IntuitionLabs

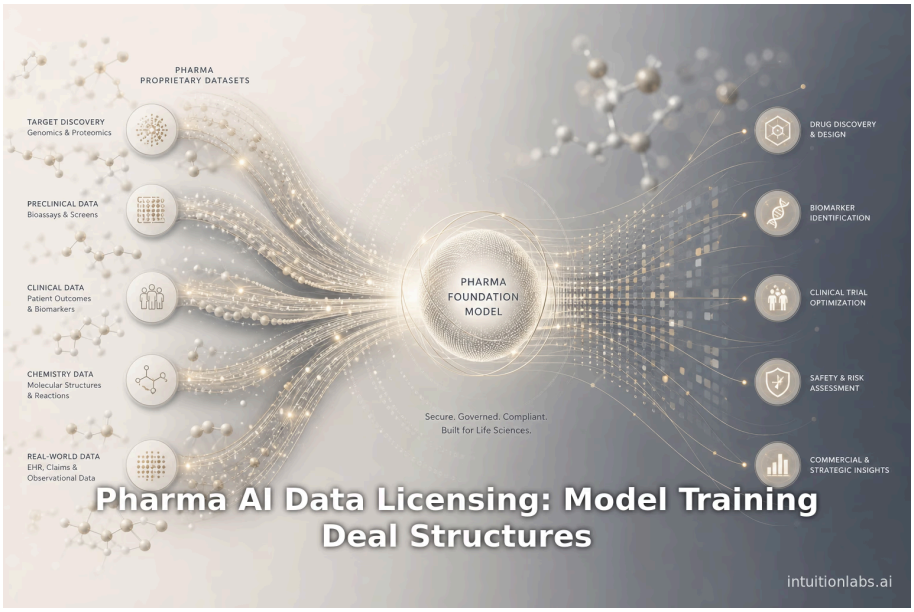
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Pharma AI Data Licensing: Model Training Deal Structures

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-06-15 · pharma data licensing · ai foundation models · drug discovery · deal structures · proprietary data · model training · biomedical ai · biotechnology

Original article: https://intuitionlabs.ai/articles/pharma-ai-data-licensing-deal-structures



In this article

1. Executive Summary
2. Introduction and Background
3. 1. Pharma Data as an AI Asset
4. 1.1 Types of Proprietary Pharma Data
5. 1.2 Why Pharma Data Matters for AI
6. 1.3 Challenges with Proprietary Data
7. 2. Foundation Models in Drug Discovery
8. 2.1 What Are Foundation Models?
9. 2.2 Why Pharma Data is Crucial for Foundation Models
10. 2.3 Foundation Model vs Traditional Models
11. 3. Deal Structures: Foundation-Model Training and Data Licensing
12. 3.1 Foundation-Model Training Agreements
13. 3.2 Pure Data Licensing Agreements
14. 3.3 Platform/Services for Data Tokenization
15. 3.4 Structured Collaboration Agreements
16. 3.5 Comparison of Deal Structures
17. 4. The Data-Licensing Playbook
18. 5. Case Studies: In-Depth Examples
19. 5.1 Incyte–Genesis Molecular AI
20. 5.2 AstraZeneca–Tempus–Pathos AI (Cancer Foundation Model)
21. 5.3 GSK–Noetik (Foundation Model Licensing)
22. 5.4 Recursion Pharmaceuticals – Tempus (Data License)
23. 5.5 Merck – Mayo Clinic
24. 5.6 CorpusAnalytiX – Ariana Pharma
25. 6. Analysis: Implications, Trends, and Evidence
26. 6.1 Market Trends and Statistics
27. 6.2 Expert Opinions and Analyses
28. 6.3 Case Study Outcomes to Date
29. 6.4 Risks and Considerations
30. 7. Future Directions and Implications
31. 8. Conclusion
32. References

Executive Summary

Pharma companies are increasingly recognizing their proprietary R&D data as a valuable asset in the era of **AI-driven drug discovery**, especially for training large-scale “foundation” AI models. Recent high-profile deals illustrate this trend. For example, Incyte’s expanded collaboration with Genesis Molecular AI has Incyte, as the paying partner, **share proprietary experimental data with Genesis to enhance its GEMS foundation-**

model platform (www.businesswire.com). In these agreements, pharma firms pay substantial upfront fees, milestones, and royalties in exchange for advanced AI capabilities built on their data. Incyte's genomics data-for-AI agreement involved $120M upfront (cash + equity) and potential billons in milestones and royalties (www.businesswire.com).

Similarly, AstraZeneca and Tempus announced a **$200M data-licensing deal** to build a multimodal oncology foundation model using Tempus's de-identified oncology data (www.tempus.com). GSK licensed Noetik's *cancer "virtual cell" foundation models* for $50M in upfront and near-term milestone payments (plus annual license fees), recognizing that "foundation models are only as good as the underlying training data" (www.businesswire.com). Other examples include Merck's collaboration with Mayo Clinic to access multimodal clinical/genomic data (www.merck.com), and Recursion Pharmaceuticals' $160M data licensing deal with Tempus for AI-ready clinical and molecular data (www.streetinsider.com).

These deals mark a **new model of pharma–AI collaboration**. Instead of traditional R&D partnerships focused on co-developing a specific drug, companies are now entering *data-licensing and model-training agreements*. In such arrangements, the pharma company grants a data/AI partner rights to use its proprietary datasets to train large AI models (the "foundation models"), while retaining rights to any resulting products. In the headline deals below, however, the pharma or biotech partner is generally the payer – licensing data or models from a vendor (AstraZeneca–Tempus, GSK–Noetik, Recursion–Tempus) or paying an AI company while contributing its own data as an input (Incyte–Genesis) – rather than a seller earning revenue from its own data. Payment structures typically include substantial upfront fees, milestone payments tied to drug development outcomes, and even equity stakes (as in Incyte-Genesis (www.businesswire.com)).

This report provides an in-depth analysis of this emerging paradigm – "pharma proprietary data as an AI asset." We describe the historical context and motivations; the types of data involved (chemical, biological, clinical, etc.); the structure of deals for foundation-model training; and the general "playbook" for licensing data to AI partners. We examine multiple case studies (Incyte–Genesis, AZ–Tempus–Pathos, GSK–Noetik, Recursion–Tempus, Merck–Mayo, etc.), present data and market statistics, and discuss regulatory, strategic, and future implications. We find converging expert views that **high-quality proprietary data is among the most valuable inputs for molecular AI** (www.businesswire.com), but that explicit licensing frameworks are needed to capture value. The report concludes with implications for pharma R&D strategy and biotech innovation as skilled stakeholders build scalable, AI-accelerated discovery pipelines using data as both fuel and commodity.

Introduction and Background

The intersection of **pharmaceutical R&D** and **artificial intelligence (AI)** has grown rapidly. New drug discovery has always generated massive datasets – from **high-throughput screening, medicinal chemistry, clinical trials, genomics, and real-world evidence** – but historically most of that data remained siloed within each company. In recent years however, advances in AI (especially in **large "foundation" models** and generative AI) have created hunger for large, diverse biomedical datasets. *Proprietary pharma data* is now being treated as a valuable **AI asset** that can speed up discovery if properly leveraged.

In parallel, the idea of “foundation models” – broadly, **big AI models pre-trained on diverse data and then fine-tuned for specific tasks** – has exploded. Originally popularized by systems like **GPT-3/4** in language, this concept extends to biology and chemistry. Models like **AlphaFold** (for protein structure) and multi-modal biotech models exemplify this trend. Building such models requires huge, high-quality datasets spanning many modalities (sequences, structures, images, assay readouts, etc.). Pharma companies sit on exactly such datasets from decades of R&D.

