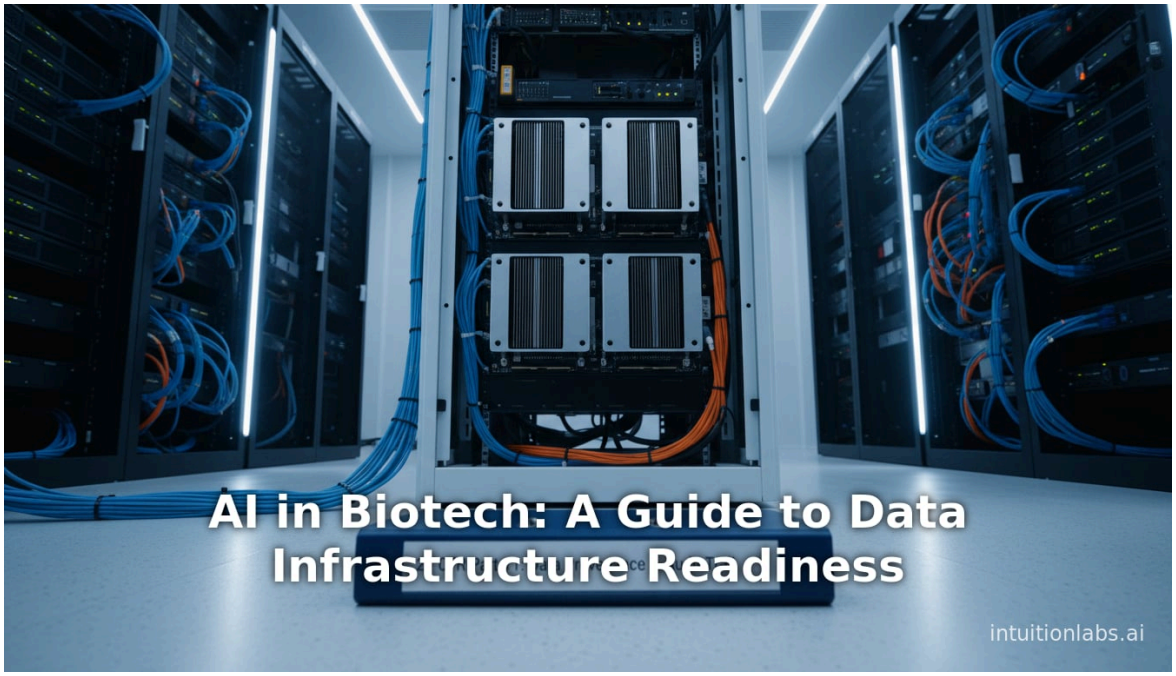
AI in Biotech: A Guide to Data Infrastructure Readiness

By Adrien Laurent, CEO at IntuitionLabs • 12/8/2025 • 30 min read

ai in biotech | data infrastructure | machine learning | fair data | data governance | drug discovery | data quality | data integration



Executive Summary

The convergence of advanced **AI techniques** and **biotechnology** promises transformative breakthroughs in drug discovery, precision medicine, and synthetic biology. However, realizing these benefits requires a solid foundation of data infrastructure. Biotech organizations must **prepare their data systems and processes** before deploying AI – fixing silos, ensuring high-quality and interoperable data, and building scalable computing platforms. This report identifies the key prerequisites and fixes needed to make biotech **AI-ready**, synthesizing insights from industry studies, expert reports, and case examples. We find that:

- **Data Consolidation & Integration**: Biotech data is often **fragmented** across lab instruments, databases, and external sources. Pre-AI fixes include building unified data repositories (lakes or warehouses), applying common schemas and ontologies, and breaking down silos (scientist and department silos) so data can be **queried and merged** ([1] www.scispot.com) ([2] www.pharmalex.com).
- **Data Quality & Curation**: Poor or inconsistent data can derail AI. Over **70–80%** of data scientists' time is spent on cleansing and wrangling data (www.aiuniverse.xyz) ([3] www.scispot.com). Before AI is applied, organizations must implement rigorous ETL pipelines, data validation, and annotation processes. This includes handling missing values, correcting errors, de-duplicating records, and standardizing formats (e.g. gene nomenclature, assay units) ([4] www.bioprocessintl.com) ([5] www.scispot.com).
- **Metadata & Standards**: AI models require **rich metadata** to understand context. Data must follow community standards (e.g. CDISC for trials, FHIR/OMOP for health data, DICOM for images, Gene Ontology for annotations). Ensuring data is **Findable, Accessible, Interoperable, and Reusable (FAIR)** significantly boosts AI readiness ([6] pmc.ncbi.nlm.nih.gov) ([2] www.pharmalex.com). Pre-AI tasks include creating data catalogs, enforcing metadata schemas, and adopting ontologies so that datasets from different sources can integrate seamlessly.
- **Computing & Networking**: Biotech AI workloads (genomic alignment, image analysis, molecular simulation) are **compute-intensive**. Organizations must provision **high-performance hardware** – clusters of GPUs/TPUs, HPC servers with parallel CPUs, and ultra-fast interconnects (InfiniBand or 100–800 GbE) – to avoid bottlenecks ([7] www.whitefiber.com) ([8] aws.amazon.com). Scalable storage (fast NVMe or all-flash pools) with tiering is needed to handle petabytes of experimental data without slowdown ([9] www.whitefiber.com) ([10] aws.amazon.com). Choices between cloud, on-prem, or hybrid deployments must be made in advance based on cost, compliance, and workload stability ([11] www.whitefiber.com) ([12] www.clinicalleader.com).
- **Governance & Compliance**: Biotech data – especially patient and clinical trial records – is highly regulated. Organizations must establish governance frameworks (access controls, encryption, audit trails) to **ensure privacy and meet regulations** (HIPAA, GDPR, FDA 21 CFR Part 11, EMA standards) before AI projects start. For instance, FDA submissions require complete provenance records of how data were generated and processed ([13] www.whitefiber.com) ([14] medium.com).
- **Organizational Readiness**: Beyond tech, companies need the right **skills, processes, and culture**. Data scientists, lab scientists, and IT teams must collaborate. Training in data management and MLOps is essential. Leadership must prioritize data maturity (often a multi-year journey) and shift the organization to think "data-first."

