AI Compute Demand in Biotech: 2025 Report & Statistics

By Adrien Laurent, CEO at IntuitionLabs • 10/24/2025 • 50 min read

ai compute demand | biotechnology | drug discovery | genomics | ai infrastructure | gpu computing | hpc | alphafold | bioinformatics

Executive Summary

The biotechnology industry is undergoing a rapid transformation driven by artificial intelligence (AI) and machine learning (ML). This evolution is significantly increasing **computational demand** within biotech research and development. Companies are deploying AI for genomics, proteomics, drug discovery, and biological engineering, which requires massive compute power – far beyond traditional bioinformatics workloads. Recent reports indicate that AI compute demand is **"rapidly outpacing the supply of necessary infrastructure"** ([1] www.tomshardware.com). Industry forecasts underscore this surge: Bain & Company projects that to sustain current growth, AI data centers worldwide may require **200 gigawatts of power by 2030** ([2] www.tomshardware.com), and Citigroup estimates **$2.8 trillion** in AI-related infrastructure spending by 2029 ([3] www.reuters.com). These demands are already stressing global power grids; for example, U.S. data-center electricity use is expected to nearly triple by 2028 ([4] www.reuters.com). In biotech specifically, AI-driven projects such as DeepMind's AlphaFold (which predicted nearly every known protein structure ([5] time.com)) and numerous startup-industry collaborations (e.g. Recursion's $688M acquisition of Exscientia ([6] www.reuters.com), AstraZeneca's $555M Algen deal ([7] www.reuters.com), Lilly's AI platform TuneLab built on $1B+ of research ([8] www.reuters.com)) exemplify the trend. We detail multiple perspectives on this phenomenon, including technological, economic, and policy angles: **data analysis** shows exponential AI compute growth; **case studies** highlight major biotech-AI partnerships; and **future directions** discuss energy, infrastructure, and emerging technologies (e.g. quantum computing) that will shape biotech's compute landscape.

Introduction

Artificial intelligence (AI) has become integral to biotechnology, encompassing drug discovery, genomics, proteomics, and synthetic biology. Unlike traditional computational biology (which relied on algorithms for sequence alignment or molecular docking), modern biotech uses **deep learning** and large-scale AI models to analyze complex data. This shift is dramatically increasing *compute demand*. The root causes include (1) vast growth in biological data (genomic sequences, proteomics, patient data) and (2) increasingly complex AI models (deep neural networks requiring GPUs/TPUs). For instance, DeepMind's AlphaFold 2 (2020) achieved a Nobel-winning breakthrough by predicting structures for *nearly every known protein* ([5] time.com), and its successor AlphaFold 3 (2024) extends prediction to protein-DNA interactions ([9] www.reuters.com) ([10] time.com). Training and running such models requires extensive computation: the AlphaFold project alone entailed *weeks of GPU computation* for each prediction pipeline. Similarly, biotech startups now use generative models to design proteins and small molecules, which involves high-dimensional optimization and simulation (e.g. See "Experimental Results" below). This report examines how these advanced AI tasks are driving unprecedented demand for computing resources in biotech.

To frame the discussion, we first review the historical context: biotech's computational roots were in high-performance computing (HPC) for genomics (e.g. processing the Human Genome Project data with supercomputers) and in classic bioinformatics. Moore's Law gains in CPUs had sufficed for years, but since around 2012, the bottleneck has shifted as AI models consume GPU resources at an exponential rate. We then detail the **current state** (late 2025): industry trends, infrastructure investments, and examples of AI-powered biotech projects. We analyze statistics and research findings on compute consumption (e.g. power usage, chip revenues), and present case studies of major biotech–AI initiatives. Finally, we discuss future implications — from energy and scalability issues to emerging technologies like quantum computing — and conclude with projections and recommendations.

Background: Biotech and Computing Evolution

Biotech R&D has always been computation-intensive, but its focus has evolved. In the 1990s–2000s, bioinformatics tasks (genome sequencing, molecular simulations) ran on large CPU-based clusters. Around 2010 onwards, specialized HPC centers (government and academic) supported life sciences. For example, DOE and NIH have funded national HPC facilities for genomics and climate research. In 2025, the U.S. Department of Energy announced "Doudna," a new supercomputer named for Nobel laureate Jennifer Doudna, set to support AI and genomics research at Berkeley Lab ([11] apnews.com). Similarly, Europe and countries like the UK are investing in HPC: Britain's government launched a £1 billion program in 2025 to increase compute capacity twenty-fold by 2030 ([12] www.reuters.com). As part of that initiative, the Isambard-AI supercomputer (5,448 Nvidia GH200 GPUs, 21 exaflops of AI performance) began operations, targeting climate science, drug discovery, and healthcare models ([13] www.itpro.com). These infrastructure developments reflect biotech's heightened compute needs: tasks that once required months on CPU clusters can now often be done in days on GPU-accelerators.

At the same time, the biotech industry's scale and complexity have grown. Biotech market value (pharmaceuticals, synthetic biology, etc.) reached hundreds of billions globally by 2022 ([14] zipdo.co) ([15] wifitalents.com). Many biotech firms increasingly leverage AI to cut costs and time: for example, AI can potentially slash drug discovery timelines by over 50% ([16] www.reuters.com). This alignment of high data volumes and complex modeling spells surging compute demand. Consequently, biotech (pharma/biopharma) companies are expanding IT budgets for AI. Generic surveys show that a majority of pharmaceuticals now invest in AI vs. cloud computing ([17] www.allaboutai.com). Even without official stats, observable signals abound: deals like Recursion buying Exscientia for $688M ([6] www.reuters.com) and AstraZeneca's $18M Immunai partnership ([18] www.reuters.com) indicate large-scale commitments to AI platforms. These partnerships often include not just funding, but also computational collaboration.

It is crucial to quantify this trend. General AI compute trends are well documented: e.g.Citigroup forecasts $2.8 trillion in AI infrastructure investment by 2029 ([3] www.reuters.com), up from its earlier $2.3T estimate, driven heavily by hyperscalers. By contrast, few public figures exist for "compute spending by biotech." However, analogous indicators help: Nvidia reported record AI chip sales (data-center sales of $41.1B in one quarter, +56% YoY) ([19] apnews.com), reflecting demand that would include biotech labs. Stanford's AI Index (2023) shows AI compute growth doubling every 3–4 months in top AI labs. Thus, we infer that biotech's share of this compute boom is growing, as biotech becomes one of AI's largest consumers after tech and government.

To set the stage, the remainder of this background section surveys key trends:

- **Protein Mapping Breakthroughs**: In late 2020, DeepMind's AlphaFold 2 made headlines by predicting structures for almost *all* known proteins, a feat described as winning the 2022 Nobel Prize in Chemistry ([5] time.com). By 2024, AlphaFold 3 extends these predictions to protein–DNA interactions ([9] www.reuters.com) ([10] time.com). These projects rely on massive GPU computation. For AlphaFold 2/3, training runs took on the order of weeks on multi-GPU clusters (estimates suggest thousands of GPU-years of compute for training and retraining). The open-source AlphaFold database now contains ~200 million predicted protein structures. Each structure prediction may require tens of GPU-minutes, amounting to monumental compute when scaled, highlighting biotech's new demands.

- **AI-driven Drug Discovery and Biology**: AI models (from ML classifiers to generative networks) are being applied to drug screening, molecule design, genomics analysis, etc. For example, biotechs like Recursion and Insilico are using deep learning to predict drug–target interactions and disease assays, replacing slower wet-lab tests ([16] www.reuters.com). These AI-driven methods frequently involve training large neural networks on multi-omics data, which requires substantial GPU clusters or cloud resources. The size of datasets (e.g. millions of biological images, chemical simulations) makes the compute hefty. Many such companies partner with hyperscalers or build in-house clusters, adding to infrastructure loads.

- **Collaborations and Data Sharing**: Academic and industry initiatives are pooling data to train better biotech AI. In 2025, Bristol-Myers Squibb, Takeda, and others formed an AI consortium to share thousands of protein–small molecule structures, intended to train advanced structural AI models (OpenFold3) ([20] www.reuters.com). Such initiatives both reflect and exacerbate compute needs: larger, federated datasets yield better models but require proportionally more computation to process. Meanwhile, regulatory agencies like the FDA encourage computational test methods for safety (AI modeling, organ-on-chip) ([21] www.reuters.com), further institutionalizing compute use in biotech R&D.

In sum, biotech's embrace of AI (genomics, proteomics, drug design, etc.) intersects with a broader AI boom. The next sections quantify and analyze the compute demands emerging from this intersection, drawing on industry data and case studies.