This confluence has given rise to novel **collaboration and licensing models**. Rather than simply buying an AI startup or embarking on a typical co-development deal, major pharmaceutical firms are now licensing their data or models to each other or to AI-specialist firms. In particular, recent *foundation-model training deals* involve pharma providing proprietary data to train AI models, and *data-licensing agreements* allow AI companies to use the data. The most cited early example is **Incyte–Genesis**: In 2025, Incyte signed an AI collaboration paying $30M up front (and potential ~$885M total) to use Genesis’s AI platform (www.fiercebiotech.com) (www.worldpharmaceuticals.net); this expanded in 2026 to a new deal with $120M up front and over $1B in later payments (www.businesswire.com).

These developments have drawn comparisons to legacy pharma–biotech licensing (e.g. J&J and biotech deals) but with a twist: the *payload* is data and AI models, not just molecules or patents. Experts note that “AI has the potential to redefine how we discover [drugs]… by combining deep expertise and significant experimental data with AI capabilities” (www.businesswire.com). In practice, Genesis’s CEO Evan Feinberg called high-quality proprietary data “among the most valuable inputs for advancing molecular AI” (www.businesswire.com).GSK’s AI head Kim Branson similarly observed that “foundation models are only as good as the underlying training data they are built upon” (www.businesswire.com).

Thus, pharma data is now being valued not only for traditional IP but as feedstock for AI acceleration. This report dives deep into this transition. We begin by surveying the **types of proprietary pharma data** and why they matter to AI (Section 1). Next, we define the **foundation-model paradigm** in drug discovery and explain why training on rich pharmaceutical datasets is compelling (Section 2). We then examine **deal structures** currently used: how companies negotiate data-for-AI collaborations, structuring payments, IP rights, exclusivity, and so on (Section 3). We provide a **“data-licensing playbook”** (Section 4) summarizing best practices and pitfalls. Throughout, we analyze **case studies** – Incyte–Genesis, AstraZeneca–Tempus–Pathos, GSK–Noetik, Recursion–Tempus, Merck–Mayo, and others – drawing lessons on terms and strategies. We also review **market data and expert commentary** on this trend. Finally, we discuss regulatory, ethical, and future considerations before concluding.

1. Pharma Data as an AI Asset

Pharma research generates **vast proprietary datasets**, ranging from **chemical libraries and assay results** to **biological omics data** and **electronic health records**. For years these data remained internal, used mainly to inform individual programs. Now, AI developers view such datasets as treasure troves to train *foundation-scale models*. Understanding this requires cataloguing what pharma data exists and why it’s valuable.

1.1 Types of Proprietary Pharma Data

- **Chemical and Screening Data**: Large pharma often screens **millions of compounds** against targets. The structures and activity results (IC50, etc.) form extensive datasets. Some companies also have proprietary **natural product libraries** or **fragment libraries**. Inventory of known molecules, including synthesized analogs.
- **Medicinal Chemistry Data**: Detailed SAR (Structure-Activity Relationship) data tying chemical modifications to biological activity. Proprietary chemical series and optimization histories.
- **Protein and Biology Data**: Internal experimental data on protein targets, including binding affinities, crystallography slides, protein-protein interaction networks, and phenotypic screens.
- **Genomics / Omics**: Patient-derived datasets, including **genomic sequences** (from trials), **transcriptomics**, **proteomics**, etc., linked to drug outcomes. Many companies have performed whole-genome or exome sequencing of trial cohorts.
- **Preclinical and Clinical Trial Data**: Detailed results from animal studies and clinical trials. This includes safety, PK/PD, biomarker changes, patient response subpopulations, etc. Such data is highly sensitive but immensely rich.
- **Real-World Evidence (RWE)**: De-identified patient data from hospitals, claims, and registries. For diseases with high prevalence, accumulated RWE (e.g. outcomes for 10,000 heart failure patients on Drug X) is proprietary.
- **Imaging Data**: Proprietary medical images (MRI, CT scans, pathology slides) collected in trials, often annotated with outcomes.
- **Procurement & Manufacturing Data**: Data on process chemistry, yields, stability. Less used for AI discovery but still proprietary know-how.
- **Literature/Doc Data**: Internal documents, patents, unpublished reports that could be text-mined (though rarely licensed as-is).

“Proprietary experimental data” thus spans all internal findings not publicly disclosed. These raw data are generally more up-to-date and higher quality than public data – e.g. public databases contain terabases of gene sequences, but pharma has bespoke datasets (e.g. patient multi-omics linked to drug response) unmatched in scale and annotation.

1.2 Why Pharma Data Matters for AI

The appeal of these data for AI lies in *scale, diversity, and specificity*. Modern AI models, especially **deep learning** and generative models, often require **massive diverse training data** to be effective (the “more you train, the better” principle (www.bristows.com)). Many **foundation models** in NLP or vision succeeded chiefly due to scale (e.g. GPT-3’s 175B parameters on 45TB of text). In drug discovery, models need equally broad training sets: e.g., IBM’s research notes training on **over a billion molecules** and multimodal biomedical data to build “foundation models” for proteins and small molecules (research.ibm.com).

Pharma data is especially valuable because it is both **high-quality** and **highly targeted**. Public chemical or biological databases (ChEMBL, PubChem, etc.) cover many activities but suffer from noise and bias. Proprietary pharma data often comes from carefully controlled studies and in-house assays, providing clearer

signals. As one analysis notes, **“AI-driven research is fundamentally only as useful as the datasets to which the AI is applied… [it] has to be substantial in size, rich and diverse, and…high quality. You truly do ‘get out what you put in’” (www.bristows.com).** Companies thus seek datasets where each datum (e.g. an assay result or patient record) has been generated under rigorous bioassays or clinical protocols.

Moreover, specialized pharma data fills gaps in public corpora. For example, no public dataset may have the exact combination of genomic variant, expression profile, and drug response that a pharma’s cancer trial data does. By providing these **unique in-house data**, pharma companies can customize AI models to their own pipelines and targets. Pharma’s data often covers disease areas they focus on (oncology, immunology, etc.), so foundation models trained on that data can be tuned to those domains.

Finally, there is a competitive aspect: historically, data was a **secret weapon**. Today, firms realize they can monetize it. The biotech community now sometimes talks of data as *“the new oil”* of pharma AI. In essence, high-quality pharma data serves as fuel for the next generation of AI tools. Indeed, Genesis CEO Evan Feinberg (formerly of Google/DeepMind) says that partnership with Incyte “marks an important moment in the evolution of AI in this vertical. **High-quality proprietary data is among the most valuable inputs for advancing molecular AI**… enabling… an industrial-scale flywheel of AI-enabled design-make-test cycles” (www.businesswire.com). This highlights the view that pharma data has intrinsic value in AI pipelines.

1.3 Challenges with Proprietary Data

While valuable, pharma data comes with challenges. Data is often siloed in disparate databases with proprietary formats. Combining multimodal data (e.g. linking assay sheets, genomic data and clinical outcomes) is difficult. Data may contain sensitive patient information (even if de-identified it must comply with HIPAA/GDPR). Quality varies by assay/campaign. Gains from licensing data must outweigh privacy and competitive risks. As Bristows LLP notes, “We need to ensure that the value in the data is protected” when licensing – typically via **limited-use licenses and confidentiality obligations** (www.bristows.com). Indeed, restrictions often include forbidding re-sharing or cross-use with other data. Also, “the main value lies in the results generated from the data, not the raw data itself… the ownership of new IP generated is of key importance” (www.bristows.com). Thus, data deals carefully negotiate IP rights on derived models or improvements.

Despite the hurdles, companies are creating ecosystems to unlock value. Data brokers and partnerships (e.g. CorpusAnalytiX/Ariana) are emerging to aggregate and syndicate data while respecting suppliers’ control (www.arianapharma.com). Moreover, federated learning initiatives like MELLODDY (EU-wide effort among 10 pharma firms) tackle data-sharing without centralizing raw data. But even federated/federated approaches require trust and standardization.