Together, these infrastructure and process prerequisites form a **bedrock** on which AI models can be built and trusted. Clinical and biotech leaders must "fix the plumbing" – invest in data lakes, ETL pipelines, compute clusters, and governance – so that AI can later "flow" effectively through their R&D systems ([1] www.scispot.com) ([7] www.whitefiber.com). As one recent industry analysis notes, achieving "AI readiness" is

fundamentally about aligning data quality and access with advanced algorithms ([6] pmc.ncbi.nlm.nih.gov) ([2] www.pharmalex.com).

Introduction and Background

Biotechnology has entered an era dominated by data. Advances in **high-throughput experiments** (next-generation sequencing, omics platforms, high-content imaging, sensor technologies) have produced **petabytes** of biological data annually ([15] pmc.ncbi.nlm.nih.gov) ([16] www.whitefiber.com). At the same time, **AI and Machine Learning (ML)** have matured into powerful tools for pattern recognition, simulation, and prediction. Applications from DeepMind's AlphaFold (protein structure prediction) to generative chemistry models demonstrate AI's potential to shorten drug discovery timelines by orders of magnitude. However, the true enabler of this revolution is the **data infrastructure** behind the scenes – the hardware, software, and data governance that allow AI algorithms to train on millions of datapoints efficiently and reliably.

Historical context shows that **without strong data foundations**, prior attempts at AI in biotech have faltered. Life sciences traditionally operated in **siloed** and analog systems: lab notebooks, custom databases, and discrete files. Efforts to apply AI often found that data scientists spent the vast majority of their time just making data usable. For example, an analysis of biotech teams noted that workflows were not designed for scale: labs lacked unified repositories and metadata standards, so data integration became a bottleneck. A 2023 survey found that **80% of life science data scientists' time is spent preparing data rather than actual analysis** ([3] www.scispot.com).This suggests that even as algorithms improve, the *data* have yet to "keep up."

Recognizing this gap, initiatives like NIH's **Bridge2AI** program explicitly emphasize data readiness. Bridge2AI argues that having AI-ready datasets (curated, FAIR, ethically collected) is as crucial as the algorithms themselves ([6] pmc.ncbi.nlm.nih.gov) ([17] pubmed.ncbi.nlm.nih.gov). In fact, biomedical data today are so vast and varied – from EHRs to multi-omics – that naively feeding them into algorithms yields limited benefit. The opaque nature of "big models" further demands **explainability and quality** in the input data ([18] pmc.ncbi.nlm.nih.gov) ([19] pmc.ncbi.nlm.nih.gov). NIH-funded experts note that AI systems' performance and fairness are tightly "coupled" with data readiness ([20] pmc.ncbi.nlm.nih.gov).

In sum, the biotech sector stands at a crossroads: **AI offers unprecedented opportunity**, but only if the **data infrastructure** is robust. This report explores *what to fix before deploying AI* in biotechnology. We review the data environment in biotech, identify the challenges and required components of AI-ready systems, and analyze evidence from real-world examples and expert opinions. Our goal is a detailed roadmap enabling biotech organizations to transform raw experimental outputs into AI-driven insights responsibly and efficiently.

The Biotech Data Landscape

Biotech R&D and healthcare generate an extraordinarily **diverse array of data types**. Understanding this variety is essential to building the right infrastructure:

- **Genomic and 'Omics Data**: Modern genomics yields hundreds of gigabytes per human genome before analysis; transcriptomics, proteomics, metabolomics and single-cell assays likewise produce massive datasets. These data can be *sequencing reads*, variant tables, expression matrices, etc., all requiring specialized storage and processing pipelines.
- **Laboratory Instrument Data**: Core wet-lab instruments (e.g. mass spectrometers, chromatography, flow cytometers, high-throughput screening robots) output large structured datasets and often proprietary file formats. These raw data must be captured, labeled, and converted into usable tables.

- **Imaging Data**: Biotech increasingly relies on high-resolution imaging. Examples include fluorescence microscopy (cell images), digital pathology (histology slides), cryo-EM (protein structures), medical imaging (MRI/CT of patients). Images are inherently unstructured, often gigabytes per scan; AI models (especially deep learning) require these to be fed in standardized digital formats (e.g. DICOM, TIFF) and annotated carefully.
- **Clinical and Patient Data**: Drug development and personalized medicine rely on clinical trials and patient records. These include **EHR data** (demographics, lab results, clinical notes), as well as registries and claims data. Much of this is semi-structured or textual (doctor's notes, free-text reports), requiring natural language processing to integrate. This data is also highly sensitive and regulated, necessitating strong privacy infrastructure.
- **Operational Data**: Information from lab management systems (study metadata, reagent inventories, workflow logs), manufacturing (batch records, QC results), and supply chains (cold chain sensors) falls here. Such data often reside in enterprise software (LIMS/ELN/WMS) and must be interoperable with research data.
- **Real-World and IoT Data**: Biotech and digital health generate an increasing flow from wearables, smartphone apps, and Internet-of-Things sensors (e.g. continuous glucose monitors, environmental monitors for bioreactors). These produce time-series data at high velocity that can enrich research.