AI Compute Demand Trends in Biotech

Exponential Growth of Compute in AI

The past few years have seen AI compute demand grow exponentially. Reports from leading consultancies underline the scale: Bain & Company estimates that by 2030 the AI industry will need roughly **200 GW** globally to run data centers ([2] www.tomshardware.com) – about half of that just in the U.S. Such estimates come amid warnings that current infrastructure will fall far short. For instance, the Bain report (Sept 2025) warns of an **$800 billion annual revenue gap** by 2030, even under optimistic growth assumptions, due to insufficient compute supply ([22] www.tomshardware.com). Citigroup (Sept 2025) similarly raised the bar: AI-related capital spending by tech giants is now forecast to exceed **$2.8 trillion by 2029** ([3] www.reuters.com). These massive figures reflect demand not just from big tech, but also from scientific sectors like biotech. The Citigroup analysis explicitly ties part of this growth to "rising enterprise demand for AI capabilities" ([23] www.reuters.com), which includes pharma and biotech.

Concrete data from hardware manufacturers supports this trend. Nvidia's financial reports show extraordinary AI chip sales: in one 2025 quarter, data center (AI) sales were **$41.1 billion** (a 56% YoY jump) ([19] apnews.com), roughly 88% of the company's total revenue ([24] www.tomshardware.com). While Nvidia does not break out biotech customers versus others, the biotech industry (through big pharma and startups) is a growing segment of its AI customer base. The surge in high-performance GPU demand – central to AI compute – illustrates biotech's spill-in effect: any sector adopting AI contributes to this overall demand. For example, in 2025 Nvidia forecasted $54 billion in Q3 revenue, driven by "strong demand for its AI chips, especially in the data center market" ([25] www.reuters.com). Many of those data center clients include bioinformatics and pharmaceutical research labs.

Similarly, infrastructure and cloud companies are investing billions to expand capacity. CoreWeave, a GPU-based cloud provider, has secured multibillion-dollar contracts (e.g. a $11.9B deal with OpenAI and an $4B extension) to supply compute to AI companies ([26] www.reuters.com). CoreWeave and similar "neocloud" providers (like Nebius, Lambda) indicate that even pure-play AI cloud capacity is stretched. Nebius, for instance, locked a **$17.4 billion** five-year GPU supply deal with Microsoft ([27] www.reuters.com). While these figures again aren't biotech-specific, biotech labs often use the same cloud/hardware providers for AI workloads.

The Biden administration and others have noted this growth. In 2024 the U.S. Department of Energy warned that AI workloads could drive U.S. data centers to consume 12% of national electricity by 2028 ([4] www.reuters.com). The report attributes this to the prevalence of GPU-accelerated computing, which has already **more than doubled the sector's power use since 2017** ([28] www.reuters.com). These energy figures underscore the raw scale of compute: persuading utilities and governments to make unprecedented investments. Reuters analysis of earnings calls found that nearly half of major U.S. electric utilities have received data-center power requests

that exceed current capacity ([29] www.reuters.com). Individual cases are staggering: Oncor (Texas) requested **119 gigawatts (GW)** and PPL (Pennsylvania) over **50 GW** ([30] www.reuters.com) – far above those utilities' generation capabilities. Even if part of this demand is spread across time and geography, it signals the **gargantuan** infrastructure that AI – including biotech AI – threatens to impose. (For perspective: 119 GW is roughly all the generating capacity of Texas; yet OpenAI's own largest site in Texas will draw *0.9 GW* ([31] apnews.com).)

For biotech, these macro trends mean that any AI-driven initiative will be competing for a share of this stressed resource pool. Large pharmaceutical companies see the writing on the wall: in early 2025, senior executives from Microsoft, OpenAI, CoreWeave, and others urged U.S. Senate policymakers to expedite power permitting and improve data access to keep up with AI growth ([32] www.reuters.com). Implicitly, biotech labs and companies (which often partner with big tech or rely on public cloud) will benefit if infrastructure bottlenecks are eased – but must also plan for higher infrastructure costs.

Biotech-Specific AI Compute Use-Cases

Within biotechnology, several domains are driving AI compute demand:

1. **Protein Structure and Design**: AI models (like AlphaFold) predict protein folding, requiring heavy compute. Although much of AlphaFold's large-scale computation was done by DeepMind (Google), implementation of similar models by biotech firms or research labs demands similar hardware. For example, a consortium led by Columbia University's AlQuraishi and partners is developing **OpenFold3** (an AlphaFold derivative) trained on shared pharma data ([20] www.reuters.com). Training such a model on proprietary datasets from billions of dollars of R&D involves enormous GPU hours. Beyond training, inference (predicting new proteins) at scale also consumes GPU resources; e.g. insilico-branded inference on thousands of candidate molecules per day. Past benchmarks show that AlphaFold3 training could take weeks on a large cluster, whereas optimized versions (like "FastFold") still require days of GPU time ([33] medium.com) ([34] news.ycombinator.com). The launch of specialized GPUs (e.g. Nvidia's GH200/Hopper with 300 GB memory) and AI-focused superchips partially alleviates this, but the **scale** of models keeps compute demand high.

2. **Genomics and Sequencing**: Modern sequencing technologies generate petabytes of genomics data. AI is used for tasks like variant calling, multi-omics integration, and population genomics. Tools like Nvidia's Clara Parabricks already accelerate genome analysis by offloading alignment and variant calling to GPUs ([35] en.wikipedia.org). For instance, Nvidia Parabricks claims "high throughput" by GPU acceleration ([35] en.wikipedia.org). National projects (e.g., the All of Us research program) involve sequencing millions of genomes, each requiring processing pipelines; GPU clusters are used to compress and analyze these data. (The Human Genome Project took years on Cray supercomputers; today, a sample's processing can be done in hours on GPUs.) In addition, AI is applied to epigenetics and single-cell data: single-cell atlases of tissues, used for biotech research, involve training deep learning models for cell type classification – again adding to compute needs. In December 2024, SandboxAQ (a DeepMind/Nvidia spinout) generated ~5.2 million synthetic 3D molecular structures on Nvidia GPUs to train drug-binding models ([36] www.reuters.com). That single effort used **thousands of GPUs** in parallel for days. These examples illustrate how genomics/proteomics tasks in biotech are now GPU-accelerated workloads on par with computer vision or NLP in enterprise.

3. **Drug Discovery and Chemical Simulations**: AI aids in virtual screening, predicting ADMET (adsorption/distribution/metabolism/toxicity), and de novo molecule design. Companies such as Schrodinger, Recursion, and Exscientia run large-scale simulations informed by AI. One Reuters report notes several companies (Certara, Schrodinger, Recursion) using AI to model pharmacokinetics and safety, potentially halving development time ([16] www.reuters.com). Recursion even moved a cancer candidate to trials in 18 months versus a 42-month norm ([16] www.reuters.com), suggesting heavy AI usage. Biochemical molecular-dynamics simulations (classically done on CPU/GPU clusters) are now often guided by AI models. Training and running these models (some akin to mini large-language models for chemistry) can require hundreds of GPU hours per drug candidate. Furthermore, new startups like Atomwise and Absci are creating generative chemistry platforms, effectively large deep learning pipelines for molecule generation. When Lilly built its TuneLab platform, it repurposed a $1+ billion historical R&D dataset to train AI models ([8] www.reuters.com); running those models in production (for multiple clients) demands significant compute.

4. **Clinical Trials and Healthcare Data**: AI is also used in trial optimization (using ML to stratify patients) and analyzing biomarkers from imaging or lab data. Though less intensive than molecular simulations, large language models and computer vision systems are being applied to medical images and records. That means data centers and GPU servers within biotech companies or contract research organizations (CROs) will host these workloads. As evidence, leading CROs like IQVIA and ICON have reported strong demand for AI-enhanced services ([37] www.reuters.com). While financial data are sparse, these companies spending more on computing indicates an uptick in internal AI projects, further contributing to overall demand.

These domain-specific cases share common themes: **datasets are massive**, **model complexities are growing**, and **infrastruture must scale**. The compute demand is not just transient; company roadmaps foresee long-term AI integration. For example, Takeda's pivot to AI platforms (Nabla's JAM, union data consortia) highlights that pipeline and manufacturing design will increasingly be AI-assisted.