Key takeaway: Proprietary pharma data – experimental results, genomic records, patient outcomes, etc. – is an *increasingly strategic asset* for AI. Companies now seek to transform these data into foundation-model training inputs through structured collaborations. The remainder of this report explores how these collaborations are structured and what frameworks govern them.

2. Foundation Models in Drug Discovery

2.1 What Are Foundation Models?

A **foundation model** is a large-scale AI model pre-trained on broad and diverse data, which can then be fine-tuned or prompted for many downstream tasks (chipp.ai). In NLP, GPT-3/4 are classic examples: trained on vast text corpora, they can be specialized for translation, summarization, etc. In biomedicine, the analogue is models like AlphaFold 2 (trained on millions of protein structures and sequences) which can predict structures of unseen proteins. More generally, foundation models for drug discovery include multimodal neural networks trained on enormous biological and chemical datasets. IBM Research describes biomedical foundation models as "trained on diverse biomedical data (antibody-antigen interactions, small molecule-protein interactions, etc.)… transforming this field" (research.ibm.com). These models can generate and evaluate novel molecules, predict properties (potency, ADME), and identify targets – far beyond traditional QSAR or docking.

Important features of foundation models **in pharma**:

- **Scale:** often billions of parameters, requiring petabytes of training data.
- **Diversity:** trained on *multimodal* inputs (chemical structures, genomics, imaging, clinical text) to capture broad biochemistry. For example, AstraZeneca's project with Tempus aims to build an oncology model from text, omics, and imaging (pharmaphorum.com).
- **Transferability:** well-trained models can be fine-tuned on specific tasks (like optimizing a series of molecules) with relatively little new data. Thus the base "foundation" can serve many discovery projects.

Domain-specific teams apply multi-modal GNNs, autoencoders, or diffusion models to drug design.

Example: IBM's Biomedical Foundation Models

IBM Research's Biomedical Foundation Models (BMFMs) project exemplifies this approach. Their targets-discovery FM is trained on genomics and gene-expression data to predict disease-relevant genes; their molecule discovery FM is trained on billions of small molecules and proteins. IBM notes these FMs "widen the search scope for novel molecules and refine it to eliminate unsuitable ones, emphasizing the detailed nuances in molecular structure and dynamics" (research.ibm.com). In practice, these BMFMs are created by pooling massive internal and public datasets.

2.2 Why Pharma Data is Crucial for Foundation Models

By their nature, foundation models benefit from *diversity and volume of training data*. Public datasets alone often lack certain classes of molecules or complex phenotypes. Proprietary pharma data fills these gaps. For example, Microsoft and AstraZeneca noted that deep, curated assays in AZ's pipeline could train models that generalize to advanced targets.

As Kim Branson of GSK states:

> "Foundation models are only as good as the underlying training data they are built upon. Noetik's approach to generating high-quality spatial data at scale to train foundation models is novel… [and] has the potential to deepen our understanding of biology" (www.businesswire.com).

This underscores that *the pharma-company's own high-quality experimental data* can make a foundation model much more powerful for drug discovery than relying solely on generic datasets. Incyte's leadership echoed this view: "by combining… our significant experimental data with Genesis' AI capabilities, we aim to more efficiently advance priority programs against high-value targets" (www.businesswire.com).

The synergy is clear: pharma labs produce detailed results (e.g. many measured IC50s on novel targets, or multi-omics profiles in a disease) that an AI model can learn from. The model can then suggest new molecules or hypotheses, which pharma can rapidly test. In effect, the **data generates its own feedback loop** (often called a "flywheel") where more experiments feed better AI, which drives more experiments. Genesis's CEO said their mission was "to unlock a new era of agentic drug design," using proprietary data to drive an "industrial-scale flywheel of AI-enabled design-make-test cycles" (www.businesswire.com).

2.3 Foundation Model vs Traditional Models

Traditionally, pharma ML involved **narrow models**: e.g. a model trained on one target's assay data to predict IC50. In contrast, foundation models trained on massive, possibly unlabeled data (using self-supervised learning) can handle **multiple targets and tasks**. They can generate candidates, predict off-target interactions, or translate between modalities (e.g., gene expression to drug response).

For example, a foundation model for small molecules might be pre-trained on all internal medicinal chemistry data from multiple projects; fine-tuning it on a new target then requires less data. This is analogous to GPT-4's success in few-shot learning.

Many big pharma are now welding foundation-model tech onto their pipelines. AstraZeneca's Pathos AI consortium aims to create a **multimodal oncology model** from Tempus's patient data (www.tempus.com). Sanofi, TrialSpark (Formation Bio) and OpenAI explicitly announced a collaboration to "provide access to proprietary data to develop AI models" with the goal of an **AI-powered drug development platform** (www.sanofi.com). These moves signal that pharma sees broad pre-training as essential to extract maximal value from their data.

To summarize, **foundation models in drug discovery** are:

- **Broadly trained** on diverse biomedical data (genomes, chemistries, images, etc.) (research.ibm.com).
- **Adaptable** to multiple drug R&D tasks (target identification, molecule design, ADMET prediction, etc.).
- **Accelerated** by access to large proprietary datasets.

Training such models is computationally intense (requiring GPUs/TPUs, often provided by tech partners), but pharma companies increasingly consider it worthwhile. The companies' own press release characterizes the Incyte–Genesis expanded deal as "one of the first major Pharma-AI collaborations to power large-scale foundation model training with a partner's proprietary experimental data" (www.businesswire.com).

3. Deal Structures: Foundation-Model Training and Data Licensing

Pharma–AI collaborations typically blend elements of **licensing**, **outsourcing**, and **co-development** deals. We categorize key models below:

3.1 Foundation-Model Training Agreements

These deals focus on using pharma's data to train a large AI model (the "foundation model") that will serve both partners. Typically, the pharma company grants the AI developer rights to use specified datasets (often securely and exclusively for model training). The output (the model and any IP from it) is then licensed back to the pharma company (often exclusively for internal development).

Key features may include:

- **Upfront Payment & Resources:** The pharma pays an upfront fee (cash, equity, or compute credits) and may fund ongoing compute costs. E.g. Incyte paid $80M cash + $40M equity to Genesis (www.businesswire.com). The AI partner usually provides the technical platform and models (like Genesis's GEMS).
- **Milestone Payments & Royalties:** Similar to drug-collaboration deals, contingent on research and development (R&D) milestones (preclinical, clinical) and product sales. In Incyte–Genesis, potential milestones exceed $1B for five targets (www.businesswire.com), and royalties on sales are also specified.
- **Exclusivity Rights:** Incyte's deal gave Incyte exclusive rights to any drugs developed. Conversely, Genesis may retain rights to underlying methods. In the Noetik–GSK model deal, GSK got a **non-exclusive** license to Noetik's foundation models (www.businesswire.com), reflecting a subscription approach.
- **Data Usage Rights:** The pharma's proprietary data (assay results, etc.) is licensed for training only, typically under strict confidentiality. It may *not* give ownership of raw data – the company retains the data. Instead, the AI firm may own improved models or algorithms (subject to negotiations).
- **Return of Results:** The resulting model (or features derived from it) is provided to the pharma for its internal use. For example, Genesis's GEMS model (improved via Incyte's data) will be used by Incyte's R&D teams.
- **Equity:** Often, AI startups may take small equity from the pharma partner (as with Genesis) or vice versa. In Incyte–Genesis, Genesis took a $40M equity investment from Incyte (www.businesswire.com), aligning incentives.
- **Governance and Collaboration:** Joint steering committees to choose targets/data and monitor progress. These deals usually cover multiple programs (Incyte-Genesis: 5+ targets to start (www.businesswire.com)).
- **Term and Termination:** Usually multi-year (often ~5-7 years) to allow model development and drug R&D. Early termination may incur penalties or require buyouts of commitments.

