Preparing Life Sciences Data for AI

Artificial intelligence is becoming imperative in almost every project and strategy within life science organizations. With AI models improving rapidly, there is added pressure to be efficient with resources. Initiatives are meticulously planned to include foundation models, large language models, digital twins, and predictive pipelines. Leadership expectations are defined and even computing resources are approved. However, progress often slows when work reaches the data.

Data from early discovery to late-stage trials typically resides across multiple systems. Data remains distributed across shared folders and repositories, S3 storage, Electronic Lab Notebooks (ELNs), legacy laboratory information management systems, DNAnexus environments, and large collections of spreadsheets. Metadata lack consistency, while ontologies differ across studies and platforms.

Variant calls and clinical annotations also exist in separate systems, and the problem is compounded by collaborators submitting data in incompatible formats. Similar challenges arise in agri-tech, where data generated across laboratory, greenhouse, and field stages is captured using disparate ontologies, normalization standards, and experimental frameworks, making aggregation and cross-stage analysis difficult. Complex R&D initiatives, therefore, require coordinated data orchestration across multiple platforms, study environments, sequencing pipelines, and downstream analytical systems. As a result, data scientists spend a large share of their time preparing inputs rather than building models.

Furthermore, AI initiatives require not only AI-ready datasets, but also decision-ready data frameworks that enable meaningful interpretation, prioritization, and downstream action. AI-ready infrastructure alone is insufficient without structured knowledge generation capabilities that help connect data, context, and downstream research outcomes.

Common Pressure Points

Data Harmonization

Fragmented legacy silos and disparate partner schemas create massive bottlenecks in training AI models. At Strand, we are experts in unifying such incompatible data stores into structured, FAIR-compliant datasets. Our AI-enabled pipelines achieve **3x faster ingestion** and **>95% automated normalization accuracy** across complex ontologies, while enabling integration of genomic, transcriptomic, environmental, and experimental datasets to support deeper biological interpretation across diverse research conditions, transforming raw data into high-quality, queryable assets.

Target Discovery and Decision Enablement

Target discovery is increasingly limited not by data generation, but by the ability to transform complex biological datasets into actionable insights. Strand's AI-powered decision enablement platforms integrate multi-omics, experimental, clinical, and environmental data with foundation-model and agentic AI approaches to identify, prioritize, and interpret high-value targets faster and more confidently. Our solutions support applications ranging from biomarker and therapeutic target discovery in diagnostics and drug development to trait prioritization, crop improvement, and precision agriculture. By combining predictive modeling, knowledge-driven analytics, and human-in-the-loop scientific workflows, we help organizations accelerate discovery, reduce experimental burden, and make better research decisions across healthcare and agritech domains.

Variant & Clinical Interpretation Automation

Manual curation tends to be the ceiling on clinical genomics throughput and time-to-insight. Strand's agentic LLM tools automate cohort building, metadata curation, and aspects of insight reduction, leading to **10–20x faster** insights overall. This includes **festiVAR**, which automates variant prioritization, genotype–phenotype correlation, and ACMG-guided classification to reduce interpretation time from six hours to under two while achieving **99.93% accuracy** in identifying reported variants within the top 25 genes, and **Blitz**, an LLM-powered literature retrieval and evidence extraction tool that uses retrieval-augmented generation or RAG over PubMed to map evidence directly to ACMG criteria, **reducing manual curation effort by 40–60%** and enabling scalable, rapid variant interpretation and reclassification workflows.

AI-Driven Scientific Decision Enablement

AI-enabled scientific decision frameworks help transform large-scale multi-omic and experimental datasets into structured, actionable knowledge that supports prioritization, hypothesis generation, translational interpretation, and downstream research decisions. Strand's agentic AI and chatbot platforms enable researchers to interrogate complex biomedical datasets using natural language, automate cohort discovery and evidence retrieval, and generate database-grounded insights with **~90% accuracy**, reducing time-to-insight by approximately **7x**.