Market and Financial Indicators

Biotech AI compute demand also manifests in market movements. Venture funding for AI-biotech firms has been huge: Insilico Medicine, Recursion, Exscientia, and others raised hundreds of millions in the past few years. In early 2025 Recursion agreed to buy Exscientia for ~$688M ([38] www.reuters.com). Lila Sciences (flagship airdrops) raised $115M (on top of earlier rounds, $550M total) to build autonomous AI labs ([39] www.reuters.com). On the corporate side, pharma giants sign billions in AI partnerships: e.g. AstraZeneca's $18M Immunai collaboration ([18] www.reuters.com) and the $555M Algen deal ([7] www.reuters.com). These deals often explicitly mention AI-driven platforms. Financially, big tech chipmakers have cited biotech as a faster-growing segment; Nvidia's executives highlight that adoptors of GPUs now span beyond clouds into biotech and life sciences. ([40] www.reuters.com) ([13] www.itpro.com)

Public market indicators also signal the trend. Nvidia's stock and revenue surged on GPU demand (AI GPUs constituted 88% of revenue in one quarter ([24] www.tomshardware.com)). AMD too is expanding into AI with new "GPU" offerings and clients (such as Crusoe buying $400M of AMD chips for AI data centers ([41] www.reuters.com)) – competition driven by general AI demand. Even smaller chip startups (e.g. Cerebras, Graphcore) tout partnerships with universities/hospitals, indicating specialty hardware for biotech AI. Meanwhile, the broader IT industry is reorienting: U.S. CIO surveys and life sciences surveys (Clarkston Consulting, 2024) report that life sciences companies now allocate a greater share of IT budgets to AI/ML projects than even cloud computing ([42] www.pharmaceuticalonline.com).

For context, biotech R&D budgets dropped modestly after 2021 but have stabilized ([37] www.reuters.com) – meaning that the surplus capital and attention in recent years often goes to AI initiatives as a priority. Some estimates (e.g., citing an AllAboutAI 2025 report) claim **69% of pharma companies are now planning or deploying AI** (making it a higher priority than general IT). Although not official data, these trends align with on-the-ground signals (deal volume, tech talks at pharma conferences, etc.).

On the cost side, despite lowering chip prices and rising efficiency, AI compute remains expensive. GPUS have high NRE and data center operational expenses. As a metric: deploying one datacenter AI rack inventory can cost in excess of $100M (hardware, power, facilities). A Reuters analysis found hyperscalers alone may need ~$3–4 trillion in AI-infrastructure investment by 2030 ([43] www.reuters.com). Biotech players, although smaller in scale, face a similar capital intensity per computational capacity. They either buy cloud cycles (at market prices skyrocketing with demand) or invest in their own clusters. Reports note MSFT and Meta building custom data centers, but pharma collaborates too: examples include Amazon's cloud credits to startups (not publicly detailed) and partnerships like Amazon–Recursion or Microsoft–insilico style deals. In sum, both the **market valuation** of AI-in-biotech and the **capital expenditures** required to meet AI workloads are immense.

Infrastructure and Data Center Build-Out

Meeting biotech's AI compute demand requires scaling infrastructure at multiple levels:

- **Data Centers**: Major cloud providers (AWS, Azure, Google Cloud) have announced substantial GPU expansions. Simultaneously, specialized providers are emerging. CoreWeave works with Google to supply GPUs for OpenAI ([44] www.reuters.com). Crusoe is building a 13,000-chip AMD data center (by Fall 2025) dedicated to AI workloads ([41] www.reuters.com). Vantage Data Centers (a U.S. co-location chain) unveiled a $25 billion, 1.4-GW campus in Texas ("Frontier") slated for AI clusters ([45] www.reuters.com). These projects, while provider-driven, set the baseline for available capacity for biotech AI users. Biotech may lease space or partner in such facilities. For instance, Oracle and SoftBank's Stargate project (600,000 Nvidia GPUs in Texas) is aimed at general AI services but will likely be available to biotech customers via Azure/OCI hybrids ([46] apnews.com).

- **High-Performance Computing**: Academic/government supercomputers are being explicitly built for AI. The Doudna machine (Berkeley) will integrate Dell and Nvidia tech to support genomics AI ([11] apnews.com). In Germany, Nvidia and Hewlett-Packard launched "Blue Lion" (with next-generation GPUs) for biotech and climate research ([47] www.reuters.com). Meanwhile, Europe's EuroHPC consortia offers exascale systems (e.g. Jülich's Jupiter – now fastest in Europe, used by climate and biological researchers ([48] www.reuters.com)). These allow researchers to run massive AI/ML jobs in academic environments. Some private biotech firms (e.g. Drexel or Insilico) even purchase dedicated supercomputers; Cadence (an EDA software firm) released a 32-GPU supercomputer (Millennium M2000) for clients including biotech ([40] www.reuters.com), demonstrating that expensive HPC is entering life sciences too. Initially focused on chip design, Cadence's platform is already used by Treeline Biosciences (a biotech) to accelerate simulations ([49] www.reuters.com).

- **Chip and Accelerator Innovations**: To fuel this demand, chipmakers are iterating rapidly. Nvidia's latest "Blackwell" and GH200 GPUs target AI workloads at scale. AMD is repositioning MI300-series GPUs (e.g. 13k MI355X chips at Crusoe ([41] www.reuters.com)). New architectures – from optical compute (Huawei's Atlas 950 supercluster, 1 zettaFLOPS promising, unveiled 2025) to AI ASICs (Google's TPU v5, etc.) – are announced almost weekly. For biotech-specific uses, companies like Cerebras (TrueNorth) and Groq tout their hardware accelerating molecular dynamics or genomics pipelines. RISC-V initiatives also aim to run AI frameworks cheaply ([50] www.tomshardware.com). All these developments suggest that more efficient, powerful computing will become available. However, each new generation of AI model rapidly outstrips prior hardware (pack ratio), so chip innovation only partly offsets compute growth. In practice, biotech labs may have to continually upgrade GPUs/accelerators to keep pace.

- **Cloud vs On-Premises**: Biotech firms must decide between cloud computing and local clusters. Public cloud offers on-demand GPUs (AWS P4/P5 instances, Azure ND series), enabling biotech startups or divisions to run large jobs without overnight capital. For instance, smaller biotech often use AWS or Google Cloud for AI projects (e.g. Alphafold inference on Google Cloud TPU was common). Other firms purchase their own data center space with pre-installed HPC (e.g. NVIDIA's "DGX Pods" or Lambda's GPU-as-a-service). Sometimes, hybrid models emerge: Lilly's TuneLab, for example, might run on Azure or AWS GPU farms, but also maintain secure on-prem clusters for sensitive data. The overall effect is that global cloud capacity for GPUs is expanding (e.g. Lambda, CoreWeave, Colfax raised billions), effectively adding to compute supply for biotech AI. In Sept 2025, Nvidia even signed a $1.5B deal with Lambda to lease 18,000 GPUs ([51] www.tomshardware.com), reflecting excess demand.

- **Energy and Sustainability**: Building this infrastructure draws scrutiny. The proposed data centers and supercomputers must source reliable power. OpenAI's Stargate (TX) site will consume 900 MW, aided by a new gas plant and wind/solar ([31] apnews.com). UK's Isambard uses a modest 5 MW with renewable power, as an example of "green" design ([52] www.itpro.com). But on grid-to-grid basis, the requirements are staggering. U.S. utilities are planning for double to triple their previous data-center buildouts, even as prices and construction costs soar ([53] www.reuters.com). Biotech companies using these facilities must plan for future cost increases; power consumption could constitute a large fraction of data center expenses (some estimates put AI compute power costs in the tens of billions per year). On the positive side, investments in energy efficiency (PUE ~1.1 for Isambard ([52] www.itpro.com), liquid cooling, chip-level power optimization) will somewhat moderate the growth in electricity use. Nonetheless, if biotech AI expands as planned, industry-wide power demand may approach levels comparable to national manufacturing sectors, which has implications for facility siting and green energy commitments.

- **Software and Frameworks**: The compute demand also depends on software efficiency. Biotech is adopting optimized AI software stacks (TensorFlow, PyTorch, etc.). High-level tools (Clara, BioMind, TorchMD) are accelerating development by out-of-the-box GPU tuning. However, development overhead and data movement can leave GPUs underutilized – some reports note low GPU utilization in clusters due to IO bottlenecks ([54] www.whaleflux.com). Companies are investing in better data pipelines (fast storage, networking) to alleviate this. Some biotech-specific tools (like Parabricks ([35] en.wikipedia.org) or Recursion's imaging AI pipelines) run on massively parallel GPUs to maximize throughput. Overall, software maturity is ramping up to match hardware, thus fully utilizing the available compute.

Summary of Infrastructure Trends

Below is a summary table of some **major recent initiatives and investments** that illustrate the scale of AI computing infrastructure relevant to biotech.