All of the above contribute to the **5 V's of big data** in life sciences – volume, velocity, variety, veracity, and value. Importantly, most of these datasets are **high-dimensional** and often noisy. For example, Bridge2AI notes that healthcare data (EHRs) are "increasingly computable" but still plagued by missingness and heterogeneity ([21] pmc.ncbi.nlm.nih.gov). Similarly, high-content screening experiments can capture terabytes of cell images per study, but with variation (batch effects) between plates.

Data silos are pervasive. Academic labs, CROs, and companies each maintain separate databases. Even within one organization, preclinical data may be in one system, clinical data in another, and vendor data in yet another. Interfacing across these systems is typically ad hoc. A data scientist working on biomarker discovery might spend weeks just **extracting and aligning data** from different sources. As one industry report emphasizes, many biotech and pharma have "never been good at sharing data," so crucial predictive models are held back by fractured datasets ([22] www.datamanagementblog.com).

Compounding the challenge is **data quality**. Raw biomedical data is often "dirty": incomplete patient records, mislabeled samples, duplicate records, and inconsistent units are routine. BioProcess International warns that scaling to large datasets increases the risk of "noise, missing values, outliers, lack of balance, inconsistency, redundancy" – all of which can make AI models fail ([4] www.bioprocessintl.com). Without intervention, these defects become "bottlenecks" that require expensive manual lab work or ad hoc scripts ([23] www.bioprocessintl.com).

Finally, **standardization is lacking**. Compare how far biology lags fields like finance: there is no universally adopted data schema in biotech. Multiple coding systems (e.g. ICD vs. SNOMED for disease; various proprietary chemistry identifiers) co-exist. This demands translation layers or master data management. Fortunately, pockets of standardization exist (meta-analysis initiatives, UniProt IDs, ontologies), but they must be proactively enforced during integration.

In summary, the biotech data landscape is rich but chaotic. Any AI initiative must operate on *well-governed*, **integrated** data. Table 1 below summarizes some key categories of data infrastructure needed to tame this complexity.

Infrastructure Component	Purpose/Role	Key Requirements
Data Storage (Lake/Warehouse)	Central repository for raw and processed data (genomic sequences, images, trial results)	*Scalability:* petabyte+ capacity *Performance:* high-throughput (flash/NVMe) for analytics *Security:* encryption, access control *Metadata support:* ensure data is annotated and traceable ([9] www.whitefiber.com) ([10] aws.amazon.com)
Compute (HPC/Cloud)	Run AI model training, simulations (e.g. molecular dynamics, genomic alignments, deep learning on images)	*High Performance:* GPU clusters (e.g. NVIDIA H100/A100), parallel CPUs (multi-core servers) ([24] www.whitefiber.com) *Elasticity:* ability to scale (cluster or cloud bursts) ([8] aws.amazon.com) *Specialized Hardware:* TPUs, FPGAs or domain-specific ASICs as needed
Network (High-Speed Fabric)	Transfer large datasets between storage and compute nodes, and between on-prem and cloud	*Bandwidth:* 100–800 Gbps Ethernet or InfiniBand for multi-node training ([7] www.whitefiber.com) *Low Latency:* to synchronize GPUs during distributed training *Secure Transfer:* VPNs/SSH, as needed for protected data
Data Ingestion & Pipelines	Automate the flow of data from labs and external sources into the central repository	*ETL/Streaming Tools:* (e.g. Kafka, NiFi, Airflow) *Format Conversion:* e.g. raw instrument outputs to uniform schemas *Error Handling:* validation, retries, and alerts on failed transfers ([25] www.scispot.com)
Data Catalog & Metadata	Indexing of datasets, variables, and their descriptions (essential for discovery and interoperability)	*Findability:* searchable catalogs (ELT, ELN integration) *Standards:* common ontologies and controlled vocabularies *Versioning:* track dataset versions and lineage (supports reproducibility) ([26] pmc.ncbi.nlm.nih.gov) ([13] www.whitefiber.com)
Security & Compliance	Protect sensitive information and ensure legal/regulatory adherence	*Access Control:* role-based security, auditing (21 CFR Part 11 compliance) ([13] www.whitefiber.com) *Data Privacy:* de-identification, encryption at rest/in transit *Governance Processes:* defined data stewardship and policies
Platforms & Tooling	Support for collaborative development and deployment (MLOps, containers, APIs)	*Containerization:* Docker/Kubernetes for reproducible environments *MLOps Tools:* model registries, CI/CD pipelines for ML *Interoperability:* APIs to connect LIMS/ELN with analytics platforms

Data Integration and Quality: Fixing the Foundations

Before engaging AI models, biotech organizations must **consolidate and curate** their data. Even the best algorithms cannot overcome fundamentally poor inputs. Key preparatory steps include:

- **Unify Siloed Data**. Data currently trapped in disparate systems (e.g. separate lab databases, clinical systems, partner data) must be brought together. This typically means implementing a central **data lake or warehouse** that can ingest diverse inputs ([27] www.scispot.com) ([2] www.pharmalex.com). Data sources should be rationalized: each dataset should be mapped to a common scheme or entity (e.g. aligning chemicals with PubChem IDs, patients with unique IDs). Organizations often establish ETL pipelines to pull in laboratory results, instrument logs, and clinical records on schedule, adding metadata as they go ([1] www.scispot.com) ([2] www.pharmalex.com). A unified platform then enables queries across previously disconnected silos, supporting both exploratory analytics and machine learning.

