IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Fine-Tuning Foundation Models for Pharmaceutical R&D

> 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 · 2026-04-15 · fine-tuning llms · foundation models · pharmaceutical r&d · ai drug discovery · medicinal chemistry · computational chemistry · large language models · insilico mmai gym

Original article: https://intuitionlabs.ai/articles/fine-tuning-foundation-models-pharma-rd



In this article

1. Executive Summary
2. Introduction and Background
3. Foundation Models in Drug Discovery and Development
4. What Are Foundation Models?
5. Previous Efforts and Domain-Specific LLMs
6. Applications of Foundation Models in R&D
7. Challenges of General LLMs and Need for Fine-Tuning
8. Insilico MMAI Gym: Architecture and Curriculum
9. Performance and Case Studies
10. Insilico's Internal Benchmarks
11. Case Study: Qwen3 Evaluation
12. Case Study: Clinical Reasoning (Target & Trial Benchmarks)
13. Comparison with Other Advances
14. Data Sources and Benchmarks
15. Business Model: AI Gym Memberships
16. Wider Industry Context and Perspectives
17. Organizational and Cultural Shifts
18. Regulatory and Safety Considerations
19. Market and Technological Trends
20. Perspectives from Experts and Press
21. Future Directions and Implications
22. Conclusion

Executive Summary

The convergence of **large foundation models** (especially large language models, or LLMs) and pharmaceutical research heralds a new era in drug discovery. Traditional drug R&D is notoriously slow and expensive – often taking over a decade and $2–5 billion per drug (www.linkedin.com) (www.deloitte.com) – prompting the industry to embrace AI and data-driven approaches. Foundation models (e.g., GPT-4, BioGPT, DrugGPT) trained on massive general or biomedical corpora have shown promise in tasks like molecule design, target identification, and clinical trial planning (www.deepgenomics.com) (pmc.ncbi.nlm.nih.gov). However, out-of-the-box LLMs often fail key pharmaceutical tasks: they hallucinate chemistry facts, ignore 3D structure, and produce "chemically implausible" reasoning (insilico.com) (www.genengnews.com). Domain adaptation should be evaluated for the intended use; fine-tuning is one approach, alongside knowledge-grounded systems such as DrugGPT. Insilico Medicine's **Science MMAI (Multi-Modal AI) Gym** represents a pioneering controlled environment for systematically fine-tuning any causal or frontier LLM, using curated chemistry, biology, and clinical data to produce a "pharmaceutical-grade" model.

Key findings include:

- **Performance Improvements:** Insilico reports gains from MMAI Gym training on medicinal-chemistry, clinical-trial, and target-discovery benchmarks. Its Qwen3-14B-MMAI results include SOTA or near-SOTA performance on 4 of 22 TDC ADMET tasks and better performance than TxGemma-27B-Predict on 12 of 22 tasks (Insilico launch release). These task-specific comparisons should be distinguished from the company's broader up-to-10× marketing claim. In an academic study, Zhang et al. (2024) reported 69–95% exact accuracy across five chemical text-mining tasks using fine-tuned ChatGPT models and minimal annotated data (Chemical Science study).
- **Curriculum-Based Training:** The MMAI Gym employs a **multi-stage curriculum** of *Medicinal Chemistry*, *Biology/Target Discovery*, and *Clinical Development* modules. It leverages Insilico's proprietary datasets (millions of medicinal-chemistry optimization chains, 100M+ synthesis descriptions, etc.) and advanced generative tools (Chemistry42, PandaOmics) to teach "scientific reasoning" in domain-specific formats (insilico.com) (www.eurekalert.org). Innovations include *Reasoning Trace* training (forcing step-by-step chain-of-thought in answers) and *Chemical Format Augmentation* (translating between SMILES, IUPAC names, SELFIES, etc., so models learn the underlying concept of molecules) (insilico.com).
- **Benchmarking and Validation:** Every training cycle is rigorously evaluated on out-of-distribution benchmarks. Public suites like the Therapeutics Data Commons (TDC) are used to quantify errors on ADMET predictions (www.eurekalert.org). Proprietary benchmarks (e.g., TargetBench for target biology, ClinBench for trial outcomes) measure real-world relevance. Wet-lab experiments can further validate predictions through Insilico's automated assay platforms. Insilico claims Gym-trained models match or exceed state-of-the-art task-specific models across multiple endpoints (insilico.com) (www.genengnews.com).
- **Business Model and Ecosystem:** The Gym is offered as a flexible *membership program*. Pharma/biotech firms or AI labs bring any causal LLM (open-source or proprietary) and select a **CSI** (Chemical Superintelligence), **BSI** (Biology/Clinical Superintelligence), or **PSI** (Pharmaceutical Superintelligence) track tailored to their pipeline (insilico.com) (www.genengnews.com). In weeks-to-months, they receive an upgraded model (promised up to 10× better) plus detailed benchmarking reports (insilico.com) (www.genengnews.com). This contrasts with prior in-house AI initiatives, emphasizing collaboration: Insilico actively invites partners to "bring [their] AI model to MMAI Gym" and even offers "CSI" and "BSI" as distinct training programs (insilico.com) (www.genengnews.com).
- **Strategic and Cultural Implications:** Experts note that simply applying LLMs as-is is insufficient; cross-disciplinary alignment and new organizational models are required (www.deepgenomics.com) (pubs.acs.org). For instance, Deep Genomics argues that shifting to a single foundation model (from dozens of narrow tools) creates synergistic gains and easier scaling (www.deepgenomics.com). However, this demands a "multilingual" workforce bridging AI and wet-lab science (www.deepgenomics.com) (pubs.acs.org). The long-term vision ("Pharmaceutical Superintelligence" or PSI) is a coordinated AI–wet-lab "system-of-systems" where an LLM orchestrates specialized engines and experiments (www.linkedin.com) (pubs.acs.org). Regulatory challenges loom large: LLMs are black boxes and ungrounded models could pose safety issues (pubs.acs.org).