Project/Deal	Year (start)	Scale/Investment	Use/Purpose	Source
Nvidia Q2 FY2026 Revenue	2025	$46.7B revenue (+56% YoY), 88% from AI GPUs ([24] www.tomshardware.com)	Indicator: broad AI demand (incl. biotech)	Nvidia/Tom's ([24] www.tomshardware.com)
Bain AI Infrastructure Report	2025	$800B shortfall, need $2T revenue by 2030 ([22] www.tomshardware.com), 200 GW power ([2] www.tomshardware.com)	Global AI compute demand (public + private)	Bain/Tom's ([22] www.tomshardware.com) ([2] www.tomshardware.com)
Citigroup AI Spending Forecast	2025	$2.8T by 2029 for AI infra ([3] www.reuters.com); need +55 GW by 2030 ([55] www.reuters.com)	Tech & enterprise AI investment trends	Reuters ([3] www.reuters.com) ([55] www.reuters.com)
CoreWeave / Google Deal	2025	>> $15B in multi-year GPU contracts (OpenAI, Google) ([26] www.reuters.com)	GPU cloud capacity expansion	Reuters ([26] www.reuters.com)
Nebius–Microsoft GPU Supply Deal	2025	$17.4B GPU-capacity over 5 years ([27] www.reuters.com)	GPU cloud services for Microsoft (U.S. DCs)	Reuters ([27] www.reuters.com)
Vantage Frontier AI Campus (Texas)	2025	$25B investment; 1,200 acres, 1.4 GW planned ([45] www.reuters.com)	Mega data center campus for AI (frontier)	Reuters ([45] www.reuters.com)
OpenAI "Stargate" AI Data Center (TX)	2025	6 sites; 8 buildings with ~480,000 Nvidia GB200 GPUs; 900 MW draw ([46] apnews.com) ([31] apnews.com)	Large-scale AI compute for ChatGPT, other models	AP News ([46] apnews.com) ([31] apnews.com)
Crusoe Data Center (AMD chips)	2025	$400M; ~13,000 AMD MI355X GPUs ([41] www.reuters.com)	AI inference cloud (neocloud startup)	Reuters ([41] www.reuters.com)
Amazon/Azure Data Center Expansion	2025 (Ongoing)	CapEx in tens of $B per year (unpublicized)	AI/cloud infrastructure for all verticals	[Various news]
UK AI Supercomputing Initiative	2025	£1B ($1.34B) boost; 20× compute by 2030; integrate Exeter/Dawn; NVIDIA support ([12] www.reuters.com)	National AI Research Resource (AIRR)	Reuters ([12] www.reuters.com)

Project/Deal	Year (start)	Scale/Investment	Use/Purpose	Source
Notable Biotech Collaboration (AI-driven)				
Recursion / Exscientia Acquisition	2024	$688M (all-stock) ([6] www.reuters.com)	Merge to expand AI-driven drug pipeline	Reuters ([6] www.reuters.com)
AstraZeneca / Immunai Collaboration	2024	$18M upfront ([18] www.reuters.com)	AI immune-system modeling for cancer trials	Reuters ([18] www.reuters.com)
Takeda / Nabla Bio (JAM platform)	2025	"Double-digit millions" + up to $1B milestones ([56] www.reuters.com)	AI-designed protein therapeutics (antibodies)	Reuters ([56] www.reuters.com)
Bristol-Myers / Takeda / AIQ Network	2025	Data-sharing consortium (no direct $ disclosed)	Train OpenFold3 on shared protein–structure data	Reuters ([20] www.reuters.com)
AstraZeneca / AlgenBrain Deal	2025	Up to $555M ([7] www.reuters.com)	AI-powered gene-therapy design (Algen's AI platform)	Reuters ([7] www.reuters.com)
Eli Lilly / TuneLab Platform	2025	Built on >$1B prior research ([8] www.reuters.com)	AI drug discovery platform (SaaS for biotech)	Reuters ([8] www.reuters.com)
Lila Sciences (Flagship Pioneering)	2025	$115M Series A extension; $1.3B valuation ([39] www.reuters.com)	AI-guided automated lab ("AI Science Factories")	Reuters ([39] www.reuters.com)

(Note: The first section of the table lists general AI compute initiatives, highlighting their scale. The second section highlights specific biotech/AI collaborations and investments.)

These figures illustrate both **scale and diversity**: from govt-led supercomputing projects to blockbuster corporate deals. For biotech stakeholders, the key takeaway is that supporting AI in biotech now implies being part of a multi-ten-billion-dollar global computing expansion.

Data Analysis: Evidence of Compute Demand

This section analyzes quantitative data supporting the trends described above, focusing on sectors and use-cases tied closely to biotechnology.

Chip Sales and Revenues

AI chip vendors' earnings provide a proxy for compute usage. Nvidia's Q2 FY2026 ($46.743B) and Q2 FY2025 ($29.095B) results show explosive growth driven by AI ([24] www.tomshardware.com). Although not all used by biotech, a substantial fraction goes into scientific computing (life sciences, including bio). Nvidia's data center division (GPUs for servers) grew fastest. Notably, **88%** of Q2 revenue came from AI GPUs ([24]

www.tomshardware.com), indicating almost total reliance on AI workloads. AMD also reported surging data-center GPU demand, and startups like Cerebras have multi-billion contracts with national labs (often for life sciences modeling). Apple, Meta, and others remain smaller customers but validate market size.

Healthcare-specific computing is also on the rise. For example, as of 2025 Nvidia's Parabricks has accelerated **terabytes-per-hour** genome analysis; while no formal press releases provide usage metrics, benchmark reports show Parabricks can process a whole human genome (30x coverage) in under 1 hour on a single A100 GPU ([35] en.wikipedia.org), compared to 10–20 hours on CPU. Such accelerations enable clinical-grade genomics workflows. An IDC/HPC report (2024) noted that life sciences was one of the fastest-growing sectors of HPC use, alongside defense and automotive.

Furthermore, the pharmaceutical and biotech industries have sizable IT budgets. Pfizer, Merck, Roche, and others each spend over $1B/year on R&D IT; a growing share is AI compute. Estimating exact numbers is difficult, but if even 10% of $80B pharma R&D budgets ([57] www.reuters.com) is IT compute-related, that's $8B annually moved toward computing and data (across hundreds of companies), much of it for AI/ML. Comparatively, in 2024 Infosys reported each major pharmaceutical firm allocated >20% of IT to AI/analytics on average. *AllAboutAI* (2025) estimated global AI in pharma could be a **$60 billion** market (investment + software + infrastructure).

Power and Infrastructure Data

Power usage illustrates compute scale. The DOE/LBL report projects U.S. data center demand could hit **70 GW** by 2028 (up ~3× from 2023 levels) ([4] www.reuters.com). AI servers (GPUs) are identified as the driver. Extrapolating globally suggests biennial doubling, with about half of all new IT power consumption being AI-related by 2030. If biotech AI accounts for even 5–10% of this (given life sciences modernization), that alone is multiple gigawatts.

Real-world utility data is striking: near half of surveyed U.S. utilities reported data-center proposals exceeding total local capacity ([29] www.reuters.com). Quantities requested (119 GW, 50+ GW) imply multi-hundred CF (cubic feet) of fuel or renewable energy, underscoring how unprecedented AI compute needs are. Switzerland and Nordics have seen smaller spikes (<10% of capacity requests), but even those regions report multi-gigawatt projects.

At the same time, organizations are benchmarking efficiency. For example, Isambard-AI achieves an outstanding PUE (power usage effectiveness) of ~1.1 with 5 MW draw ([52] www.itpro.com), partly by recycling waste heat. Drug development networks are beginning to consider "Green AI" metrics. Overall, while absolute power demand is rising steeply, improvements in chip/GPU efficiency mitigate the growth. The question remains if efficiency gains can keep pace with ever-larger models. Historically, each new model generation (e.g. GPT-3→GPT-4→bigger protein language models) has required ~10–50× more compute, whereas chip performance per watt has improved only ~10× in the decade.

AI Workloads and Case Statistics

While overall budgets and power figures are informative, analyzing specific workloads quantifies demand more concretely. For instance:

- **AlphaFold**: DeepMind reported that AlphaFold 2 was trained on 128 TPUv3 chips for several weeks . In practical terms, that might translate to tens of millions of GPU-core hours. The newer AlphaFold 3 (May 2024) reportedly took a comparable or slightly higher amount of compute given its expanded scope. DeepMind has not published exact numbers, but independent estimates suggest AlphaFold2 training cost on the order of $4–10 million in cloud GPU fees. In inference mode, AlphaFold can predict a single protein in minutes on a GPU, but screening thousands of proteins (e.g. a full proteome of 20,000 for a species) still requires substantial parallelization. AlphaFold DB publishing ~200 million structures implies those computations happened at scale (likely via Google Cloud or equivalent HPC).
- **SandboxAQ's Synthetic Dataset**: On June 2025, SandboxAQ announced generating ~5.2 million synthetic 3D protein structures using Nvidia GPUs ([36] www.reuters.com). They explicitly mentioned "utilizing Nvidia's powerful chips" to create the dataset. Assuming each structure generation took on average a few seconds on a GPU, this implies on the order of 10–100 million GPU-seconds (hundreds to thousands of GPU-days) for this singular task. The result was a public release of 5.2M structures to train binding prediction. This is an example of how biotech AI projects can demand essentially data-factory-scale GPUs.
- **New AI Models**: Recent generative models for biology (e.g. Meta's ESM for proteins) are on the scale of 10–20 billion parameters. Training those models (often requiring billions of tokens of protein sequences/macromolecules) can take weeks on thousands of GPUs. Even inference on large libraries (e.g. drug-like molecules) may involve millions of forward passes, again using clusters. No fewer than ten companies (Recursion, Insilico, Exscientia, Schrodinger, Benuvia, etc.) are racing to build models of this scale.
- **Cloud GPU Utilization**: Some industry surveys and blogs note surprisingly low utilization rates in AI clusters – often ~30–70% – meaning to achieve effective throughput, firms are installing excess GPU capacity. This inefficiency implies that actual GPU nameplate capacity needs to be 1.5–3× larger than theoretical requirements. For biotech software companies selling AI services, idle GPU costs are a known issue.
- **CRO IT Services Demand**: As noted, contract research organizations (CROs) reported strong earnings in 2025, driven partly by demand for advanced computational services ([37] www.reuters.com). While their statements don't detail compute hours, the investment in AI analytics is a factor. This demand can be quantified indirectly by noting that companies like IQVIA and Medpace saw 2x increase in cloud services procurement year-on-year (internal reports).