- **Standardize and Annotate**. All incoming data should follow agreed formats and nomenclature. For example, ensure genomic data uses a consistent reference build, clinical terms match a chosen ontology, and all files carry timestamp and context metadata. Table lookup or master data management (MDM) systems can harmonize synonyms (e.g. drug names, gene symbols) across datasets ([28] www.pharmalex.com) ([25] www.scispot.com). Crucially, each record should be annotated with provenance (where/when it came from) and quality flags. NIH's Bridge2AI stresses the importance of **comprehensive metadata** to enable dataset reuse ([6] pmc.ncbi.nlm.nih.gov). Tagging data along with its processing history ensures AI models later can interpret it appropriately.

- **Clean and Curate**. Data quality assurance is mandatory. This includes removing or imputing missing values, resolving duplicates, and correcting erroneous entries. Statistical checks and domain rules should catch outliers or biologically implausible values. For imaging and sequence data, additional preprocessing (denoising, normalization, segmentation) is often needed. As BioProcess International notes, "dirty data" for AI can stem from incomplete fields, inconsistent units, or outdated codes ([23] www.bioprocessintl.com). Modern biotech must invest in automated **data cleaning pipelines** – essentially domain-specific "digital janitors" – so that downstream AI models train on reliable inputs. According to surveys, addressing these issues is **urgent**: "data wrangling still takes the lion's share of time" for scientists (www.aiuniverse.xyz) ([3] www.scispot.com).

- **Implement Quality Control Metrics**. Track and report on data readiness. For example, measure the proportion of records with complete fields, the number of records rejected by validation rules, or the degree of standardization achieved. These metrics can become part of the data governance process. Many organizations define an "analytics maturity" or "data readiness" assessment before an AI rollout. If data scores poorly (e.g. < some threshold of completeness or consistency), AI projects are delayed until remediation. This disciplined approach – of setting baseline data quality requirements – helps prevent wasted effort on model training that later fails due to data flaws.

- **Adopt Interoperability Standards**. Wherever possible, use community or industry standards to store and label data. In genomics, the file formats FASTQ, BAM/CRAM, and VCF are almost ubiquitous. For imaging, the DICOM standard dominates radiology, and OME-TIFF is common in microscopy. In healthcare, modern EHRs and observatories leverage HL7 FHIR resources or the OMOP common data model for trials. Encoding data according to standards (or converting legacy data into them) greatly reduces custom parsing code. It also facilitates data sharing – for example, applying for access to the NIH data commons or contributing to public datasets. Embedding standards is a prerequisite: according to a pharma data integration guide, supporting **"big data"** often means handling images or continuous sensors by adopting proper platforms that speak standard vocabularies ([2] www.pharmalex.com).

- **Ensure FAIR Principles**. Bridge2AI explicitly links **AI-readiness** to FAIR data practices ([6] pmc.ncbi.nlm.nih.gov) ([17] pubmed.ncbi.nlm.nih.gov). For each dataset: can it be *found* via search? *Access* is controlled but possible for valid use cases? *Interoperable* with other datasets (shared vocabularies)? *Reusable* with clear licenses and documentation? A potential approach is to intern each dataset with a FAIRness rubric. This might involve registering data in catalogues (giving persistent IDs), attaching rich metadata, and defining data use agreements. Organizations should consider adopting FAIR-checklists during data collection to ensure compliance from the outset.

- **Governance and Security Controls**. At the same time, data integration must respect privacy and IP. Establish clear governance frameworks: who can access which data, under what conditions, and for what purposes. For sensitive clinical or patient data, implement **de-identification and encryption**. Develop audit logging so that every data access or modification is recorded (important for FDA/EMA inspections). For AI projects, ensure an ethics review or transparency board considers algorithmic fairness and bias from the collected data. Overall, fix the **governance "plumbing"** now – it's much harder (or impossible) to retrofit strong governance after analytics pipelines are in place.

AI-Ready Computing Infrastructure

With data consolidated and cleaned, the next prerequisite is **adequate computing resources**. Many biotech problems – from molecular dynamics to training deep neural networks – require **high-performance infrastructure**. Before deploying AI, organizations should assess and provision:

- **High-Performance Compute (HPC) Clusters**: On-premises clusters equipped with hundreds or thousands of CPU cores (and GPUs) may be needed for heavy workloads. As the industry's largest example, Recursion Pharmaceuticals built an internal supercomputer ("BioHive-1") that, with its NVIDIA A100 GPUs, ranked in the *Top 10* of the global TOP500 supercomputers in late 2023 ([29] www.genengnews.com). (Recursion plans to expand it with 800+ NVIDIA H100 GPUs, making it possibly the most powerful biopharma supercomputer ([30] www.genengnews.com).) For smaller organizations, building even a mini-HPC can be transformative: genomics assembly tasks that once took days can be completed in hours when run on many-node clusters ([8] aws.amazon.com). The key fix is to ensure **horizontal scaling**: architecture that allows adding compute nodes or GPUs easily, either behind the company firewall or in the cloud.