- **Outlook:** Early users report highlighting advantages: dramatically faster migration of powerful models into niche drug tasks, and democratization of AI tools. Similar domain-specific LLM efforts (BioGPT, DrugGPT, etc.) are also advancing, underscoring a broad trend. Yet caution is warranted: success depends on data quality, interpretability, and rigorous testing. The MMAI Gym exemplifies the direction of pharmaceutical AI – blending massive compute/training with deep domain expertise – and suggests a future where LLMs become integral to drug pipelines, with appropriate domain adaptation, use-specific validation and human oversight. Knowledge-grounded approaches such as DrugGPT may also be suitable.

In sum, **Insilico's MMAI Gym** represents a novel, systematic approach to fine-tuning foundation models for pharmaceutical R&D. By combining high-quality scientific data, advanced fine-tuning techniques, and robust validation, it promises to turn general "giant brains" into precise "scientific engines" for drug discovery (insilico.com). This report reviews the historical context, technical methods, case studies, and implications of this approach, situating it within a rapidly evolving AI-driven pharma landscape. All claims below are supported by published data, expert analysis, and illustrative examples (see Citations throughout).

Introduction and Background

Pharmaceutical R&D faces a productivity crisis: new drug approvals remain low relative to soaring costs, despite unprecedented needs. AI and machine learning have long been touted as solutions to accelerate drug discovery, but early efforts yielded mixed results (www.deepgenomics.com) (pmc.ncbi.nlm.nih.gov). The recent advent of **foundation models** – deep neural networks pre-trained on vast, multi-billion token corpora – has reinvigorated hope. Models like OpenAI's GPT-4, Google's Bard/Gemini, Meta's LLaMA, and specialized variants (e.g., BioGPT) can process complex inputs and generate fluent text, code, or biology-informed outputs (pmc.ncbi.nlm.nih.gov) (pubs.acs.org). Their versatility has transformed sectors from customer service to creative art, and increasingly into scientific domains (www.deepgenomics.com) (pmc.ncbi.nlm.nih.gov).

In drug discovery, however, the challenges are unique. Precision medicinal chemistry, multi-step synthetic pathways, biological mechanisms, and clinical endpoints all involve specialized knowledge and opaque data formats (molecular structures, assay results, patient cohorts). A general LLM, trained mainly on internet text, typically lacks the biochemical grounding to answer questions like "design a novel kinase inhibitor with improved ADMET properties" or "predict the probability of success for a Phase II trial of this compound." Indeed, Insilico's benchmarks show that without adaptation, flagship LLMs fail to predict key pharmacokinetic/toxicology endpoints (hERG blockage, drug-induced liver injury, etc.) and often give "chemically implausible" answers (insilico.com) (www.genengnews.com). Prompt engineering (few-shot examples, chains-of-thought prompts) can help in general NLP tasks, but numerous studies report that LLMs still hallucinate or produce misleading rationales when faced with specialized chemistry or biology prompts (insilico.com) (pubs.acs.org). In summary, the intrinsic **agency and reasoning** of LLMs does not automatically transfer to the complexities of pharmaceutical science.