Case Study: Incyte–Genesis (Foundation Model Collaboration)

- **Initial deal (Feb 2025):** Incyte paid **$30M upfront** (to Genesis) for discovery collaboration on 2 targets (www.fiercebiotech.com). It granted Incyte rights to use Genesis's GEMS platform on those targets, with up to $295M in milestones *per target* (www.fiercebiotech.com). Incyte also obtained exclusive drug rights.
- **Expanded deal (May 2026):** After generating encouraging results, Incyte expanded the collaboration. Now Genesis will get $80M cash + $40M equity (= $120M) upfront (www.businesswire.com), plus recurring research funding. They'll cover at least 5 targets (with options for more, and Incyte sole commercialization). Genesis can earn up to $232M in development, regulatory and sales milestones per program, and over $1B if all milestones are achieved across the five initial targets (www.businesswire.com), with even more possible if more targets are added.

The Incyte–Genesis agreements explicitly describe **"fast-training data usage"**: Incyte will **securely share proprietary experimental data to enhance Genesis' foundation models** (www.businesswire.com). The press release highlights that this is *"one of the first major pharma-AI collaborations to power large-scale foundation model training with a partner's proprietary experimental data"* (www.businesswire.com).

3.2 Pure Data Licensing Agreements

In some cases, a company may simply **license datasets** to an AI partner (who may then train models on it), without a defined discovery collaboration. The pharma's data, possibly anonymized, is provided under contract for specified uses. This is often done when the AI company already has its own platform or model, and just needs high-quality data.

Example: On November 3, 2023, Recursion Pharmaceuticals (a biotech) entered a "Tempus Agreement" with Tempus Labs to license Tempus's de-identified clinico-molecular database (www.streetinsider.com). Under this agreement:

- Recursion gets the right to use Tempus's **proprietary clinical and molecular data** to develop and train its own ML models (www.streetinsider.com).
- Recursion paid an **initial $22M license fee** plus annual fees ($22–42M) for up to $160M over 5 years (www.streetinsider.com). The license fees are payable, at Recursion's option, in cash, Class A common stock, or a combination of the two.
- Recursion also gained access to Tempus's analytics platform (LENS) for data exploration (www.streetinsider.com).
- This deal did not directly involve co-developing drugs; rather, Recursion can leverage the data in its AI drug discovery work. Termination clauses and usage restrictions (only for "therapeutic development purposes") are spelled out (www.streetinsider.com).

This exemplifies a **data-as-subscription** model: pharma (or a pharma-backed biotech) treats data as a licensed service. Cronus Data, Datavant, and others pursue similar data marketplaces. Notably, Recursion's SEC filing highlights that "Permitted Uses include… train, improve, modify, and create derivative works of Company's machine [learning]" (www.streetinsider.com), confirming that training AI is within scope.

3.3 Platform/Services for Data Tokenization

A variant in this space is licensing *technology* that enables new uses of pharma data, rather than the data itself. For example, Datavault AI's 2025 license to Scilex grants Scilex exclusive rights to Datavault's AI-driven **tokenization platform** for genomic and drug data (www.sec.gov). Under this agreement:

- Datavault gave Scilex a worldwide license (with sublicensing) to its platform and patents to "create and operate a Biotech Exchange platform" for genomics and pharma data (www.sec.gov).
- Scilex paid $10M upfront and may owe up to $2.55B in milestones/royalties based on sales of the tokenized assets (www.sec.gov). This is less a direct pharma partnership than an attempt to create a data marketplace using blockchain. It shows that even broader **data-tokenization and monetization** ideas are attracting huge valuations. (We do not count this as a typical pharma-AI R&D deal, but it highlights the growing value placed on healthcare data.)

3.4 Structured Collaboration Agreements

Many pharma–AI deals are hybrids combining elements of licensing, CRO services, and co-investment. They often follow the "drug discovery alliance model" with adaptation for AI. Key points in structuring such deals include:

- **Scope (Targets/Programs):** Define disease indication or target family. Incyte initially defined two targets (www.fiercebiotech.com), then expanded to five. Option clauses let the pharma add targets.
- **Data Sharing:** Specify exactly what data is shared, in what format, and under what security. For instance, Incyte says it will "securely [use] data to train" Genesis's platform (www.businesswire.com). Often, pharma data is uploaded to a secure enclave or compute in place.
- **Term and Exclusivity:** Usually longer term (e.g. 5 years or until patents expire). Exclusivity clauses cover target areas. In the Tempus–Recursion deal, Recursion may not assign the agreement to a large pharmaceutical company without Tempus's consent.
- **Payment Structure:** As noted, it's common to have a moderate upfront fee, plus heavy milestone and royalty potential. For foundation-model deals, payments may be structured more like R&D payments than pure license fees, since outcomes are uncertain. (E.g., Genesis can earn over $1 billion if all milestones are achieved across the five initial targets, provided aggregate peak annual net sales of the five products exceed specified milestones (www.businesswire.com).)
- **Equity and Options:** Sometimes deals include equity (Incyte's $40M equity purchase in Genesis) to align interests. Licensing agreements may also include exclusivity termination if milestones or payments aren't met (Datavault–Scilex required minimum royalty payments or lose exclusivity (www.sec.gov)).
- **Governance:** Joint steering committee to set AI training priorities, architecture, and to review milestones.
- **Model Governance:** Clarify who owns resulting IP. Often the AI company claims IP on its platform and resulting model, while pharma gets rights to use results and learns. The balance is negotiated case-by-case (often mirroring traditional biotech co-licensing deals in spirit).

3.5 Comparison of Deal Structures

Category	Foundation-Model Training Deal	Data Licensing Deal	Tokenization/Platform Deal
Purpose	Joint R&D: train AI model on pharma data for drug discovery.	License raw data for AI use (e.g. model training).	License tech/platform to manage data as assets.
Partner Roles	Pharma provides data + funds; AI co develops models.	Data provider (pharma/healthcare org) vs. AI developer (data consumer).	Tech provider (e.g. Datavault) vs. pharma or biotech user (tokenizer).
Upfront Payment	Large lump sum (sometimes part cash, part equity) – e.g. $80M+ (www.businesswire.com).	Moderate (e.g. Recursion $22M) (www.streetinsider.com).	Usually lower (e.g. Datavault $10M) (www.sec.gov).
Milestones & Royalties	Very high potential. Incyte–Genesis:>$1B+ for 5 targets (www.businesswire.com).	Based on usage or revenue (Recursion: $160M max over 5 yrs (www.streetinsider.com)).	Potentially huge (Scilex up to $2.55B on token sales) (www.sec.gov).
Equity	Common (e.g. pharma equity in AI co) – aligns incentives (www.businesswire.com).	Sometimes (Recursion may pay Tempus in cash or stock at its option) (www.streetinsider.com).	Rare, usually license fee only.
Exclusivity	Often exclusive rights to model outputs. Pharma may have exclusive commercialization rights.	Restricted to authorized use – often non-transferable.	Exclusive rights to use tech in field (e.g. biotech industry).
Data Ownership	Pharma retains raw data rights; AI co owns derivative models (with license back).	Pharma retains data; AI co gets a “license” to use it.	Pharma retains data; tech co retains platform tech.
Example	Incyte–Genesis (AI model training) (www.businesswire.com); AstraZeneca–Tempus–Pathos (www.tempus.com).	Recursion–Tempus (data license) (www.streetinsider.com).	Datavault–Scilex (tokenization tech) (www.sec.gov).

Table 1: Typical structures of pharma–AI data/model deals, with illustrative terms.

These deal types are **not mutually exclusive**. Complex collaborations may blend elements (e.g., pharma funds AI R&D, gets data back, plus licenses in tech platform).

