, R.; Giaretta, D.; De Giusti, M.; L'Hours, H.; Hugo, W.; Jenkyns, R.; et al. The TRUST Principles for digital repositories. *Sci. Data* **2020**, *7*, 144. [CrossRef] [PubMed]
18. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K.; Chan, A.W.; Darzi, A.; Holmes, C.; Yau, C.; Ashrafian, H.; et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nat. Med.* **2020**, *26*, 1364–1374. [CrossRef] [PubMed]
19. DECIDE-AI Expert Group; Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.; Collins, G.; Denaxas, S.; Denniston, A.; Faes, L.; Geerts, B.; et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [CrossRef] [PubMed]
20. Gallifant, J.; Afshar, M.; Ameen, S.; Aphinyanaphongs, Y.; Chen, S.; Cacciamani, G.; Demner-Fushman, D.; Dligach, D.; Daneshjou, R.; Fernandes, C.; et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **2025**, *31*, 60–69. [CrossRef] [PubMed]
21. Gjorgjioski, B.; Bukovec, D.; Vichentijevikj, I.; Kitanovski, I.; Mishev, K.; Simjanoska Misheva, M. [Dataset] Multi-Agent Readiness Score (MARS) Evaluation of Genomic LLMs. 2026. Available online: https://zenodo.org/records/21404473 (accessed on 26 June 2026).
22. Fishman, V.; Kuratov, Y.; Petrov, M.; Shmelev, A.; Shepelin, D.; Chekanov, N.; Kardymon, O.; Burtsev, M. GENA-LM: A Family of Open-Source Foundational DNA Language Models for Long Sequences. *bioRxiv* **2023**. [CrossRef]
23. Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Carranza, N.L.; Grzywaczewski, A.H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B.P.; Sirelkhatim, H.; et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *Nat. Methods* **2025**, *22*, 287–297. [CrossRef] [PubMed]
24. Brixi, G.; Durrant, M.G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G.A.; King, S.H.; Li, D.B.; Merchant, A.T.; et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv* **2025**. [CrossRef]
25. Chen, T.; Vahab, N.; Tyagi, N.; Cummins, E.; Peleg, A.Y.; Tyagi, S. genomicBERT: A Light-weight Foundation Model for Genome Analysis using Unigram Tokenization and Specialized DNA Vocabulary. *bioRxiv* **2025**. [CrossRef]
26. Ellington, C.N.; Sun, N.; Ho, N.; Tao, T.; Mahbub, S.; Li, D.; Zhuang, Y.; Wang, H.; Song, L.; Xing, E.P. Accurate and General DNA Representations Emerge from Genome Foundation Models at Scale. *bioRxiv* **2025**. [CrossRef]
27. Nguyen, E.; Poli, M.; Durrant, M.G.; Thomas, A.W.; Kang, B.; Sullivan, J.; Ng, M.Y.; Lewis, A.; Patel, A.; Lou, A.; et al. Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv* **2024**. [CrossRef]
28. Huang, K.; Qu, Y.; Cousins, H.; Johnson, W.A.; Yin, D.; Shah, M.; Zhou, D.; Altman, R.; Wang, M.; Cong, L. CRISPR-GPT: An LLM Agent for Automated Design of Gene-Editing Experiments. *arXiv* **2024**. [CrossRef]
29. Benegas, G.; Batra, S.S.; Song, Y.S. DNA language models are powerful predictors of genome-wide variant effects. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2311219120. [CrossRef] [PubMed]
30. Nguyen, E.; Poli, M.; Faizi, M.; Thomas, A.; Birch-Sykes, C.; Wornow, M.; Patel, A.; Rabideau, C.; Massaroli, S.; Bengio, Y.; et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv* **2023**, arXiv:2306.15794.