Multi-Modal & Multi-Omic Integration

Linking genomics, transcriptomics, and real-world clinical data is a significant engineering hurdle for drug discovery. Strand builds the end-to-end infrastructure to harmonize these disconnected layers, having successfully integrated longitudinal RWD for 500,000+ patients records and unified single-cell and proteomics data into analysis-ready disease atlases.

Strand's orchestration frameworks also support integration of molecular, experimental, and real-world datasets to enable deeper biological interpretation across diverse research conditions. These capabilities help connect multi-omic insights with downstream phenotypic and translational outcomes, supporting more effective discovery and development workflows.

Strand prepares scientific data for AI applications so research teams can work with datasets that support reliable analysis and model development.

Illustrative past case studies

1. Data Ingestion & Harmonization for Multi-Omics

Need: A unified, analysis-ready dataset from UK Biobank's primary care records to support robust longitudinal modeling for an innovative biotech company.

Challenge: GP data spans multiple formats and coding systems, requiring complex transformation, ontology mapping, and normalization to ensure interoperability, accuracy, and completeness across a large, heterogeneous patient record base.

Solution: Deployed a multi-layered automated pipeline using GPT-4o for dataset identification and NLP-based metadata extraction, combined with semantic search to identify mislabeled "edge case" datasets.

Impact: Substantially **accelerated curation timelines** with a significant reduction in manual effort, achieving **high sensitivity and specificity** in identifying usable datasets for deeper therapeutic insights.

2. AI-Enabled Data Orchestration and Metadata Harmonization

Need: Rapid discovery and curation of specific scRNA-seq samples and clinical metadata to fuel AI-driven disease modeling and target prioritization, for a leading biotechnology organization.

Challenge: Fragmented legacy R&D silos and manual curation bottlenecks delayed critical insights by weeks, making it impossible to scale data ingestion across disparate partner schemas.

Solution: Strand deployed a custom platform, utilizing agentic AI for automated cohort building and RAG-based ontology mapping. This unified the various multi-omic datasets into a structured, FAIR-compliant database.

Impact: Achieved **3x faster data ingestion** and a **10–20x reduction** in time-to-insight, transforming manual curation from weeks into hours with **>95% automated normalization accuracy**.

3. Data Ingestion & Harmonization for Multi-Omics

Need: A unified, analysis-ready dataset from UK Biobank's primary care records to support robust longitudinal modeling for an innovative biotech company.

Challenge: GP data spans multiple formats and coding systems, requiring complex transformation, ontology mapping, and normalization to ensure interoperability, accuracy, and completeness across a large, heterogeneous patient record base.

Solution: Deployed a multi-layered automated pipeline using GPT-4o for dataset identification and NLP-based metadata extraction, combined with semantic search to identify mislabeled "edge case" datasets.

Impact: Substantially **accelerated curation timelines** with a significant reduction in manual effort, achieving **high sensitivity and specificity** in identifying usable datasets for deeper therapeutic insights.

4. Accelerating CRISPR Screen Data Curation

Need: An automated, reliable pipeline to curate and standardize CRISPR screening data from heterogeneous sources for one of the world's leading biopharmaceutical companies.

Challenge: Inconsistent metadata across databases, growing data volumes, and the expert effort required for manual extraction make curation slow, error-prone, and difficult to scale.

Solution: Strand designed a tailored metadata schema and operational workflow that integrated reference ontologies for organism, genotype, and spatial context, consolidating data from public databases, cloud repositories, and CROs.

Impact: Harmonized over 40 diverse datasets and **filtered 25,000+ datasets (~96,000 samples)**, accelerating downstream ML processes with a 5-day turnaround for smaller batches.

5. Empowering Curation Scalability and Insights Through Chatbots

Need: A scalable, intelligent interface to explore and query a large, multi-dimensional neurological disorder target database efficiently for one of the world's leading biopharmaceutical companies.