4. The Data-Licensing Playbook

Pharma companies entering the AI-data licensing arena should follow strategic guidelines. Based on emerging practice and expert advice (www.bristows.com), a “playbook” for pharma data licensing might include:

- **Identify High-Value Data Assets:** Catalog datasets (bioassays, clinical, genomics) that could train useful AI. Consider uniqueness (e.g. a large private cancer-patient genomics dataset). Evaluate quantity, quality, and strategic importance.
- **Define Objectives and Scope:** Decide what goals to achieve (improve internal R&D, create joint products, or purely monetize data). Choose partner(s) accordingly. For example, a foundation-model partnership (like Incyte–Genesis) aims at internal drug discovery acceleration (www.businesswire.com),

whereas a licensing to an AI startup might be revenue-driven.

- **Partner Selection:** Look for AI collaborators with complementary expertise. This could be established AI firms (OpenAI, Google), specialized biotech-AI companies (Genesis, Noetik), or data aggregators (Recursion, Tempus, CorpusAnalytix). Each partner offers different assets – e.g. Genesis had a strong generative chemistry platform (www.fiercebiotech.com), while Tempus had unmatched oncology patient data (www.tempus.com).
- **Structuring the Agreement:** Craft a deal balancing risk and reward. Key considerations include:
- **Upfront vs Contingent:** Upfront fees secure partner commitment and data access rights, but pharma may prefer more contingent (milestone-based) payments if outcomes are uncertain (www.bristows.com). Incyte's deal used both: $120M upfront plus milestones (www.businesswire.com).
- **Milestones and Royalties:** Set milestones tied to concrete R&D events (e.g. IND, pivotal trials). Royalties on product sales can share upside but need careful cap/terms. Notice that Genesis is eligible for over $1B across the five initial targets (up to $232M per program), with further payments if additional targets are nominated (www.businesswire.com).
- **Data Rights:** Clearly define licensed data (types, scope, duration). Include strong confidentiality protections. Bristows advises limiting uses and forbidding combining with external data unless allowed (www.bristows.com).
- **IP Ownership:** Agree who owns AI models, data improvements, and downstream inventions. Often, pharma retains rights to any new drug IP, while AI partner keeps model IP (or vice versa if structured differently). The playbook must align these with each party's goals.
- **Exclusivity:** Decide if the partner has exclusive data rights (often yes, to incentivize investment) or if concurrent licensing is allowed. Exclusive rights typically command higher payment.
- **Governance:** Establish joint committees to oversee data use, model outputs, and compliance. Agree on reporting, audits, and review procedures.
- **Regulatory/Privacy Compliance:** Ensure de-identified clinical data complies with HIPAA/GDPR. Obtain all necessary patient consents for secondary use. Automate audits if possible.
- **Termination Terms:** Define consequences if targets or payments are not met. E.g., Datavault's license reverts to non-exclusive if Scilex fails minimum royalties (www.sec.gov). Early termination clauses, while standard, can include penalties or intellectual property releases.
- **Data Preparation:** In many deals, pharma must prepare data (cleaning, annotation) or allow the partner to do so. This is often funded via the deal (Incyte provided recurring research funding for compute and data preparation (www.businesswire.com)).
- **Technical Implementation:** Data often stays on a secure platform (to protect IP). Models are trained in a "flywheel" mode: Incyte's data feeds to GEMS, which is iteratively refined (www.businesswire.com).
- **Capture Value:** Companies should consider how to "share in the value" of downstream innovations. For example, even if a pharma partner doesn't want IP in AI algorithms, it may still want a share of the economics. Bristows suggests license fees, milestone payments, royalties or equity

(www.bristows.com). Incyte's deal offers Genesis over $1B in milestones across the five initial targets plus royalties on sales of any approved collaboration products, capturing part of the drug proceeds (www.businesswire.com).

- **Iterate and Expand:** Many pioneers structure deals with options to expand scope. Incyte can nominate additional targets over time (www.businesswire.com). Agreements can be extended if early success is demonstrated.

In summary, **treat data as a negotiable asset**. Approach these deals with the same rigor as traditional IP licensing or M&A. Bristows warns: while the entire data may not be sold, pharma should definitely **extract value** for it (e.g., license fees, royalties on derived products) (www.bristows.com). As one advisor puts it, ensure you *"share, in some way, in the value that [a] partner derives from commercialising the data or the resulting IP"* (www.bristows.com). This mindset underpins the new data-licensing playbook: proactive valuation, structured monetization, and legal protection.

5. Case Studies: In-Depth Examples

To illustrate these principles, we analyze several representative deals:

5.1 Incyte–Genesis Molecular AI

Initial Feb 2025 Collaboration: Incyte (a mid-sized biotech) partnered with Genesis to use the GEMS AI platform for *small-molecule drug discovery*. Incyte agreed to:

- Pay **$30M upfront**.
- Collaborate on 2 targets (with option for a 3rd).
- Grant Genesis a license to use GEMS on Incyte-chosen targets.
- Incyte secured exclusive rights to any resulting drugs.
- Genesis could get up to $295M in milestones *per target*, plus royalties (www.fiercebiotech.com).

This structure mirrors a typical R&D collaboration with AI twist: moderate upfront, high contingent payments, exclusive pharma rights. The announcement stressed fusion of **Incyte's expertise and data with Genesis's AI** (www.fiercebiotech.com).

Expanded May 2026 Agreement: Following promising early work, they expanded the deal:

- **Upfront and Equity:** $80M cash + $40M stock (total $120M) to Genesis (www.businesswire.com).
- **Compute Funding:** Incyte provides ongoing funding for AI compute workloads (www.businesswire.com).
- **Targets:** Expanded to at least five initial targets, with rights to add more (www.businesswire.com).
- **Milestones:** Up to $232M *per program* (development, regulatory and sales milestones) and over $1B total if all milestones are achieved across the five initial targets (www.businesswire.com). Additional billions possible for extra targets.
- **Royalties:** Yes, on drug sales.

- **Rights:** Incyte retains exclusive development/commercial rights to collaboration products (www.businesswire.com).
- **Quotes:** Incyte touted combining “deep expertise and significant experimental data” with Genesis’s AI to speed high-value programs (www.businesswire.com); Genesis called hi-quality data “among the most valuable inputs” to an AI-driven drug-design flywheel (www.businesswire.com).

This expansion signifies Incyte’s confidence and Genesis’s standing: a big bet on AI. (Fierce reports the total potential value could reach ~$885M + more (www.fiercebiotech.com), with Incyte eyeing “exotic” early-stage tech (www.fiercebiotech.com).)

Deal Significance: Many observers call this a **paradigm-setting deal**. Here Incyte is the payer, not a data seller: it pays Genesis $80M in cash and buys $40M of Genesis equity, and it also shares proprietary experimental data to enhance Genesis’s GEMS platform. The deal also includes high milestones and a multi-program pipeline – analogous to a big biotech licensing deal, but traded for AI services. Separately, Noetik’s Chief Business Officer said its GSK deal “defines a new asset class in biotech” (www.businesswire.com) – a notion born from this space.

5.2 AstraZeneca–Tempus–Pathos AI (Cancer Foundation Model)

In April 2025, AstraZeneca announced a **three-way collaboration** with Tempus and Pathos AI to build a **multimodal oncology foundation model** (www.tempus.com). This is exemplary for clinical/AI collaborations:

- **Data Owner (Tempus):** Tempus supplies the de-identified oncology data. Under the terms, AZ and Pathos will use Tempus’s de-identified data to train a cancer foundation model (www.tempus.com).
- **AI Developer (Pathos):** Pathos AI will develop the model (using Tempus’s platform). Pathos is an AI-enabled drug-development company co-founded by Tempus founder and CEO Eric Lefkofsky, so it is a Tempus-related party rather than an unrelated third party.
- **Pharma User (AZ):** AZ will use the model to accelerate oncology R&D across its pipeline.
- **Financials:** Tempus revealed the agreement includes **$200M in data licensing and model development fees** (www.tempus.com) payable to Tempus. (It isn’t clear how this is structured, but likely a mix of upfront and installments.)
- **Shared Outcome:** The final model will be shared among AZ, Pathos, and Tempus. AZ hopes it will “increase the probability of clinical success” in diverse oncology programs (www.tempus.com).
- **Governance:** AZ’s oncology AI head highlighted this as continuing AZ’s strategic partnership with Tempus (initially announced in 2021) and a step toward precision oncology at scale (www.tempus.com) (research.ibm.com).