31. Wu, W.; Li, Q.; Zhang, Y.; Zhan, Z.; Chen, R.; Li, M.; Fu, K.; Qi, J.; Bao, Y.; Wang, C.; et al. GENERator: A Long-Context Generative Genomic Foundation Model. *arXiv* **2026**. [CrossRef]
32. Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R.V. DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **2021**, *37*, 2112–2120. [CrossRef] [PubMed]
33. Li, Z.; Subasri, V.; Shen, Y.; Li, D.; Zhao, Y.; Stan, G.B.; Shan, C. Omni-DNA: A Unified Genomic Foundation Model for Cross-Modal and Multi-Task Learning. *arXiv* **2025**. [CrossRef]

34. Zvyagin, M.; Brace, A.; Hippe, K.; Deng, Y.; Zhang, B.; Bohorquez, C.O.; Clyde, A.; Kale, B.; Perez-Rivera, D.; Ma, H.; et al. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. *Int. J. High Perform. Comput. Appl.* **2023**, *37*, 683–705. [CrossRef]
35. Yang, M.; Huang, L.; Huang, H.; Tang, H.; Zhang, N.; Yang, H.; Wu, J.; Mu, F. Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res.* **2022**, *50*, e81. [CrossRef] [PubMed]
36. Duan, Q.; Huang, B.; Song, Z.; Lehmann, I.; Gu, L.; Eils, R.; Wild, B. JanusDNA: A Powerful Bi-directional Hybrid DNA Foundation Model. *arXiv* **2025**. [CrossRef]
37. Schiff, Y.; Kao, C.H.; Gokaslan, A.; Dao, T.; Gu, A.; Kuleshov, V. Caduceus: Bi-directional equivariant long-range DNA sequence modeling. In Proceedings of the 41st International Conference on Machine Learning, ICML'24, Vienna, Austria, 21–27 July 2024.
38. Jin, Q.; Yang, Y.; Chen, Q.; Lu, Z. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *arXiv* **2023**. [CrossRef]
39. Cornman, A.; West-Roberts, J.; Camargo, A.P.; Roux, S.; Beracochea, M.; Mirdita, M.; Ovchinnikov, S.; Hwang, Y. The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv* **2024**. [CrossRef]
40. Schmidinger, N.; Schneckenreiter, L.; Seidl, P.; Schimunek, J.; Hoedt, P.J.; Brandstetter, J.; Mayr, A.; Luukkonen, S.; Hochreiter, S.; Klambauer, G. Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences. *arXiv* **2024**. [CrossRef]
41. Mao, L.; Tian, Y.; Song, Y. DNAZEN: Enhanced Gene Sequence Representations via Mixed Granularities of Coding Units. *arXiv* **2025**. [CrossRef]
42. Zhang, D.; Zhang, W.; Zhao, Y.; Zhang, J.; He, B.; Qin, C.; Yao, J. DNAGPT: A Generalized Pre-trained Tool for Versatile DNA Sequence Analysis Tasks. *arXiv* **2023**. [CrossRef]
43. Vishniakov, K.; Ben Amor, B.; Tekin, E.; Elnaker, N.; Viswanathan, K.; Medvedev, A.; Singh, A.; Nadeem, M.; Sayeed, M.; Kanithi, P.; et al. Gene42: Long-Range Genomic Foundation Model With Dense Attention. *arXiv* **2025**. [CrossRef]
44. Zhu, Y.H.; Liu, Z.; Liu, Y.; Ji, Z.; Yu, D.J. ULDNA: Integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein–DNA binding site prediction. *Brief. Bioinform.* **2024**, *25*, bbae040. [CrossRef] [PubMed]
45. Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. Big Bird: Transformers for Longer Sequences. *arXiv* **2021**. [CrossRef]
46. Malusare, A.; Kothandaraman, H.; Tamboli, D.; Lanman, N.A.; Aggarwal, V. Understanding the natural language of DNA using encoder–decoder foundation models with byte-level precision. *Bioinform. Adv.* **2024**, *4*, vbae117. [CrossRef] [PubMed]
47. Liu, J.; Shen, H.; Chen, K.; Li, X. Large language model produces high accurate diagnosis of cancer from end-motif profiles of cell-free DNA. *Brief. Bioinform.* **2024**, *25*, bbae430. [CrossRef] [PubMed]
48. Bai, Z.; Zhang, Y.z.; Miyano, S.; Yamaguchi, R.; Fujimoto, K.; Uematsu, S.; Imoto, S. Identification of bacteriophage genome sequences with representation learning. *Bioinformatics* **2022**, *38*, 4264–4270. [CrossRef] [PubMed]
49. Li, S.; Wang, Z.; Liu, Z.; Wu, D.; Tan, C.; Zheng, J.; Huang, Y.; Li, S.Z. VQDNA: Unleashing the power of vector quantization for multi-species genomic sequence modeling. In Proceedings of the 41st International Conference on Machine Learning, ICML'24, Vienna, Austria, 21–27 July 2024.