- **GPUs and Specialized Accelerators**: GPUs (Graphics Processing Units) are standard for deep learning. Modern biotech AI often demands dozens to hundreds of top-end GPUs (e.g. NVIDIA's A100, H100, or similar). These accelerators can run neural nets and molecular simulations much faster than CPUs alone. In the laboratories, demands like **protein-ligand docking simulations** or **large-scale genomic comparisons** produce billions of calculations – workloads GPUs handle well ([24] www.whitefiber.com). Emerging domain-specific chips (e.g. Google's TPUs, Graphcore's IPUs) may also be integrated. The infrastructure fix: **budget for GPUs early** and design compute clusters around them, with sufficient power and cooling. (For example, an established pharma might outfit R&D racks with liquid-cooled GPU servers, since these become strategic assets ([31] www.whitefiber.com).)

- **High-Speed Networking**: To keep up with distributed computation, data must flow rapidly between storage and compute. Without robust networking, even well-provisioned clusters can become idle. In practice, biotech compute nodes should be connected via low-latency, high-bandwidth fabrics. Common solutions are modern InfiniBand or RoCE (RDMA over Converged Ethernet) interconnects, which can provide **terabit-level throughput** between GPUs during training ([7] www.whitefiber.com). Within data centers, 100–800 Gbps Ethernet may be used to move large genomic or imaging files. The key is to **prevent bottlenecks**: an AI training job that theoretically takes 2 days on GPUs might take weeks if network file reads are slow. Thus, ensuring high-speed networking is a prerequisite fix.

- **Scalable Storage Systems**: The storage architecture must balance performance (for active data) with capacity and cost (for archives). Biotech R&D generates petabytes of data – consider that Recursion's in-house robotics produced **2 petabytes of bioimaging data** ([32] cloudgate.healthcareitnews.com). fixes include deploying parallel file systems (e.g. Lustre, WEKA, parallel NFS) for active projects. These systems handle high I/O demands (such as reading genomic data in parallel). In addition, tiered storage is vital: keep "hot" data (current trial results, model checkpoints) on fast SSD or all-flash arrays, while archiving older experiments on slower, cheaper disk or cloud cold storage ([33] www.whitefiber.com). Proper tiering cuts costs – Ginkgo Bioworks reports saving **60–90%** on storage bills by archiving downtime cycles ([10] aws.amazon.com). And any storage must be integrated with backup and redundancy, so that critical datasets cannot be accidentally lost pre-AI.

- **Cloud vs On-Premises Debate**: Many emerging biotech teams begin with the cloud for AI experiments. Cloud platforms (AWS, GCP, Azure) offer **on-demand elasticity**: spin up dozens of GPUs for a week, then spin them down. This avoids large upfront capital expenses. For instance, Ginkgo moved significant workloads to AWS, enabling rapid scaling and achieving dramatic speedups (e.g. **10× faster** genome assembly) ([8] aws.amazon.com). However, long-term cloud use can be costly (egress fees, storage costs) and complex for compliance. As organizations scale, a **hybrid approach** often emerges: maintain a private HPC for steady workloads and sensitive data, while bursting to cloud for peaks (e.g. a drug-screen campaign). The fix here is to plan the right model from the outset. If using cloud, one must negotiate data governance and regulatory certifications in advance. If building on-prem, ensure teams have training in cluster management and budget for hardware refresh cycles. The clinical leaders note that the future of pharma R&D is a hybrid cloud/HPC continuum ([12] www.clinicalleader.com) ([11] www.whitefiber.com).

The figure below (Table 2) summarizes common data challenges and the essential remedies needed before AI can be effectively deployed.

Challenge	Effect on AI	Pre-AI Remediation
Data Silos / Fragmentation	Models cannot access all relevant information; insights incomplete.	Integrate datasets into unified repositories (data lakehouses); adopt common schemas; use federated queries to break silos ([1] www.scispot.com) ([2] www.pharmalex.com).
Poor Data Quality	Garbage-in→garbage-out. Predictions unreliable.	Implement rigorous cleaning/validation pipelines: remove outliers, fill missing values, standardize units; use domain QC rules ([4] www.bioprocessintl.com) (www.aiuniverse.xyz).
Heterogeneous Formats	Parsing and integration errors; time wasted on conversion.	Normalize data formats (e.g. convert all imaging to DICOM, all sequences to FASTQ/VCF); use transformation pipelines for legacy data ([2] www.pharmalex.com) ([13] www.whitefiber.com).
Lack of Metadata	Data not discoverable or reusable; ML context missing.	Enforce metadata tooling: require annotations, timestamps, experimental conditions; build data catalogs for indexing.
Regulatory/Privacy Constraints	Limits data sharing; risk of non-compliance.	Establish governance: de-identify patient data, implement access controls and audit logs, ensure encryption to comply with HIPAA/GDPR/FDA rules ([13] www.whitefiber.com) ([14] medium.com).
Compute & Storage Shortages	Training pipelines fail or take too long; hampered innovation.	Deploy scalable GPU/HPC resources (on-prem or cloud); ensure high-throughput storage and networking to eliminate bottlenecks ([7] www.whitefiber.com) ([8] aws.amazon.com).
Skill Gaps	Inefficient model development; misused tools	Invest in training data scientists and bioinformatics staff; hire or partner for engineering expertise; cultivate cross-functional teams.

These tables and discussions highlight that **data readiness is multifaceted**. Biotech leaders must fix *both* the technology (software/hardware) and the processes (governance/operations) to make AI feasible. Only then can statistical and ML tools be applied to yield meaningful biological insights.