This deal differs from Incyte’s in that it is explicitly broad and non-exclusive (all partners use the model). It resembles a *pre-competitive consortium* within one pharma’s network. Still, the $200M price tag marks pharma paying top dollar for data-driven model-building. AZ’s insiders stressed that analyzing “vast amounts of rich data” via AI is transforming cancer R&D (pharmaphorum.com) and that Tempus invested “billions” collecting the necessary data (www.tempus.com). The scale is thus enormous, aiming to serve the entire oncology division.

5.3 GSK–Noetik (Foundation Model Licensing)

On January 8, 2026, GSK announced it would **license Noetik's foundation models** for cancer research (www.businesswire.com). Key points:

- GSK obtains a **non-exclusive license** to Noetik's "OCTO-VC" model in two indications (NSCLC and colorectal cancer) (www.businesswire.com).
- **Payment:** GSK committed **$50M in upfront and near-term milestone payments** (www.businesswire.com), structured as part of a five-year collaboration. GSK will also pay annual licensing fees (subscription model) to access updated models (www.businesswire.com).
- **Approach:** Noetik had trained "virtual cell" generative models on *its own proprietary spatial biology dataset* (hundreds of millions of cells). GSK will integrate these models into its pipeline (www.businesswire.com).
- **Pioneer Theme:** Noetik's Chief Business Officer Shafique Virani said the deal "defines a new asset class in biotech" – **licensing human foundation models** rather than just AI services (www.businesswire.com). Noetik's CEO Ron Alfa likewise described it as "a new paradigm in biotech: the licensing of human foundation models."
- **Rationale:** GSK's AI chief Branson echoed that foundation models need high-quality data; Noetik's data-driven models are seen as novel and promising (www.businesswire.com).

Unlike Incyte or AZ deals, this *licenses existing pretrained models* rather than co-building from scratch. It suggests a different structure: pharma can simply acquire models trained on other entities' data, rather than providing its own data. GSK is essentially paying for **AI infrastructure** (Noetik's models) that complements its own datasets. The non-exclusive nature and recurring fees indicate GSK is comfortable accessing, but not owning, this AI asset.

5.4 Recursion Pharmaceuticals – Tempus (Data License)

This SEC-filed agreement (Nov 2023) is a pure **data-license** and serves as an example of terms:

- **Scope:** Recursion is granted a *limited right to access Tempus's proprietary database of de-identified clinical and molecular data*, with a cap on records downloaded at any one time, an aggregate cap on unique records, and a 180-day retention period per downloaded record. Permitted uses include developing, training, improving, and creating derivative ML models (www.streetinsider.com). Essentially, Recursion can use the data to drive discovery tools.
- **Duration:** 5-year term.
- **Fees:** $22M initial, plus $22–42M/yr (total up to $160M) (www.streetinsider.com). Under the filing, the initial and annual license fees are payable at Recursion's option in cash, Class A common stock, or a combination – not a fixed equity stake.
- **Software Access:** Recursion gets Tempus's LENS data analysis software for an additional fee (www.streetinsider.com), akin to SaaS in data deals.
- **Termination:** Recursion can exit after 3 years (with fee), subject to penalties.
- **Uses:** Exclusively for internal R&D (with rights to share in collaboration as needed).

- **Implications:** Recursion essentially treats Tempus as a *data vendor*. The structure resembles a biotech licensing a specialized database or platform (like Bloomberg Terminal vs. stock data) rather than a co-development project.

This deal highlights the **data value**: Recursion was willing to commit up to $160M in license fees over the term (payable in cash or Class A stock at its option). Recursion's filings note the focus on AI/R&D use cases. The key element is the license grant for *model training* ("train, improve, modify…and create derivative works of Company's ML" (www.streetinsider.com)), confirming that mainstream drug companies now explicitly license data for AI training under contract.

5.5 Merck – Mayo Clinic

In Feb 2026, Merck (MSD) and Mayo Clinic announced a multi-year R&D partnership to integrate Mayo's extensive clinical/genomic data with Merck's AI platform (www.merck.com):

- **Data:** Merck will gain access to Mayo Clinic's de-identified multimodal data – lab results, imaging, clinical notes, and genomics – via the Mayo Clinic Platform (www.merck.com).
- **Technologies:** The collaboration uses Mayo's "Platform Orchestrate" program and analytics tools.
- **Scope:** The goal is improving Merck's drug discovery (target ID, early development) by validating AI models against real patient data with Merck's research.
- **Significance:** This is Mayo's first such wide-scale deal. Unlike explicit license deals, this is more of a joint R&D arrangement. The announcement emphasizes combining Mayo's data with Merck's virtual-cell AI ("AI-enabled virtual cell technologies" (www.merck.com)).
- **Financials:** Not publicly detailed – likely funded internally by Merck as a strategic investment.

Though a non-transactional tone, it nonetheless fits the pattern of pharma seeking proprietary health data for AI purposes. Merck's CEO underscored that integrating "high-quality clinical data and AI-enabled insights into discovery research" should improve program success (www.merck.com). In effect, Mayo's patient database – essentially a national treasure – becomes part of Merck's AI pipeline. The absence of money talk suggests either it's all internal or undisclosed.

5.6 CorpusAnalytiX – Ariana Pharma

This Oct 2024 deal involves **a data marketplace and an AI drug developer**:

- **Partners:** CorpusAnalytiX, a broker of diverse healthcare datasets (from SMEs, labs, clinics) (www.arianapharma.com); and Ariana Pharma, an AI drug discovery company.
- **Agreement:** CorpusAnalytix provides Ariana access to a range of previously siloed datasets (pathology, oncology, genomics, etc.) through its marketplace (www.arianapharma.com). Ariana will use its KEM eXplainable AI platform on these data to generate biomarkers and trial designs (www.arianapharma.com).
- **Value Sharing:** Ariana will feed its derived insights (and possibly small drug asset deals) back into the marketplace, creating a data–AI feedback loop.

- **Structure:** No money amounts are public; this is pitched as a strategic alliance rather than a pure license.
- **Notable Point:** The partnership emphasizes empowering *data suppliers* to gain revenue: “the supplier-centric model… empowers data generators (startups, labs) to maintain control over their datasets while monetizing them” (www.arianapharma.com). In effect, pharma/biotechs with data can plug into this ecosystem.

While not an Incyte-level deal, it is a practical example of the data strategy playbook: pharma (or any data owner) can monetize by contributing to a platform, receiving AI insights, and still retaining data ownership. It illustrates a **market-driven model** for licensing and exchange of diverse biomedical data.

6. Analysis: Implications, Trends, and Evidence

6.1 Market Trends and Statistics

- **Surging Investment:** The fact that pharma is committing hundreds of millions to these deals (e.g. $120M+ with Genesis (www.businesswire.com), $200M with Tempus (www.tempus.com), $50M with Noetik (www.businesswire.com)) indicates high expectations. Investors have noted that such *AI licensing deals* are creating new biotech valuations. Noetik raised a $40M Series A in 2024 to train its models, and sees the GSK deal as validating AI infrastructure licensing (www.businesswire.com).
- **Data Monetization Market:** One vendor forecast, from the market-research firm Towards Healthcare, supports growth: it projects the **global healthcare data monetization market** growing from ~$0.48B in 2024 to $1.92B by 2034 (www.towardshealthcare.com). This encompasses licensing of clinical data, analytics services, AI insights, etc. The CAGR is ~15% (www.towardshealthcare.com). Smaller analyses similarly forecast robust growth in “Healthcare Data Monetization” and “AI in Healthcare Data” markets (www.mordorintelligence.com) (www.towardshealthcare.com). While still a fraction of pharma R&D spend, these forecasts highlight a new market niche centered on data-as-asset.
- **Deal Volume:** Precise counts are hard, but dozens of announcements have come in various sizes. Besides the mega-deals above, smaller pharma (e.g. Recursion, Ariana’s pharma, etc.) and many startups are active. Industry newsletters (e.g., *Fierce Pharma/BIotech*) routinely report new “AI collaborations” each month. This suggests a rapid secular trend – likely accelerating as generative AI hype meets pharma’s R&D challenges.
- **Shift in R&D Strategy:** C-suite executives comment that post-2024, every pharma pipeline slide includes an AI-powered target hunt or generative platform. CEOs like Incyte’s then-CEO Hervé Hoppenot (since succeeded by Bill Meury) said they will invest in “new technologies we don’t have—so, early-stage exotic stuff” (www.fiercebiotech.com), referring to AI.