50. Zhao, Q.; Zhang, C.; Zhang, W. dnaGrinder: A lightweight and high-capacity genomic foundation model. *arXiv* **2024**. [CrossRef]
51. Gwak, H.J.; Rho, M. ViBE: A hierarchical BERT model to identify eukaryotic viruses using metagenome sequencing data. *Brief. Bioinform.* **2022**, *23*, bbac204. [CrossRef] [PubMed]
52. Sanabria, M.; Hirsch, J.; Poetsch, A.R. The human genome's vocabulary as proposed by the DNA language model GROVER. *bioRxiv* **2023**. [CrossRef]
53. Ma, J.; Song, J.; Young, N.D.; Chang, B.C.H.; Korhonen, P.K.; Campos, T.L.; Liu, H.; Gasser, R.B. 'Bingo'—A large language model- and graph neural network-based workflow for the prediction of essential genes from protein data. *Brief. Bioinform.* **2023**, *25*, bbad472. [CrossRef] [PubMed]
54. Zeng, W.; Huson, D.H. Enhanced 5mC-Methylation-Site Recognition in DNA Sequences using Token Classification and a Domain-specific Loss Function. *bioRxiv* **2025**. [CrossRef]
55. An, W.; Guo, Y.; Bian, Y.; Ma, H.; Yang, J.; Li, C.; Huang, J. Advancing DNA Language Models through Motif-Oriented Pre-Training with MoDNA. *BioMedInformatics* **2024**, *4*, 1556–1571. [CrossRef]
56. Ma, M.; Liu, G.; Cao, C.; Deng, P.; Dao, T.; Gu, A.; Jin, P.; Yang, Z.; Xia, Y.; Luo, R.; et al. HybriDNA: A Hybrid Transformer-Mamba2 Long-Range DNA Language Model. *arXiv* **2025**. [CrossRef]
57. Trotter, M.V.; Nguyen, C.Q.; Young, S.; Woodruff, R.T.; Branson, K.M. Epigenomic language models powered by Cerebras. *arXiv* **2021**. [CrossRef]
58. Träuble, F.; Stuart, L.; Georgiou, A.; Notin, P.; Mehrjou, A.; Schwessinger, R.; Chevalley, M.; Branson, K.; Schölkopf, B.; van Duijn, C.; et al. Multi-megabase scale genome interpretation with genetic language models. *arXiv* **2025**. [CrossRef]

59. Mo, S.; Fu, X.; Hong, C.; Chen, Y.; Zheng, Y.; Tang, X.; Shen, Z.; Xing, E.P.; Lan, Y. Multi-modal Self-supervised Pre-training for Regulatory Genome Across Cell Types. *arXiv* **2021**. [CrossRef]
60. Lyu, Y.; Wu, Z.; Zhang, L.; Zhang, J.; Li, Y.; Ruan, W.; Liu, Z.; Yu, X.; Cao, C.; Chen, T.; et al. GP-GPT: Large Language Model for Gene-Phenotype Mapping. *arXiv* **2024**. [CrossRef]
61. Li, T.; Shetty, S.; Kamath, A.; Jaiswal, A.; Jiang, X.; Ding, Y.; Kim, Y. CancerGPT for few shot drug pair synergy prediction using large pretrained language models. *npj Digit. Med.* **2024**, *7*, 40. [CrossRef] [PubMed]
62. Vichentijevikj, I.; Mishev, K.; Simjanoska Misheva, M. Prompt-to-Pill: Multi-Agent Drug Discovery and Clinical Simulation Pipeline. *Bioinform. Adv.* **2025**, *6*, vbaf323. [CrossRef] [PubMed]