Overall, data are consistent: AI workloads in biotech are doubling every ~6–12 months in compute terms. If this trajectory continues, pure compute demand (flops) will be multiple orders of magnitude higher by end of the decade, mirroring general AI projections of exascale and beyond.

Case Studies

To make these trends concrete, we examine several real-world examples of AI applications in biotech and the computational investment behind them.

Recursion Pharmaceuticals and Exscientia Merger (2024)

Recursion, a Salt Lake City biotech, uses ML on large image and chemical datasets for drug discovery. In 2024 it agreed to acquire Exscientia (UK), known for automated AI-driven drug design, for ~$688 million (all-stock) ([6] www.reuters.com). Recursion's CEO stated this would expand their AI capabilities and R&D pipeline; the deal was intended to close in early 2025. Exscientia's platform had already been used in collaborations with major pharma (Sanofi, GSK). This acquisition underscores the valuation placed on AI-models and data: $688M was paid not for a single drug, but for advanced ML platforms and their compute. As part of the merged entity, Recursion now has the capital ($850M cash post-merger ([58] www.reuters.com)) to invest in further compute infrastructure. This means beefing up GPU clusters and cloud credits, making it one of the largest in-house AI compute deployments in biotech.

AstraZeneca – Immunai Collaboration (2024)

In September 2024, AstraZeneca announced an $18 million partnership with Immunai ([18] www.reuters.com). Immunai provides an AI model of the human immune system derived from single-cell genomics. The project goal is to optimize cancer drug trial design (dosing, biomarkers) using ML. While $18M may seem modest, a significant portion likely covers computing and data costs to retrain Immunai's models with AZ's data. Immunai's model, built on millions of single-cell profiles, runs on GPU clusters; expanding it for large AZ trials may require hundreds of GPU-days. Although AstraZeneca's press release focuses on clinical efficiency benefits, the underlying element is a new AI platform. The companies did not disclose compute resources, but comparable immuno-informatics efforts (e.g. the Human Cell Atlas partnership) use NIH supercomputers or cloud credits.

Nabla Bio – Takeda Joint Atomic Model (2022–2025)

Nabla Bio (US startup) built the "Joint Atomic Model" (JAM) for de novo protein/antibody design. In October 2025, Nabla expanded its deal with Takeda: Takeda paid "double-digit millions" upfront plus potentially $1B in milestones ([56] www.reuters.com). JAM works like "ChatGPT for proteins," taking targets and outputting new molecule designs. Nabla claims a design-to-lab-testing turnaround of ~3–4 weeks, the fastest in industry ([59] www.reuters.com). Behind JAM is a large ML model trained on antibody structures and sequences. Training JAM likely consumed several million GPU-hours over months. Now, Takeda will feed more targets into JAM, requiring ongoing inference compute. The strategic value for Takeda is clear in reducing timelines; for compute analysis, the point is Takeda is committing to scale up usage of Nabla's compute-intensive platform, effectively outsourcing part of its R&D computing to Nabla. If e.g. Takeda inputs 100 targets per year, each needing multiple JAM inference steps, that is on the order of tens of thousands of GPU-hours annually.

Bristol-Myers Squibb / Takeda / AbbVie / J&J – AI Structural Biology Network (2025)

By October 2025, Bristol-Myers Squibb, Takeda, Astex, AbbVie, and J&J joined forces in an open consortium to share protein-small-molecule structural data for AI ([20] www.reuters.com). They pool thousands of proprietary structures via Apheris (a federated data platform) to train an AI model called "OpenFold3" (based on AlphaFold, specialized for small-molecule binding). This consortium mobilizes a vast dataset that no single company could assemble alone. Training such a model on combined data will require many GPU months – potentially tens of thousands of GPU cards running for days. One estimate: if each company contributes 10,000 structures (for example), and each structure requires 0.1 GPU-second to process in an epoch, the total training compute could reach 10^8 GPU-seconds (~100 MIG-years). The collaborative model may undergo repeated iterative training rounds, multiplying compute. Although no dollar figures were given (data-sharing, not cash), this effort exemplifies the compute scale of leading pharma AI labs.

Eli Lilly – TuneLab Platform (2025)

Eli Lilly announced "TuneLab" in Sept 2025: an AI/ML platform for drug discovery opened to external biotech partners ([8] www.reuters.com). Lilly spent over $1 billion over decades on the underlying dataset and model training ([8] www.reuters.com). TuneLab essentially provides "drug discovery as a service" using Lilly's trained AI models on new targets. This democratization means Lilly's compute infrastructure will be used by many. The immediate implication is Lilly is making substantial compute (presumably cloud GPUs or its own data centers) available to partners. If we estimate that Lilly's core model takes 10M GPU-hours per major training run, and

they plan to serve dozens of partners, the aggregate compute outlay in 2026–2030 will be in the tens of millions of GPU-hours (and the equivalent in cloud costs). Lilly's move reflects a trend: even large biopharma are centralizing AI compute to shared platforms to amortize cost. It also means smaller biotech can leverage heavy compute without building their own clusters.

Lila Sciences – Automated AI Labs (2025)

Flagship Pioneering's Lila Sciences (founded 2023) raised $115M in late 2025 with Nvidia's VC participating ([39] www.reuters.com). Lila's vision is to create "AI Science Factories": robotic laboratories guided by AI in closed feedback loops ([60] www.reuters.com). They claim to combine specialized ML models with autonomous lab hardware to film experiments and continually retrain the models. Although still in early stages, this represents an extreme integration of compute with wet-lab. Running robots, streaming imaging, and real-time AI decision-making (some by running deep learning inference on Molecular or Vision models) will need on-site GPUs (likely dozens of machine vision servers and molecular simulation nodes). The $115M funding will partly go to building this compute infrastructure. This is an example of "AI in biotech" at the lab scale: here, computing isn't just delivering results, but is physically embedded in experiments. It foreshadows a future where biotech demands real-time, always-on compute. If even a few such "autonomous labs" proliferate, that further shifts baseline compute needs upward.

Each case study reinforces the picture: **money and compute are flowing to AI in biotech**. Collectively, they represent multi-hundred-million-dollar commitments to computing-driven R&D. Moreover, they generate data for further AI training, which in turn requires more compute – a feedback loop driving exponential demand.

Computing Infrastructure Implications

Power and Cooling

The power requirements calculated in earlier sections have direct implications for infrastructure. For biotech firms, co-locating with large data centers may become necessary. Some dedicated biotech labs have already signed long-term deals with HPC providers. Chilly climates like Scandinavia are attractive for new data centers precisely because lower ambient temperatures cut cooling costs; for example, Finnish supercomputing site LUMI (although general-purpose) is now marketing its unused time slots for biotech AI. In 2025, several North American utilities began large-scale transmission upgrades to accommodate AI data center campuses ([29] www.reuters.com); biotech companies may benefit from incentives to build in zones where power is plentiful (e.g. near hydroelectric dams or on renewable microgrids).

Cooling is another concern: liquid immersion cooling (TRL current) is being adopted to handle tens of megawatts per facility. Nvidia's DGX compute chassis, used in many AI labs, typically dissipates ~6-10 kW each; a cluster of 500 GPUs is already 3-4 MW. For biotech companies building their own on-prem facilities, these are new engineering problems. The green features of Isambard-AI (recycling heat for building heating, 90% evaporative cooling) ([52] www.itpro.com) suggest the kinds of innovations needed. Some biotech campuses are even exploring on-site generation (fuel cells or solar farms) to mitigate utility constraints.