6.2 Expert Opinions and Analyses

- **Importance of Data Quality:** Industry consultants and AI experts uniformly stress data quality. As Bristows explains, accessible data is valuable only if it's cleaned, well-annotated, and sizable (www.bristows.com). The Stanford CRFM report that coined the term "foundation models" defines them as models "trained on broad data at scale" that are adaptable to a wide range of downstream tasks (arxiv.org). This underlies the push to mobilize pharma's high-fidelity datasets.
- **Data vs. IP:** Analysts note a philosophical shift: historically pharma guarded data as proprietary trade secrets, but now they see a revenue path. The Bristows article highlights that firms used to focus on owning drug IP and not worry about who owns algorithmic "improvements" (www.bristows.com). Now, companies "should consider whether it is appropriate… to share in the value" from their data (www.bristows.com). The Incyte and GSK deals reflect this – they ensure pharma gets value (milestones/royalties) from any resultant drug IP.
- **New Asset Class:** The Noetik–GSK deal illustrates a novel concept: *licensable AI models as assets*. Noetik's executives describe the move as one "from AI services collaborations to licensing AI infrastructure," and Noetik's CEO says that "the licensing of human foundation models" is a new paradigm in biotech (www.businesswire.com). This suggests biotech products are not just drug candidates but also trained AI models.
- **Economic Scale:** Some caution about inflated numbers. While milestone totals reach the billions, skeptics note that actual probability of success is low, so the *expected value* may be far lower. For example, Incyte's $1B+ potential milestone across five targets does not guarantee actual payments. Likewise, data markets of ~$2B by 2034 are small compared to multi-trillion pharma sales. However, experts see these deals more as ecosystem foundations than immediate profit centers.
- **Federation vs Licensing:** There is debate whether data pooling (e.g. MELLODDY) or licensing yields more value. Federated learning keeps data private but requires coordination across many parties. Licensing gives tangible returns per deal. Both approaches may coexist. For example, MELLODDY (an EU project with GSK, Merck, etc.) is not exactly a deal, but warns companies about data-sharing vs IP security. The current "foundation deals" circumvent federated complexity by using contracts.

6.3 Case Study Outcomes to Date

These deals are too recent for published outcomes, but a few notes:

- **Preliminary AI Results:** Incyte reported "strong results from initial programs" pre-expansion (www.businesswire.com), implying the models delivered useful leads. Pathos has started dosing a drug discovered via AI (though not necessarily in the Tempus collaboration) (pharmaphorum.com), hinting at tangible drug candidates. Noetik claims its models can simulate patient biology at unprecedented resolution (www.businesswire.com).
- **Stock Movements:** Markets sometimes react. Recursion's shares jumped after announcing Tempus partnership (www.streetinsider.com). Tempus itself is publicly traded; on news of the AZ/Pathos deal its stock spiked nearly 15% (pharmaphorum.com). Such moves highlight investors taking these

partnerships seriously.

- **Strategic Momentum:** The pace of announcements has quickened. Across 2025 and early 2026, we saw multiple new deals (e.g. Incyte-Genesis Feb 2025 and May 2026, AZ-Tempus/Pathos Apr 2025, GSK-Noetik Jan 2026, Merck-Mayo Feb 2026). This momentum suggests a snowball effect: once a few large pharmas head down this path, peers follow to avoid missing out.

6.4 Risks and Considerations

While promising, experts caution about risks:

- **Data Privacy and Security:** Leaked training data could reveal sensitive info. Pharma must ensure models don't memorize patient data (model inversion attacks). Comprehensive de-identification and access logs are needed.
- **Regulatory Hurdles:** How AI-driven findings integrate with regulatory submissions is still evolving. Agencies may eventually require disclosure of AI model provenance or training data.
- **Failure Mode:** If AI models underperform or drug programs fail, milestone payments evaporate – leaving big upfronts as sunk cost. Companies hedge by doing multiple targets.
- **Equity Dilution:** Giving equity to AI partners (as in Incyte's $40M Genesis stock) risks undervaluing the startup if it grows fast, but ignoring equity may reduce startup's commitment.
- **Technology Obsolescence:** AI models improve rapidly; an exclusive license to a model may expire by the time valuable results come. Hence, iterative licensing (as with subscription fees) is emerging.
- **Monopoly vs Collaboration:** Pharma want to leverage data but may fear empowering potential competitors (e.g. Microsoft/OpenAI, Google). This geopolitical concern factors in deals – e.g. GSK's non-exclusive model may reflect careful positioning.

Despite hurdles, the **strategic imperative is pressing**: generative AI is reshaping industries, and "pharma R&D via AI" is widely viewed as inevitable. As one review noted, "pharma AI data pools" hold huge promise but require careful management (www.p05.org). Companies that effectively apply these data-as-a-service strategies expect accelerated discovery and new revenue streams, while laggards may be disrupted.

7. Future Directions and Implications

Looking ahead, several developments are likely:

- **Normalization of Data Deals:** The Incyte and AZ examples signal these deals will become commonplace. Within a few years, almost every large pharma will have one or more AI-data partnerships. Smaller pharmas and biotech will too, using data marketplaces or consortia if needed. We may see franchise-level AI deals (e.g. "AI collaboration for oncology portfolio" with explicit pipelines).
- **Standardization:** As volume grows, industry standards for data licensing terms may emerge (perhaps via trade groups or consortia). Standard contracts could expedite deals. Common data formats and ontologies (e.g. through the Pistoia Alliance's efforts on ELN data (www.pistoiaalliance.org)) will be pushed to facilitate AI training.

- **Tech Evolution:** AI platforms will mature. Cloud providers (AWS, Azure) are pitching ML infrastructure to pharma. Partners like NVIDIA are investing in generative biology. Alphafold 3's release (requiring special licensing) hints at big AI models in life sciences. We can expect specialized biotech LLMs (e.g. trained on patents, literature and data) to integrate with models trained on proprietary data for compound design.
- **Extended Applications:** Beyond R&D, pharma will explore AI in manufacturing (supply chain, quality control) and post-market (pharmacovigilance), which will open new data licensing fronts. Clinical AI (diagnostics, personalized medicine) might license data similarly. For instance, UK's NHS is being nudged to share data for AI research, which may lead to future data deals with big tech.
- **Policy and Ethics:** Governments may step in. Data privacy regulations (e.g. GDPR) might get updates specifically addressing AI training on personal health data. There may even be proposals to treat pharma data as a public good for AI (especially RWD), requiring compensation to donors. Ethical AI guidelines will become important as models suggest harmful drugs or biases emerge. Companies will need transparency – perhaps publishing model "data provenance" in filings.
- **New Business Models:** Startups will proliferate to help pharma manage this. Data marketplaces and data-collaboration platforms are likely to raise funding. AI companies will diversify: besides being research collaborators, they may offer "AI-as-a-Service" (AIaaS) platforms where pharma can commission custom models trained on provided data.
- **Academic–Industry Ties:** We may see academic consortia using pharma data for foundation models, akin to how OpenAI collaborated with Sanofi. Or large-scale federated models funded by NIH/IMI combining industry data ethically. The interface between public and private data could blur, with dual public-private models.