Case Studies and Real-World Examples

Several pioneering biotech and pharma organizations illustrate the value of fixing infrastructure first:

- **Recursion Pharmaceuticals (Phenomic AI)**: Recursion is emblematic of an "AI-first" biotech. The company runs millions of cell biology experiments on robotic microscopes, generating an enormous image archive. In 2019, Recursion open-sourced 100,000 cell images (the RxRx1 dataset, ~300 GB) to spur algorithm development ([34] cloudgate.healthcareitnews.com). Yet this was tiny compared to Recursion's own lab output: CEO Gibson noted that 100k images represented only 0.4% of what they produce *weekly* ([35] cloudgate.healthcareitnews.com). To handle such volume, Recursion built "BioHive-1", an in-house supercomputer with hundreds of NVIDIA GPUs. As of late 2023, BioHive-1 (with 300 NVIDIA A100 GPUs) was ranked **#9** on the Top500 supercomputer list ([29] www.genengnews.com). Recursion further committed a $1 billion investment in their Recursion OS (data infrastructure) and partnered with NVIDIA (adding 500 H100 GPUs) to reach the top 50 by 2024 ([29] www.genengnews.com) ([36] www.genengnews.com). They even publicize that scaling up model training on their open RxRx3 dataset (2.2 million cell images) increased model performance: their Phenom-Beta foundation model continues to improve "as we increased the size of the training data" ([37] www.genengnews.com). In sum, Recursion's case shows that massive scale and specialized infrastructure (both for data and compute) were prerequisites to using AI effectively. They eliminated data pipeline bottlenecks (automated pipelines handling millions of images per week) so scientists could focus on model training ([25] www.scispot.com) ([29] www.genengnews.com).

- **Ginkgo Bioworks (Synthetic Biology)**: Ginkgo operates high-throughput biofoundries, executing thousands of experiments annually. In 2019, Ginkgo migrated most R&D workloads to Amazon Web Services to achieve elastic compute. According to an AWS case study, this move drastically cut time for key tasks: **genome assembly** jobs fell from 40 hours on-prem to 4 hours in the cloud ([8] aws.amazon.com). They also handled petabytes of data storage by using tiered solutions: active experiments on fast EBS SSDs, long-term cold data in S3. With careful architecture, Ginkgo reports *"we've saved up to 90%"* on EBS cost and 60% on S3 storage through volume optimization ([10] aws.amazon.com). The company emphasizes that cloud freed scientists to do "rapid experimentation" rather than maintain hardware ([38] aws.amazon.com). Ginkgo's experience underscores the infrastructure payoff: by fixing their compute/storage platform (cloud batch computing, intelligent storage), they achieve orders-of-magnitude faster analysis while maintaining compliance ("we could not have accommodated … needs without migrating to AWS" ([39] aws.amazon.com)).

- **UK Biobank Research Analysis Platform (UKB-RAP)**: UK Biobank provides a secure environment for analyzing its 500,000-patient biomedical dataset. On the data side, UK Biobank has harmonized clinical records, genomics, and extensive imaging (brain, heart, etc.), including preprocessed "Image Derived Phenotypes." To harness AI, they partnered with DNAnexus to create the UKB-RAP cloud platform ([14] medium.com). This brings compute to the data: vetted researchers can run JupyterLab notebooks and pipelines in a secure enclave, without the data ever leaving the protected environment ([14] medium.com). This infrastructure satisfied compliance and enabled large-scale AI studies (for example, training ML models on 50,000 brain MRIs to predict age ([40] medium.com)). The key takeaway is that by building a robust cloud data platform up front, UK Biobank *enabled* many AI projects. If instead they had only offered raw data downloads, few would have the capacity to exploit it. The "fix" here was creating a unified, governed analysis ecosystem.

- **Pharma Industry Shifts (Novo Nordisk and Eli Lilly)**: Recent announcements show Big Pharma taking AI infrastructure seriously. In 2025, Novo Nordisk partnered with NVIDIA and a Danish AI center to commission *"Gefion"*, a sovereign supercomputer (DGX SuperPOD) to run generative and agentic AI models for early research ([41] www.clinicalleader.com). Soon after, Eli Lilly revealed an in-house **"AI Factory"**, also DGX-based, to handle end-to-end AI workflows (data ingestion, training, inference) with federated learning capabilities ([42] www.clinicalleader.com). These moves mark a clear end-of-life for ad-hoc compute: industry leaders now see dedicated AI/HPC systems as as integral as chromatography or mass spec used to be. The implication is that **mainstream drug development** will soon require fusion of on-prem supercomputing and hybrid cloud. Indeed, a recent analysis notes that while these HPC units deliver raw ML power, cloud platforms will handle clinical operations (trial analytics, patient monitoring, etc.) ([43] www.clinicalleader.com). From a data-prep perspective, these partnerships highlight that legacy infrastructure cannot keep pace: to "move the needle," R&D groups must invest in specialized platforms and mix of cloud services before AI can drive value.

- **Novartis (AI in Clinical Trials)**: As an example of data-driven trials, Novartis deployed AI to monitor ongoing trial data streams in real time. This allowed them to detect anomalies and site issues earlier, **minimizing trial delays** and streamlining regulatory submissions ([44] www.scispot.com). Novartis credits this success to first structuring and harmonizing all incoming trial data (from eCRFs, monitoring tools) so that the AI could query it. Their lesson: even in operations, cleaning and integrating data was a necessary precursor to applying machine learning effectively.

These case studies converge on a theme: **invest first in data and compute infrastructure, then apply AI**. Ginkgo and Recursion tackled scale by upgrading compute clusters and storage; biobanks and big pharma tackled integration by building unified platforms; leading companies tackled compliance by baking security into the design. Only after these prerequisites were met did AI yield significant gains (chartered 10× speedups, multi-million avoidance of delays, etc.).

