

JULY 9, 2026

Proprietary Data and Specialist AI: The Evolving Market in Life Sciences Transactions

In life sciences, three structures are emerging for AI-enabled research transactions — but each raises legal and strategic questions, requiring adaptations to existing deal frameworks.

Noelle C. Dubiansky, Erini R. Svokos, Bethany P. Withers

Molecules have long been the primary asset in life sciences. Companies develop therapeutics, protect them with patents, and license or sell them. With the rise of artificial intelligence (AI), the data generated in the course of researching and developing molecules has become an asset in its own right.

A market has formed around that shift. AI-native biotech companies have spent years generating the proprietary data needed to train AI models — through clinical genomics and patient sample networks, automated wet labs, and spatial biology platforms — and building the integrated platforms to do it. They now license those models as products and services to life sciences companies that are sitting on decades of their own accumulated research, which AI can analyze to identify drug candidates, predict treatment response, and map disease biology.

In one sense, these deals are relatively straightforward: A life sciences company pays for access to a model that can analyze its data. In some cases, the company may grant the model owner rights to use a portion of its data to train or improve the model. But these deals raise complex issues that deal practice was not designed to handle.

How does each party protect the proprietary assets it brings to the transaction? What are each party's rights to the model's outputs and the drug candidates, targets, and development programs they enable? And if the model is improved or fine-tuned through use of the licensed data, who owns the resulting system?

Three structures have emerged for these deals. They differ in how access to the model is granted and what each party can do with the data and outputs, and each concentrates the legal risk in a different place.

1. The Model Owner Runs the Analysis

In the most common arrangement, the model owner performs R&D activities using its AI platform and delivers therapeutic discoveries and insights to another life sciences company. The other company never touches the model.¹ This structure may take two forms.

In the first form, a life sciences company identifies a research objective (a biological target, a compound of interest, a therapeutic area), then a model owner runs analysis using its own proprietary or public datasets to generate results. The legal questions relate to what the model owner must deliver and what the recipient company can do with the results, including how they can be used and whether they can be deployed to train a competing model.

In the second form, a company provides proprietary data (compound libraries, inactive program records, clinical safety records) to the model owner for it to use and analyze. In some of these deals, the model owner is also permitted to use that data to train or improve its platform. Many consider access to this proprietary data as part of the consideration for the transaction.

The legal exposure expands significantly here. For what purposes can the model owner use the contributed data? Can it retain the data after the collaboration ends? Can it use insights derived from that data in future work with additional life sciences companies? These questions are addressable in contract through use restrictions, retention and deletion obligations, audit rights, and explicit prohibitions on using the provided data for purposes such as model training.

When the contract permits the model owner to use contributed data for training, the life sciences company providing the data should understand that the contribution is difficult to meaningfully reverse. Once absorbed into a model's weights, its influence persists. The model's behavior may continue to reflect what it learned even if the contract requires deletion of the original files. The contract must address not just what happens to the original data but what the model owner can do with a platform that has been trained on or improved through that data's use.

2. The Model Owner Licenses the Model

In a second structure, a model owner grants a licensee direct access to the model itself, sometimes via API, sometimes as a deployable instance. The licensee runs analysis on its own data, and, in some cases, does so without the model owner's involvement at all. There's been a trend in favor of this structure recently, including for cancer research² and biologics discovery.³

In this structure, legal exposure runs in both directions. For the model owner, the risks are several. Direct access can enable the licensee to probe, extract, or reverse engineer the model's underlying training data and architecture, and it provides unfettered access to the model-generated output. Models can be made to leak training data through documented technical attacks, and trade secret law doesn't protect against reverse engineering unless a contract explicitly prohibits it. The contract should address what the licensee can do with the model's outputs, including whether they can be sublicensed, published, or used in regulatory submissions, and what happens if the licensee uses those outputs to train a competing model. If the structure enables the licensee to fine-tune the model using its own data, the agreement should address each party's rights to the fine-tuned model upon termination.

The risks are also significant for the licensee. When a licensee runs its own proprietary data through a hosted model, the contract should explicitly state what the model owner can retain or learn from that interaction. Provenance is the other critical exposure: If the model was trained on data the model owner didn't have the right to use, the licensee's downstream work, including drug-development decisions and

regulatory submissions, may be compromised. What warranties cover the training data, and who bears the liability if those warranties turn out to be wrong, are among the most consequential terms in any model licensing agreement.

3. Federated Learning and Consortium Platform Agreements

In a third structure, a group of companies collaborate to develop a shared AI model, training it on proprietary data they provide, in exchange for access to the shared model.