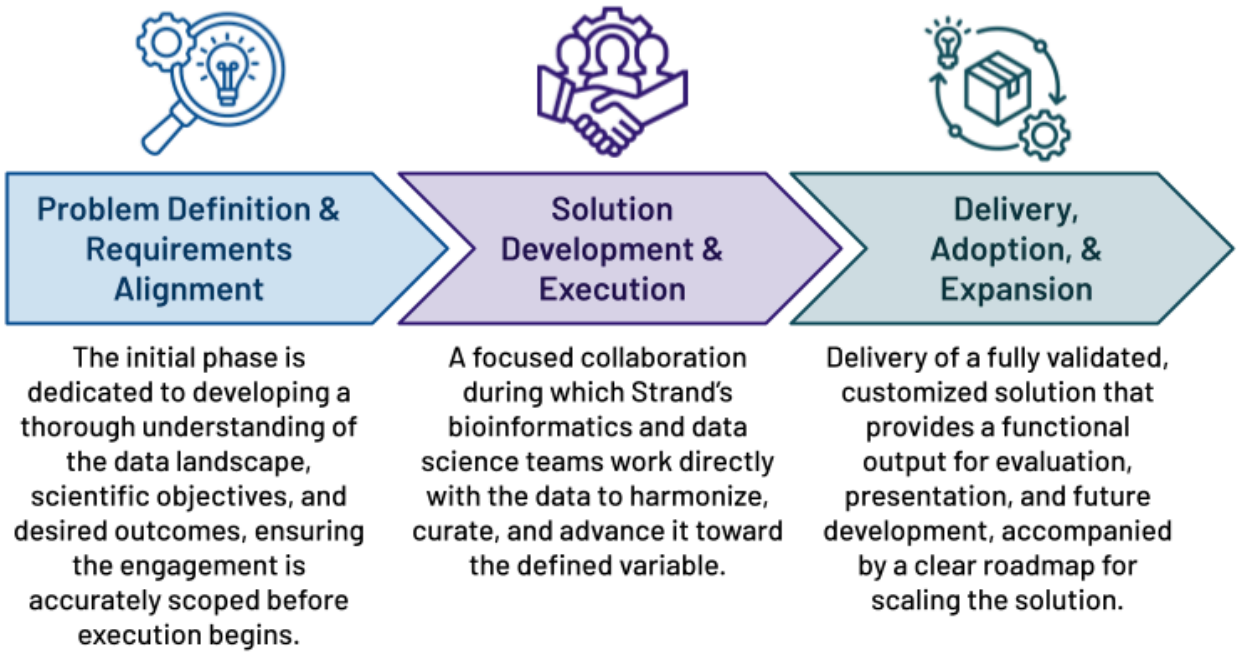
Challenge: The volume and complexity of 1200+ entries across 40+ metadata fields makes manual or spreadsheet-based exploration impractical for identifying patterns at scale.

Solution: Analyzed claims for ~1,000 patients across core diagnostic areas (Oncology, Rare Disease, NIPT), using data-driven questions to evaluate clinical effectiveness and redundant billing patterns.

Impact: I**dentified a 50% redundant billing rate** for certain panels and a **3.5% false positive** rate in NIPT, empowering the payer with a strategic framework for policy revision and cost optimization.

Engaging with Strand

At Strand we work with you closely throughout the project from start to finish and provide support beyond.



WHY STRAND

Strand Life Sciences is a bioinformatics technology company with 25+ years of deep expertise in bioinformatics, clinical genomics, and data science. The company develops scalable bioinformatics and AI-driven platforms for genomic analysis, including omics-focused AI foundation models for target discovery, phenotype-driven analysis, variant interpretation, assay development, and single-cell and spatial omics, enabling scalable interpretation of genomic data for diagnostics, precision medicine, and translational research.

With a multidisciplinary team of around 500 professionals, Strand supports global pharmaceutical, diagnostic, healthcare, agricultural biotechnology, and research organizations through structured knowledge generation, and modular, scalable solutions that accelerate discovery, improve clinical trial efficiency, support product development decisions, and deliver actionable biological insights.

For more information, please visit us.strandls.com.