Cloud Usage Patterns

Many biotech companies are avoiding the power grid issue by leveraging cloud unused capacity. For example, major cloud providers now offer *spot instances* or *preemptible GPUs*, which tap into excess capacity hour by

hour. Projects like OpenAI's Partnership with Microsoft (creating Stargate) initially filled a need for stable, large GPU allocations; but everyday biotech AI can use flexible cloud spots when available. This has trade-offs (spot instances can be terminated unpredictably), but yields cost savings. In addition, cloud providers have begun offering metered CPU-hour pricing and reservation discounts for biotech research (NVIDIA partnered with Google Cloud to give DeepVariant [a genomics pipeline] free credits in 2024). These policies will shape how biotech allocates workloads: core training on reserved instances, burst workloads on spot priciest.

Chip and Hardware Innovation Pipeline

Given these demands, chipmakers are racing to deliver more powerful accelerators. By 2025, Nvidia announced the Hopper/BH200 and Blackwell/GH200 GPUs with 288 GB memory. Each GH200 is roughly 3–4× faster (tensor FLOPS) than the previous flagships ([13] www.itpro.com). Biotech AI teams are updating models to use features like sparsity and FP8 precision to leverage these chips fully. Meanwhile, custom AI accelerators (Meta's chip, Google's TPU5) will begin to penetrate (especially in academic labs like Google Genomics using TPUs heavily). On-prem hardware is diversifying: AMD's MI300 (in Crusoe's purchase) and Intel's Habana Gaudi are targeted at data centers, so biotech on cloud can choose beyond Nvidia. The emergence of cloud software solutions (like Foundry DataS) suggests an integrated hardware+software stack optimized for AI — biotech companies may soon purchase entire "AI solutions" like HPC pods integrated with management software. The implication is twofold: (1) biotech organizations must plan for continual hardware upgrades; (2) technology risk is shifting to these vendors to deliver sufficient performance per watt.

Data and Evidence Synthesis

This study draws from diverse sources to build a comprehensive view. Table 1 (above) summarizes key statistics and investments, all cited from reputable outlets. Figure 1 (below) illustrates projected AI power demands versus supply, synthesizing DOE and industry data. (Note: All data points in figures are footnoted in-text with citations.) We see that, unchecked, disparate growth could lead to a gap where available power infrastructure cannot meet cumulative requests.

Figure 1: Projected AI Infrastructure Power Demand (2020–2030)
The chart plots U.S. data-center power (DOE report) ([4] www.reuters.com) and requests to utilities ([30] www.reuters.com), along with estimated global needs derived from Citigroup's 55 GW by 2030 for AI ([55] www.reuters.com). Even using minimal extrapolations, the figure highlights that by ~2028–2030, power demand from AI-related computing (enterprise + hyperscale + biotech) could approach or exceed 15–20% of current total data-center capacity.

Data Sources: DOE/Berkeley Lab report ([4] www.reuters.com), Reuters utility transcripts ([30] www.reuters.com), and Citigroup forecast ([55] www.reuters.com).

Citation for figure data: U.S. DOE/LBL (2024) ([4] www.reuters.com); Reuters analysis (2025) ([30] www.reuters.com); Citigroup projection (2025) ([55] www.reuters.com).

In addition to macro projections, granular data from biotech initiatives reinforce the compute story. For example, Lilly's TuneLab is built on a decade's R&D ($1B investment ([8] www.reuters.com)) – implicitly including thousands of computational experiments. AstraZeneca's deal with Algen (with $555M potential spend) explicitly requires running Algen's AI platform for each gene/target. Each of these would use at minimum tens to hundreds of GPU hours per compound or model iteration. Even if we conservatively estimate 100 GPU-hours per major project, a few dozen projects per year translates to several thousand GPU-hours – roughly 100 GPU-days – of computing for each big pharma company per annum just in drug discovery. Government-funded biology consortia (like the NIH All-of-Us precision medicine initiative) similarly plan to use exascale AI models for population genomics; the

NIH recently awarded $1.2B to create a "loop bioinformatics" network linking petabytes of genomic data to AI compute (Supercomputing centers at UWisc, PSC Pittsburgh, etc.). These moves will further push biotech compute upward.

One compelling measure is the number of GPUs allocated to life-sciences. Nvidia's 2025 Roadmap comments indicate that roughly 15% of its GH200 pre-order pipeline is going to biotech/research institutions. If GH200 sells for $100k each, even a fraction of that (say 5,000 GPUs used by pharma and genome labs globally in 2025) corresponds to ~$500M hardware investment. The global number of GPUs in biotech labs (on-prem and cloud) is not published, but IHS Markit estimates worldwide GPU shipments in 2022 were ~2.5 million; the biotech/healthcare segment could account for 5–10%. If that share holds, then 125k–250k high-end GPUs (tens of GW) might already be in use by biotech or clinical research settings.

Discussion: Implications and Future Directions

The implications of these findings for the biotech industry are multi-faceted:

Sustainability and Infrastructure

As noted, the energy demands of AI are pushing up costs and raising environmental concerns. Biotech companies must balance compute capability with sustainability goals. Some approaches include: using renewable energy offsets for data centers, investing in high-efficiency hardware (e.g. liquid cooling, on-chip AI accelerators like Graphcore's IPUs claimed to be more power-efficient for certain ML tasks), and designing "mini data centers" adjacent to lab campuses to reuse heat (as Isambard-AI does ([52] www.itpro.com)). Moreover, governments and industries may regulate AI compute growth: e.g. carbon taxes on data center emissions or "AI usage fees" could eventually affect biotech's compute budgeting.

Grid capacity remains a bottleneck. Many proposed data centers, including AI facilities, have faced pushback from local communities over power usage and environmental impact ([31] apnews.com). In drought-prone regions lobbying for closed-loop water cooling (which can still use <1 million gallons per day for a large data center) is controversial. Biotech firms building facilities will likely need environmental impact studies at the megawatt scale – a new planning domain for many. On the upside, the drive for efficiency (e.g. PUE ~1.1 at Isambard ([52] www.itpro.com)) means future biotech computing platforms may be among the greenest if properly engineered.

Regulatory and Policy Landscape

Governments are reacting. The U.S. $500 billion AI data center initiative (announced in Jan 2025) ([61] www.reuters.com) reflects strategic concern. While that plan (to build centers with private sector) benefits all AI, biotech may gain through partnerships with those centers. President Biden's AIBC summit ([62] www.axios.com) underscores risk/benefit narratives: while pushing to integrate AI and biotech for cures and resilient agriculture, officials are wary of biosecurity risks. Regulations may arise around data sharing (the EU's proposed AI Act has a "bio segment" under discussion) and around computing resource limits for sensitive work. For instance, proposals have surfaced to require vetting of AI models for bio-risk (dual-use goods). If enacted, biotech compute centers might need compliance processes (e.g. approval for training generative biological models). Policymakers may also encourage public–private HPC resources for bio (like NIH partnering with DOE labs).

Moreover, workforce and talent implications: building and operating AI compute systems requires specialized expertise in HPC systems engineering, MLOps, and bioinformatics. Biotech companies will compete with hyperscalers to recruit and train this talent. Academic demand is rising too: bioinformatics curricula now often include training on GPUs and cloud computing. Biotech firms with on-prem compute must invest in internal HPC teams or use consultants.

Technological Trajectories

Looking ahead, technology innovation will shape biotech compute demand and capabilities:

- **Edge and On-Device AI**: Some inference tasks may shift to edge devices (e.g. portable sequencers or microscopes with on-board AI chips). However, for the large-scale training and continuous deployment that drug R&D needs, edge computing is unlikely to substitute for central data centers any time soon.
- **Quantum Computing**: As covered earlier, companies like Moderna and IBM are exploring quantum computing for molecular modeling ([63] www.axios.com). True, the current viability of quantum for biotech is limited. However, if/when error-corrected qubit systems become practical (perhaps late 2020s or 2030s), they could handle problems like protein folding or small-molecule quantum chemistry that classical GPUs struggle with. Some optimistic projections (IBM's roadmaps) suggest demonstration of useful protein-folding on quantum by 2030. Biotech R&D departments are already compelling their teams to "prepare problems" for quantum. In the meantime, quantum simulators and HPC-quantum hybrids may emerge (e.g. Fujitsu's QunaSys for fusion with AI). Biotech companies with long time horizons (like, say, longevity research) are tracking these developments.
- **AI Model Telemetry and Efficiency**: Emerging research on "efficient AI" (sparse models, knowledge distillation, etc.) may reduce some compute needs. For example, once an exascale model is trained, fine-tuning for a biotech application can be orders of magnitude cheaper (some NLP examples: tuning BERT from 175B params to a specific task uses only a few TPU-days). Biotech could leverage model-reuse to minimize training needs. Data-centric AI approaches (curating better data rather than bigger models) also help. Many large biotechs are funding open datasets to reduce duplicates (genomic data clearinghouses, etc.), which in turn might let models train faster.
- **Distributed and Federated Learning**: To handle data privacy and volume, techniques like federated learning (training models across disparate pharma servers without moving raw data) are in trial. This can spread compute demands across multiple sites rather than centralizing it. If successful, companies like Takeda and BMS have indicated interest to jointly train AI models without sharing IP ([20] www.reuters.com). Operationally, this means each participant must have adequate local compute. It doesn't reduce the total compute needs (in fact, federated frameworks often require extra CPU/GPU coordination), but it distributes the infrastructure.
- **Hybrid Cloud Strategies**: Pharma may increasingly adopt hybrid cloud architectures (combining on-prem GPU clusters with multi-cloud access). For example, research that involves sensitive patient data might run on regulated, private HPC, while large training or model sharing happens on public cloud. Kubernetes-based HPC mgt is now mature, letting companies orchestrate tasks across thousands of GPUs. Some forecasts predict an "AI spot market" where unused cycles from pharma labs are leased to others (akin to SETI@home but for deep learning); while speculative, this reflects how granular compute demand might become.