Lilly TuneLab⁴ is an example of this structure. Eli Lilly and Company (“Lilly”) trains its models and other third-party models on its own historic preclinical, ADME (absorption, distribution, metabolism, and excretion), and safety data (datasets it values at more than \$1 billion) and has made those models available to biotech partners at no cost.⁵ In exchange, partners contribute their own data through federated learning. Rather than sending raw data to Lilly, each partner runs the model locally against its own datasets, returning only model updates. In a well-structured federated architecture, partner-provided data doesn’t leave its own environment, and no partner sees another partner’s data.

The legal exposure here is distinct from the first two structures, though it shares some features with on-premises deployments under the second structure, in which the licensee hosts the model locally and the data doesn’t cross a boundary. The questions are about what rights accumulate over time. Who owns improvements to the model generated through use by partners? If a partner’s data has shaped the model in ways that can be traced back to that partner, what rights does it retain in the resulting system? What rights are granted with respect to the outputs, and do those rights affect partners’ underlying intellectual property (IP)? These questions sit at the intersection of IP ownership, contract law, and technical architecture. There are no established answers. Parties entering these structures need to examine carefully how each issue has been addressed in the specific agreement.

Getting the Right Deal

These transactions sit at the intersection of two bodies of practice that have historically operated separately: technology licensing and life sciences dealmaking. Structuring them well requires fluency in software and data licensing, the technical vulnerabilities that create legal exposure, and how AI models actually work. The regulatory constraints, milestone structures, and IP frameworks that govern life sciences partnerships require deep domain experience in drug development. Deals structured without both tend to leave significant gaps in provisions that weren’t drafted, risks that weren’t anticipated, and questions that weren’t asked until something went wrong.

The most economically important IP in these collaborations often sits downstream of the model itself, in the drug candidates, targets, biomarkers, and development programs generated through use of the AI platform. AI-enabled discovery complicates traditional ownership and attribution analysis, raising questions about inventorship, derivation, and rights in AI-assisted outputs. Agreements, therefore, need to address not only ownership of models and training data but also rights in discoveries generated through use of the platform.

The deals that hold up combine technical architecture that makes unauthorized extraction harder with contractual provisions that make it unlawful. But technical controls alone can’t resolve questions of

ownership. Who controls outputs that the model generates across engagements with different partners? If the model owner's proprietary training data covers the output, the recipient party typically seeks a broad unblocking license, granting the right to use outputs without infringing the model owner's underlying IP. Negotiating that scope, ensuring it's narrow enough to protect the model owner and broad enough to be useful to the recipient party, is among the most contested points in these deals.

* * *

Visit the Forces of Law hub and discover our full collection of insights into legal dynamics disrupting modern business strategy.

[1] "Alnylam and Inceptivo Form Strategic AI Collaboration to Accelerate the Discovery of RNAi Therapeutics," Business Wire (June 3, 2026); "insitro and Bristol Myers Squibb Collaboration Expanded with Nomination of New Targets," insitro (March 23, 2026).

[2] "GSK Licenses Noetik's AI Foundation Models in Anchor Partnership to Transform Cancer Therapeutic Research and Development," Business Wire (January 8, 2026).

[3] "Chai Discovery Announces Collaboration with Eli Lilly and Company to Accelerate Biologics Discovery," Business Wire (January 9, 2026); "Chai Discovery Announces License Agreement with Pfizer to Accelerate Drug Discovery with AI," Business Wire (June 4, 2026).

[4] "Lilly launches TuneLab platform to give biotechnology companies access to AI-enabled drug discovery models built through over \$1 billion in research investment," PR Newswire (September 9, 2025).

[5] "Chai Discovery Collaborates with Lilly TuneLab to Offer AI Capabilities to Select Biotechs," Business Wire (June 18, 2026); "insitro partners with Lilly to build first-in-kind machine learning models to advance small molecule drug discovery," insitro (September 9, 2025).

This informational piece, which may be considered advertising under the ethical rules of certain jurisdictions, is provided on the understanding that it does not constitute the rendering of legal advice or other professional advice by Goodwin or its lawyers. Prior results do not guarantee similar outcomes.

GET AI INSIGHTS DELIVERED

Subscribe for actionable insights, industry perspectives, events, and practical guidance for companies building and investing in AI.

[Subscribe](#)

CONTACTS

Noelle C. Dubiansky

Partner

Technology

ndubiansky@goodwinlaw.com

Erini R. Svokos

Partner

Technology

esvokos@goodwinlaw.com

Bethany P. Withers

Partner

Technology

bwithers@goodwinlaw.com