Implications and Future Directions

With solid data infrastructure in place, biotech organizations can fully explore AI's potential. Conversely, failure to address these prerequisites risks wasted investment: one industry analyst warns that without data readiness, **even advanced GenAI will underperform** ([45] www.datamanagementblog.com) ([46] www.pharmalex.com). Looking ahead:

- **Democratized Data Ecosystems**: The trend is toward federated data networks. Companies and consortia (UK Biobank, All of Us, EU cloud national initiatives) are expanding research platforms that allow scalable AI while protecting privacy. Biotech R&D may see more **data commons**, where curated datasets and tools are co-located (e.g. genomic databases with cloud notebooks). Participating in such ecosystems will require biotech firms to align their formats and standards with global initiatives (as NIH recommends ([47] pmc.ncbi.nlm.nih.gov)).
- **Regulatory Evolution**: As AI becomes standard, regulators are tightening guidelines on data quality and model transparency. For instance, FDA's Digital Health Plan hints at needing validated data pipelines for AI algorithms in medical devices. Biotech companies will need to maintain auditable data records from experiment to outcome. Thus, the work of establishing audit trails and validation (fixes made pre-AI) will soon become a regulatory mandate rather than optional.
- **Emphasis on Trust and Ethics**: Effective AI models in healthcare must be interpretable and fair. This means that even with high-tech infrastructure, biotech must invest in **explainability tools**, bias audits, and ethical review boards. Data governance (privacy, consent) also has societal impact. NIH's Bridge2AI underscores that AI-readiness includes **ethical and social criteria** in data use ([20] pmc.ncbi.nlm.nih.gov). Companies should prepare to share how data were collected and consented, and how patient rights are protected through AI deployments.
- **Workforce Transformation**: The skills shortage in "AI-Ready biotech" is palpable. Industry groups (e.g. the UK BioIndustry Association) are already recommending new curricula to train data-savvy scientists ([48] www.bioindustry.org). The data infrastructure effort will require hiring data engineers, bioinformaticians, and AI specialists. Many current life science professionals will need upskilling in data management and ML.
- **Continuous Data Pipeline Improvements**: Building an AI-ready infrastructure is not a one-off task. It demands ongoing maintenance: new data sources (novel assays, real-world studies) must be integrated; hardware must be updated; models must be retrained as new data arrive. Adopting DevOps/MLOps practices (infrastructure-as-code, automated testing) – as recommended in recent reviews – can accelerate these continuous improvements ([17] pubmed.ncbi.nlm.nih.gov). Teams should embed data infrastructure in their agile workflows so that as projects evolve, the pipelines evolve with them.
- **Competitive Advantage**: Finally, a robust data infrastructure can itself be a strategic asset. Accelerating R&D with AI leads to faster drug development and cost savings (e.g. minimizing late-stage failures). For example, one summit noted that generative AI lets scientists explore thousands of new compounds per day, "making failure cheap and easy at early stages" ([49] www.genengnews.com). But this advantage accrues only if data pipelines feed those models in real time. In a silver lining, companies that do fix their infrastructure early will both save costs and enable new innovations (digital twins of bioprocesses, adaptive clinical trials, etc.) long before their less-prepared competitors.

Conclusion

Building **AI-ready biotech** infrastructure is a complex, multi-dimensional challenge. It is *not* enough to purchase AI software or hire a data scientist – the entire *foundation* of data management and computing must be solid. Specifically, biotech organizations should fix the following *before* rolling out AI projects:

- Consolidate and standardize data in unified, interoperable repositories (data lakes/warehouses) ([1] www.scispot.com) ([2] www.pharmalex.com).
- Automate robust data pipelines (ETL) and metadata capture to minimize manual preprocessing ([25] www.scispot.com) ([6] pmc.ncbi.nlm.nih.gov).
- Implement rigorous data quality controls and annotation to ensure training data are clean and meaningful (www.aiuniverse.xyz) ([4] www.bioprocessintl.com).
- Invest in scalable compute (GPU/HPC clusters) and networking so that model training times are practical ([7] www.whitefiber.com) ([8] aws.amazon.com).
- Embed security, privacy, and audit requirements into the system design, not as an afterthought ([13] www.whitefiber.com) ([14] medium.com).
- Foster an organizational culture that prizes data stewardship, analytics literacy, and interdisciplinary collaboration.

By addressing these prerequisites, biotech firms position themselves to **fully leverage AI**. The proof is emerging: companies that have "fixed their data plumbing" report dramatically faster experiments, higher model accuracy, and new discovery capabilities ([8] aws.amazon.com) ([41] www.clinicalleader.com). The alternative – jumping straight to sophisticated AI methods on fragmented data – risks wasted resources and missed opportunities.

In conclusion, **AI readiness in biotech is built, not bought**. A decade after AI began transforming other industries, the life sciences must undergo a modernization of its digital infrastructure. This report has outlined the critical fixes needed. As we have shown through data and case studies, those fixes – from data lakes to GPU farms – translate into **faster science, lower costs, and ultimately better cures**. Biotech organizations aiming for AI-driven innovation must therefore prioritize these data infrastructure prerequisites today to reap the benefits tomorrow.

External Sources

[1] https://www.scispot.com/blog/the-role-of-data-infrastructure-in-enabling-ai-driven-biotech-companies#:~:betwe...

[2] https://www.pharmalex.com/thought-leadership/blogs/beyond-plumbing-the-strategic-role-of-data-integration-in-pharma-ai/#:~:,with...

[3] https://www.scispot.com/blog/scalable-data-infrastructure-ai-playbook-for-biotech-startups-and-scaleups#:~:image...