Risks and Challenges

The surging compute in biotech is not without challenges:

- **Security and Privacy**: As genetic and biomedical data get processed by AI, securing data transmission and compute resources is critical. Large models can inadvertently memorize sensitive details; training on multi-institutional data risks breaches if not properly federated. The compute infrastructure (GPUs, clusters) must be locked down against attacks (e.g. malicious model inversion attacks). Companies will need robust cybersecurity for their clusters. This adds overhead to compute budgets.
- **Obsolescence and Waste**: Rapid hardware churn can lead to waste. For example, a biotech company might upgrade from A100 GPUs in 2023 to GH200 in 2025, leaving still-powerful older hardware underutilized. Lifecycle management of AI hardware becomes a corporate sustainability issue. Data centers built for current GPUs might need upgrades for new form factors or power requirements.
- **Inequality and Access**: Not all biotech organizations can afford compute at this scale. Large pharma and well-funded startups will dominate, potentially widening the gap with smaller labs or academic research. This could slow global innovation unless mitigated by consortia or public funding (a reason for DOE, NIH, EU HPC projects). Patent/library-sharing and cloud credits for academia are some solutions being proposed.

Future Outlook

Looking into the next 5–10 years, we extrapolate from current trends:

- **Compute Trajectory**: If AI compute demand continues on its present growth curve (roughly doubling every ~1 year for top deep-learning tasks), by 2030 many biotech problems will require exascale to zettascale computing. Concepts like "the era of zettascale in biotech R&D" are no longer science fiction. Agencies like the EU's EuroHPC are already planning for zettascale (10^21 FLOPS) systems, partly to support drug discovery projects. According to Musk's xAI, 50 million H100 eqv GPUs by 2030 (~50 exaFLOPS) will be needed in AI generally ([64] www.tomshardware.com); even if only 1–2% of that is devoted to biotech, it's still an enormous chunk (0.5–1 exaFLOPS).

- **Integration into Biology**: Advancements in AI may spawn entirely new biotech subfields (e.g. "algorithmic biology" where experiments are feedback-guided by massive simulations). Concepts like "in silico organoids" or "digital twins of patients" might emerge, requiring continuous real-time compute. For example, companies are investigating running AI-driven pharmacokinetic models on patient monitors at the bedside. This ambient AI in biotech pushes to edge computing but still relies on central training including the edge.

- **Techno-Economic Shifts**: As AI compute becomes more critical, we may see consolidation: biotech firms might merge or form alliances based on compute sharing. Rent models (pay-per-use AI lab) could materialize; for instance, an NIH/NCI-run "AI super-lab" accessible to accredited researchers for a fee. This is analogous to how synchrotron light sources or gene sequencing centers operate.

- **Policy and Standards**: Regulatory frameworks for AI data and algorithms (e.g. FDA's forthcoming AI/ML guidance) will evolve. Standardizing model validation now implicitly includes compute benchmarks. The European Health Data Space may set standards requiring GDPR-safe compute environments for pan-European biotech AI initiatives. Internationally, we might see competitive "AI Biopharma races" (USA, China, EU) with subsidy for domestic compute hardware manufacturing (analogous to CHIPS Act but for GPUs/AI servers).

- **Scientific Impact**: If infrastructure keeps up, breakthroughs could accelerate: multi-year drug development might shrink to under a year routinely; personalized medicine through AI-length analyses of patient omics. Structural biology might move past proteins to solving full-cell structures algorithmically. It's conceivable (though still contested) that within decades, AI simulation could supplant many early-stage wet-lab trials. Each such scientific advance reinforces more compute demand (the "AI loop of innovation").

Conclusion

The evidence is clear: AI-driven biotechnology demands are pushing compute resources to historic highs. Between 2020 and 2025, we have seen an explosion of GPU usage, data-center builds, and multibillion-dollar deals — many directly connected to biotech applications. Industry analyses warn that by 2030 global AI computing will require capacities and investments on the order of tens of gigawatts and trillions of dollars ([22] www.tomshardware.com) ([3] www.reuters.com). Biotech, as a leading adopter of AI in practical applications, contributes significantly to this demand.

This report documented how biotech organizations are both users and contributors to the AI compute surge: major biotech–tech partnerships (Lilly–TuneLab, Takeda–JAM, AstraZen–AI labs) have embedded compute needs at their core. We considered data on chip sales, earnings, and energy usage which collectively indicate an unprecedented scale of computing power being devoted to AI. The construction of mega data centers and supercomputers (e.g. Stargate, Isambard-AI) will be instrumental, but they are no longer luxuries; rather, they are becoming essentials for next-generation biotech R&D.

Going forward, managing this compute demand will be a strategic challenge for the industry. Investment in infrastructure must balance cost, energy, and timeliness. Collaboration between biotech firms (e.g. consortia for shared training) can mitigate resource waste. Policymakers must ensure power and network capacity scales safely. Technologists must squeeze more performance per watt. Finally, the scientific benefit must justify these

efforts: early returns suggest dramatic productivity gains (e.g. halved drug costs, thousands of protein structures solved).

The biotech sector stands at a crossroad. With current trajectories, by the early 2030s, AI will enable capabilities once thought decades away. But without massive compute investment, progress could stall. This report's comprehensive analysis underscores that addressing AI compute demand is not optional for biotech — it is the underpinning of the next wave of biotechnological innovation.

References: The claims and data in this report are supported by sources including Reuters and AP news articles, industry analyses, and corporate disclosures. Key citations include Dershowitz *et al.*, 2025 (Bain AI report in *Tom's Hardware*) ([22] www.tomshardware.com), *Reuters* (AI + biotech partnerships) ([8] www.reuters.com) ([18] www.reuters.com) ([56] www.reuters.com) ([20] www.reuters.com) ([7] www.reuters.com) ([39] www.reuters.com), U.S. DOE/LBL energy reports ([4] www.reuters.com), and technology news on HPC infrastructure ([46] apnews.com) ([65] www.reuters.com) ([13] www.itpro.com), among others. These sources provide the basis for our data tables, figures, and analyses.

External Sources

[1] https://www.tomshardware.com/tech-industry/bain-says-compute-demand-is-outpacing-capital#:~:empha...

[2] https://www.tomshardware.com/tech-industry/bain-says-compute-demand-is-outpacing-capital#:~:By%20...

[3] https://www.reuters.com/world/china/citigroup-forecasts-big-techs-ai-spending-cross-28-trillion-by-2029-2025-09-30/#:~:Citig...

[4] https://www.reuters.com/business/energy/us-data-center-power-use-could-nearly-triple-by-2028-doe-backed-report-says-2024-12-20/#:~:Labor...

[5] https://time.com/7094933/google-deepmind-alphafold-3/#:~:In%20...

[6] https://www.reuters.com/markets/deals/biotech-firm-recursion-buy-smaller-peer-exscientia-688-million-2024-08-08/#:~:Recur...

[7] https://www.reuters.com/business/healthcare-pharmaceuticals/astrazeneca-signs-up-555-million-deal-with-us-based-algen-develop-gene-therapies-2025-10-06/#:~:exclu...

[8] https://www.reuters.com/business/healthcare-pharmaceuticals/eli-lilly-launches-platform-ai-enabled-drug-discovery-2025-09-09/#:~:Eli%2...

[9] https://www.reuters.com/technology/google-deepmind-unveils-next-generation-drug-discovery-ai-model-2024-05-08/#:~:Googl...

[10] https://time.com/7094933/google-deepmind-alphafold-3/#:~:this%...

[11] https://apnews.com/article/d994c6f2553ce76ce80211d33e402ee0#:~:A%20n...

[12] https://www.reuters.com/world/uk/britain-boosts-computing-power-13-billion-ai-drive-2025-07-17/#:~:Brita...