8. Conclusion

Pharmaceutical proprietary data is rapidly transitioning into a **valuable AI asset**, changing how industry collaborations are structured. The Incyte–Genesis expanded collaboration (with ~$1B+ potential) is emblematic of a new era where **foundation AI models are co-built using pharma's secret experimental data** (www.businesswire.com). GSK's deal and AstraZeneca's Tempus alliance, among others, show that large players are committing eight- and nine-figure sums to access data and models.

These emerging deals marry the **financial structures of traditional drug licenses** (milestones, royalties) with the **data-sharing needs of AI development**. Successful agreements carefully define use-cases, protect intellectual property, and ensure pharma firms capture value for their data inputs. Our analysis finds recurring themes: pharma data commands strong upfront and success-based payments (www.businesswire.com); exclusivity is negotiated (non-exclusive licensing also appears); and equity can align interests. Experts unanimously note that *without proprietary, high-quality data, advanced AI has limited impact* (www.businesswire.com).

Going forward, as more companies adopt these models, we expect a proliferation of pharma-AI partnerships. Frameworks will evolve from these pioneering deals into standardized playbooks for data licensing. The potential payoff is huge: accelerated discovery and new medicines. But challenges – data governance, privacy, equitable sharing of gains – will be critical. Pharma executives and legal teams must tread carefully yet boldly: by formalizing their data's value, they can turn an R&D expense into a lucrative asset.

In sum, leveraging proprietary data in AI foundation models represents a **strategic shift in pharma innovation**. The field is in its early innings, but the "data-versus-IP" calculus is changing. As Noetik's Chief Business Officer put it, the GSK deal "defines a new asset class in biotech" (www.businesswire.com). Companies that refine this data-licensing playbook stand to lead the next wave of drug discovery.

References

- Incyte Corporation. *Incyte and Genesis Expand Molecular AI Collaboration to Accelerate Drug Discovery*. BusinessWire, May 20, 2026 (www.businesswire.com).
- Waldron, James. "Incyte pens $885M biobucks AI pact to use Genesis' GEMS to develop new drugs." *FierceBiotech*, Feb. 20, 2025 (www.fiercebiotech.com).
- Maddela, Vidya Sagar. "Incyte, Genesis Therapeutics enter AI-focused research collaboration." *WorldPharma News*, Feb. 21, 2025 (www.worldpharmaceuticals.net).
- Tempus AI, Inc. *"Tempus Signs Expanded Strategic Agreements with AstraZeneca and Pathos to Develop the Largest Multimodal Foundation Model in Oncology."* Tempus (Press Release), April 23, 2025 (www.tempus.com).
- Taylor, Nick Paul. "GSK inks model deal in $50M bet on Noetik's cancer AI platform." *FierceBiotech*, Jan. 8, 2026 (www.fiercebiotech.com).
- Noetik, Inc. *"GSK Licenses Noetik's AI Foundation Models in Anchor Partnership to Transform Cancer Therapeutic R&D."* BusinessWire, Jan. 8, 2026 (www.businesswire.com).
- Recursion Pharmaceuticals. SEC Form 8-K, Nov. 9, 2023 (Tempus Labs collaboration) (www.streetinsider.com).
- Bristows LLP. *"AI in drug discovery: do you know the value of your healthcare data?"* Legal Insight, 2023 (www.bristows.com).
- IBM Research. *"Biomedical Foundation Models."* IBM Research (project overview) (research.ibm.com).
- Kasianov, Roman. "Tempus, AstraZeneca, and Pathos Partner to Build Oncology Foundation Model Using Multimodal Data." *BioPharmaTrend*, April 23, 2025 (covering Tempus press release) (pharmaphorum.com) (www.tempus.com).
- Tanzi, Giulia. "Bayer and Recursion expand oncology research partnership." *Pharmaceutical-Technology*, Jan. 25, 2024 (www.pharmaceutical-technology.com).
- *Healthcare Data Monetization Market Trends 2025…* towardshealthcare.com (Towards Healthcare), Apr. 2026 (www.towardshealthcare.com).
- *Datavault AI Inc.:$10M licensing deal with Scilex.* StockTitan (from SEC filings), Nov. 2025 (www.sec.gov).
- Sanofi S.A. *"Press Release: Sanofi, Formation Bio and OpenAI announce first-in-class AI collaboration."* May 21, 2024 (www.sanofi.com).

- Arora, Aneesh et al. *“CorpusAnalytiX and Ariana Pharma Announce Collaborative Alliance to Accelerate AI-Driven Drug Development.”* Ariana Pharma (Press Release), Oct. 15, 2024 (www.arianapharma.com).

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://www.businesswire.com/news/home/20260520680248/en/Incyte-and-Genesis-Expand-Molecular-AI-Collaboration-to-Accelerate-Drug-Discovery
2. https://www.tempus.com/news/tempus-signs-expanded-strategic-agreements-with-astrazeneca-and-pathos-to-develop-the-largest-multimodal-foundation-model-in-oncology/%3Fsrsltid%3DAfmBOoqeuvOp7b7KCGTtUvr0kPg6EtFaP1KX_qVHyLYzmJXtduxj3mIY
3. https://www.businesswire.com/news/home/20260108468293/en/GSK-Licenses-Noetiks-AI-Foundation-Models-in-Anchor-Partners_to-Transform-Cancer-Therapeutic-Research-and-Development
4. https://www.merck.com/news/merck-and-mayo-clinic-announce-new-research-and-development-collaboration-to-support-ai-enabled-drug-discovery-and-precision-medicine
5. https://www.streetinsider.com/SEC%2BFilings/Form%2B8-K%2BRECURSION%2BPHARMACEUTICAL%2BFor%3A%2BNov%2B09/22387008.html
6. https://www.fiercebiotech.com/biotech/incyte-pens-900m-biobucks-ai-pact-use-genesis-gems-develop-new-drugs
7. https://www.worldpharmaceuticals.net/news/incyte-genesis-therapeutics-enter-ai-focused-research-collaboration
8. https://www.businesswire.com/news/home/20260108468293/en/GSK-Licenses-Noetiks-AI-Foundation-Models-in-Anchor-Partnership-to-Transform-Cancer-Therapeutic-Research-and-Development
9. https://www.bristows.com/news/ai-in-drug-discovery-do-you-know-the-value-of-your-healthcare-data
10. https://research.ibm.com/projects/biomedical-foundation-models
11. https://www.arianapharma.com/corpusanalytix-and-ariana-pharmaannounce-collaborative-alliance-toaccelerate-ai-driven-drugdevelopment
12. https://chipp.ai/ai/glossary/foundation-model
13. https://pharmaphorum.com/news/astrazeneca-joins-tempus-pathos-cancer-ai-project
14. https://www.sanofi.com/en/media-room/press-releases/2024/2024-05-21-05-30-00-2885244
15. https://www.sec.gov/Archives/edgar/data/0001682149/000110465925106780/tm2530277d1_8k.htm
16. https://www.merck.com/bdl_item/merck-and-mayo-clinic-announce-new-research-and-development-collaboration-to-support-ai-enabled-drug-discovery-and-precision-medicine
17. https://www.towardshealthcare.com/insights/healthcare-data-monetization-market-sizing
18. https://www.mordorintelligence.com/industry-reports/healthcare-data-monetization-market
19. https://arxiv.org/abs/2108.07258
20. https://www.p05.org/pharma-ai-data-pools-promise-and-pitfalls
21. https://www.pistoiaalliance.org/projects/current-projects/semantic-enrichment-of-eln-data
22. https://www.fiercebiotech.com/biotech/gsk-inks-model-deal-50m-bet-noetiks-cancer-ai-platform
23. https://www.pharmaceutical-technology.com/news/bayer-recursion-oncology-research

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.