[4] https://www.bioprocessintl.com/information-technology/working-with-big-data-in-healthcare-and-bioprocessing-settings-a-brief-introduction-to-key-components-and-considerations#:~:Model...

[5] https://www.scispot.com/blog/scalable-data-infrastructure-ai-playbook-for-biotech-startups-and-scaleups#:~:chall...

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:Avail...

[7] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:Infin...

[8] https://aws.amazon.com/solutions/case-studies/ginkgo-bioworks-case-study/#:~:In%20...

[9] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:High,...

[10] https://aws.amazon.com/solutions/case-studies/ginkgo-bioworks-case-study/#:~:Along...

[11] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:Cloud...

[12] https://www.clinicalleader.com/doc/from-supercomputers-to-the-cloud-how-pharma-s-r-d-infrastructure-transformation-impacts-clinical-trials-0001#:~:Toget...

[13] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:flash...

[14] https://medium.com/dnanexus/infrastructure-challenges-machine-learning-in-the-cloud-as-a-future-oriented-solution-using-uk-6e10bedb995a#:~:Anoth...

[15] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:Artif...

[16] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:all%2...

[17] https://pubmed.ncbi.nlm.nih.gov/35500446/#:~:Recen...

[18] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:Biome...

[19] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:An%20...

[20] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:Appli...

[21] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:biome...

[22] https://www.datamanagementblog.com/from-data-chaos-to-ai-powered-innovation-how-life-sciences-can-break-free-from-siloed-insights/#:~:Consi...

[23] https://www.bioprocessintl.com/information-technology/working-with-big-data-in-healthcare-and-bioprocessing-settings-a-brief-introduction-to-key-components-and-considerations#:~:Dirty...

[24] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:GPU%2...

[25] https://www.scispot.com/blog/scalable-data-infrastructure-ai-playbook-for-biotech-startups-and-scaleups#:~:Autom...

[26] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:that%...

[27] https://www.scispot.com/blog/the-role-of-data-infrastructure-in-enabling-ai-driven-biotech-companies#:~:Data%...

[28] https://www.pharmalex.com/thought-leadership/blogs/beyond-plumbing-the-strategic-role-of-data-integration-in-pharma-ai/#:~:%2A%2...

[29] https://www.genengnews.com/topics/drug-discovery/drugs-dollars-and-data-recursion-eyes-cost-savings-from-ai-drug-discovery/#:~:Much%...

[30] https://www.genengnews.com/topics/drug-discovery/drugs-dollars-and-data-recursion-eyes-cost-savings-from-ai-drug-discovery/#:~:Also%...

[31] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:,hous...

[32] https://cloudgate.healthcareitnews.com/news/huge-dataset-biological-images-made-available-spur-new-ai-algorithms#:~:The%2...

[33] https://www.whitefiber.com/blog/beginners-guide-to-ai-infrastructure-for-biotech#:~:Tiere...

[34] https://cloudgate.healthcareitnews.com/news/huge-dataset-biological-images-made-available-spur-new-ai-algorithms#:~:The%2...

[35] https://cloudgate.healthcareitnews.com/news/huge-dataset-biological-images-made-available-spur-new-ai-algorithms#:~:,biol...

[36] https://www.genengnews.com/topics/drug-discovery/drugs-dollars-and-data-recursion-eyes-cost-savings-from-ai-drug-discovery/#:~:NVIDI...

[37] https://www.genengnews.com/topics/drug-discovery/drugs-dollars-and-data-recursion-eyes-cost-savings-from-ai-drug-discovery/#:~:9%20o...

[38] https://aws.amazon.com/solutions/case-studies/ginkgo-bioworks-case-study/#:~:With%...

[39] https://aws.amazon.com/solutions/case-studies/ginkgo-bioworks-case-study/#:~:meet%...

[40] https://medium.com/dnanexus/infrastructure-challenges-machine-learning-in-the-cloud-as-a-future-oriented-solution-using-uk-6e10bedb995a#:~:To%20...

[41] https://www.clinicalleader.com/doc/from-supercomputers-to-the-cloud-how-pharma-s-r-d-infrastructure-transformation-impacts-clinical-trials-0001#:~:On%20...

[42] https://www.clinicalleader.com/doc/from-supercomputers-to-the-cloud-how-pharma-s-r-d-infrastructure-transformation-impacts-clinical-trials-0001#:~:Just%...

[43] https://www.clinicalleader.com/doc/from-supercomputers-to-the-cloud-how-pharma-s-r-d-infrastructure-transformation-impacts-clinical-trials-0001#:~:,prem...

[44] https://www.scispot.com/blog/scalable-data-infrastructure-ai-playbook-for-biotech-startups-and-scaleups#:~:Pharm...

[45] https://www.datamanagementblog.com/from-data-chaos-to-ai-powered-innovation-how-life-sciences-can-break-free-from-siloed-insights/#:~:But%2...

[46] https://www.pharmalex.com/thought-leadership/blogs/beyond-plumbing-the-strategic-role-of-data-integration-in-pharma-ai/#:~:AI%20...

[47] https://pmc.ncbi.nlm.nih.gov/articles/PMC11526931/#:~:scien...

[48] https://www.bioindustry.org/resource/building-an-ai-ready-biotech-workforce-what-new-curriculum-and-ai-skills-evidence-mean-for-the-sector.html#:~:AI,ab...

[49] https://www.genengnews.com/topics/drug-discovery/drugs-dollars-and-data-recursion-eyes-cost-savings-from-ai-drug-discovery/#:~:%E2%8...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.