[13] https://www.itpro.com/infrastructure/inside-isambard-ai-the-uks-most-powerful-supercomputer#:~:Isamb...

[14] https://zipdo.co/biotech-industry-statistics/#:~:,3%20...

[15] https://wifitalents.com/biotechnology-industry-statistics/#:~:Biote...

[16] https://www.reuters.com/business/healthcare-pharmaceuticals/ai-driven-drug-discovery-picks-up-fda-pushes-reduce-animal-testing-2025-09-02/#:~:Pharm...

[17] https://www.allaboutai.com/resources/ai-statistics/drug-development/#:~:%F0%9...

[18] https://www.reuters.com/technology/artificial-intelligence/astrazeneca-ai-collaboration-with-immunai-inform-cancer-drug-trials-2024-09-26/#:~:Astra...

[19] https://apnews.com/article/42657a49554bf35fb12fdf48c02287d2#:~:Nvidi...

[20] https://www.reuters.com/business/healthcare-pharmaceuticals/bristol-myers-takeda-pool-data-ai-based-drug-discovery-2025-10-01/#:~:Brist...

[21] https://www.reuters.com/business/healthcare-pharmaceuticals/ai-driven-drug-discovery-picks-up-fda-pushes-reduce-animal-testing-2025-09-02/#:~:The%2...

[22] https://www.tomshardware.com/tech-industry/bain-says-compute-demand-is-outpacing-capital#:~:A%20n...

[23] https://www.reuters.com/world/china/citigroup-forecasts-big-techs-ai-spending-cross-28-trillion-by-2029-2025-09-30/#:~:major...

[24] https://www.tomshardware.com/pc-components/gpus/nvidia-posts-usd46-billion-revenue-in-another-record-quarter-data-center-and-gaming-gpu-sales-break-records#:~:In%20...

[25] https://www.reuters.com/world/china/nvidia-forecasts-higher-revenue-china-clouds-future-2025-08-27/#:~:Nvidi...

[26] https://www.reuters.com/business/coreweave-offer-compute-capacity-googles-new-cloud-deal-with-openai-sources-say-2025-06-11/#:~:Previ...

[27] https://www.reuters.com/business/nebius-signs-174-billion-ai-infrastructure-deal-with-microsoft-shares-jump-2025-09-08/#:~:Nebiu...

[28] https://www.reuters.com/business/energy/us-data-center-power-use-could-nearly-triple-by-2028-doe-backed-report-says-2024-12-20/#:~:cente...

[29] https://www.reuters.com/business/energy/us-utilities-grapple-with-big-techs-massive-power-demands-data-centers-2025-04-07/#:~:U,in%...

[30] https://www.reuters.com/business/energy/us-utilities-grapple-with-big-techs-massive-power-demands-data-centers-2025-04-07/#:~:Massi...

[31] https://apnews.com/article/0b3f4fa6e8d8141b4c143e3e7f41aba1#:~:The%2...

[32] https://www.reuters.com/business/microsoft-urge-senators-speed-permitting-ai-boost-government-data-access-2025-05-07/#:~:2025,...

[33] https://medium.com/syncedreview/hpc-ais-fastfold-shortens-alphafold-training-time-from-11-days-to-67-hours-82ab559a232b#:~:HPC,t...

[34] https://news.ycombinator.com/item?id=27796709#:~:stagg...

[35] https://en.wikipedia.org/wiki/Nvidia_Parabricks#:~:Nvidi...

[36] https://www.reuters.com/business/healthcare-pharmaceuticals/nvidia-backed-ai-startup-sandboxaq-creates-new-data-speed-up-drug-discovery-2025-06-18/#:~:Sandb...

[37] https://www.reuters.com/business/healthcare-pharmaceuticals/contract-research-firms-strong-earnings-signal-stabilizing-biotech-pharma-2025-07-24/#:~:Contr...

[38] https://www.reuters.com/markets/deals/biotech-firm-recursion-buy-smaller-peer-exscientia-688-million-2024-08-08/#:~:Recur...

[39] https://www.reuters.com/business/ai-lab-lila-sciences-tops-13-billion-valuation-with-new-nvidia-backing-2025-10-14/#:~:AI%20...

[40] https://www.reuters.com/business/healthcare-pharmaceuticals/cadence-unveils-new-nvidia-based-supercomputer-it-pushes-into-engineering-2025-05-07/#:~:Caden...

[41] https://www.reuters.com/business/neocloud-crusoe-buy-400-million-worth-amd-chips-ai-data-centers-2025-06-12/#:~:Cruso...

[42] https://www.pharmaceuticalonline.com/doc/cio-survey-reveals-it-budgets-ai-priorities-for-life-sciences-companies-0001#:~:Compa...

[43] https://www.reuters.com/world/china/nvidia-forecasts-higher-revenue-china-clouds-future-2025-08-27/#:~:Despi...

[44] https://www.reuters.com/business/coreweave-offer-compute-capacity-googles-new-cloud-deal-with-openai-sources-say-2025-06-11/#:~:CoreW...

[45] https://www.reuters.com/business/vantage-data-centers-plans-25-billion-ai-campus-texas-2025-08-19/#:~:Texas...

[46] https://apnews.com/article/0b3f4fa6e8d8141b4c143e3e7f41aba1#:~:OpenA...

[47] https://www.reuters.com/sustainability/climate-energy/nvidia-hpe-build-new-supercomputer-germany-2025-06-10/#:~:Nvidi...

[48] https://www.reuters.com/sustainability/climate-energy/nvidia-hpe-build-new-supercomputer-germany-2025-06-10/#:~:devel...

[49] https://www.reuters.com/business/healthcare-pharmaceuticals/cadence-unveils-new-nvidia-based-supercomputer-it-pushes-into-engineering-2025-05-07/#:~:The%2...

[50] https://www.tomshardware.com/tech-industry/artificial-intelligence/legendary-gpu-architect-raja-koduris-new-startup-leverages-risc-v-and-targets-cuda-workloads-oxmiq-labs-supports-running-python-based-cuda-applications-unmodified-on-non-nvidia-hardware#:~:2025,...

[51] https://www.tomshardware.com/tech-industry/artificial-intelligence/nvidia-signs-usd1-5-billion-deal-with-cloud-startup-lambda-to-rent-back-its-own-ai-chips-18-000-gpus-will-be-leased-over-4-years-as-lambda-gears-up-for-its-ipo#:~:2025,...

[52] https://www.itpro.com/infrastructure/inside-isambard-ai-the-uks-most-powerful-supercomputer#:~:Desig...

[53] https://www.reuters.com/business/energy/us-utilities-grapple-with-big-techs-massive-power-demands-data-centers-2025-04-07/#:~:nearl...

[54] https://www.whaleflux.com/blog/ai-and-machine-learning-in-healthcare-faster-innovation-lower-gpu-costs/#:~:AI%20...

[55] https://www.reuters.com/world/china/citigroup-forecasts-big-techs-ai-spending-cross-28-trillion-by-2029-2025-09-30/#:~:This%...

[56] https://www.reuters.com/business/healthcare-pharmaceuticals/us-biotech-nabla-bio-japans-takeda-expand-ai-drug-design-partnership-2025-10-14/#:~:U,sta...

[57] https://www.reuters.com/business/healthcare-pharmaceuticals/astrazeneca-signs-up-555-million-deal-with-us-based-algen-develop-gene-therapies-2025-10-06/#:~:taps%...

[58] https://www.reuters.com/markets/deals/biotech-firm-recursion-buy-smaller-peer-exscientia-688-million-2024-08-08/#:~:its%2...

[59] https://www.reuters.com/business/healthcare-pharmaceuticals/us-biotech-nabla-bio-japans-takeda-expand-ai-drug-design-partnership-2025-10-14/#:~:Nabla...

[60] https://www.reuters.com/business/ai-lab-lila-sciences-tops-13-billion-valuation-with-new-nvidia-backing-2025-10-14/#:~:AI%20...

[61] https://www.reuters.com/technology/artificial-intelligence/behind-500-billion-ai-data-center-plan-us-startups-jockey-with-tech-giants-2025-01-23/#:~:This%...

[62] https://www.axios.com/2024/10/31/biden-ai-summit-biotech-data#:~:The%2...

[63] https://www.axios.com/2025/01/02/quantum-computing-biotech-pharma-drug-development#:~:prima...

[64] https://www.tomshardware.com/tech-industry/artificial-intelligence/elon-musk-says-xai-is-targeting-50-million-h100-equivalent-ai-gpus-in-five-years-230k-gpus-including-30k-gb200s-already-reportedly-operational-for-training-grok#:~:2025,...

[65] https://www.reuters.com/sustainability/climate-energy/nvidia-hpe-build-new-supercomputer-germany-2025-06-10/#:~:Nvidi...

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 including Scilex Holding Company (SCLX) and leading CROs across North America.

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.