Domain-specific fine-tuning is the natural countermeasure: by exposing a foundation model to curated chemical/biological data and tasks, one can "teach" it to use the language and logic of medicinal chemistry and biology rather than generic text patterns. This approach aligns with broader trends. In genomics and biology,

experts argue that narrow models have had limited impact, and that “AI foundation models” trained on broad datasets are beginning to show promise in drug discovery, clinical trials, and end-to-end pipelines (www.deepgenomics.com). Notably, Microsoft Research’s BioGPT and other specialized LLMs (e.g. Med-PaLM) demonstrate that pre-training on scientific corpora can vastly improve biomedical question answering and relationship extraction (academic.oup.com) (www.nature.com). Likewise, Zhang *et al.* (2024) fine-tuned GPT-3.5-turbo for five chemical text-mining tasks. GPT-3.5-turbo is proprietary and was accessed through OpenAI’s API; the study also evaluated open-source models including Mistral and Llama (Chemical Science study).

Within this context, Insilico Medicine (a clinical-stage AI biotech firm) has launched the **Science MMAI Gym** to systematically fine-tune any LLM for pharmaceutical R&D. The concept of an “AI Gym” draws on analogies to athletic training: models undergo structured “workouts” (fine-tuning) to build “chemical and biological superintelligence” (insilico.com) (www.genengnews.com). This initiative builds on Insilico’s decade of experience in AI-driven drug discovery (27 preclinical candidates, 10+ INDs, multiple clinical trials) and massive proprietary datasets (insilico.com). By opening this training infrastructure to external partners, the company aims to transform general-purpose LLMs (e.g. GPT, Claude, Llama, as well as open-source models) into specialized engines that can reason and predict with pharma-grade precision (insilico.com) (www.genengnews.com).

This report provides an in-depth look at the Insilico MMAI Gym and its relevance. We begin by reviewing the capabilities and limitations of foundation models in life sciences, citing key studies and expert commentary (pubs.acs.org) (pmc.ncbi.nlm.nih.gov). We then describe Insilico’s fine-tuning approach in detail: data sources, training curriculum, and architecture (CSI/BSI tracks) (insilico.com) (www.eurekalert.org). Case studies illustrate the dramatic gains possible (e.g. Qwen variants) (www.genengnews.com). We compare Insilico’s framework with other LLM-based efforts in pharma (such as BioGPT and DrugGPT) and survey the evolving industry perspective (www.nature.com) (www.deepgenomics.com). Finally, we discuss business models, organizational challenges, regulatory considerations, and the future “pharmaceutical superintelligence” vision (www.linkedin.com) (www.deepgenomics.com). Throughout, specific data, benchmark results, and expert analyses are cited to support every statement. Tables summarize key projects and MMAI Gym components for clarity.

Foundation Models in Drug Discovery and Development

What Are Foundation Models?

“Foundation models” refer to very large neural networks (often transformer-based) pre-trained on broad data (text, images, code, protein sequences, etc.), which can then be fine-tuned or adapted to downstream tasks. The term underlines that these models serve as a generic base or “foundation” for many applications. Representative examples include GPT-n (OpenAI), Google's PaLM and Gemini, Meta’s LLaMA, and domain-specific variants like BioGPT, GatorTron, or Google’s Med-PaLM (academic.oup.com) (pubs.acs.org). They have achieved human-like performance in language tasks and have multi-modal extensions (e.g. images and

language in one model). In healthcare, regulatory agencies and researchers are actively exploring foundation models for medical coding, literature review, and even clinical decision support (pmc.ncbi.nlm.nih.gov) (www.nature.com).

In drug discovery and development, foundation models are appealing because they promise to **integrate knowledge across disciplines**. A single model, if sufficiently skilled, could potentially answer questions about chemistry, biology, pharmacology, and trial design. The alternative ("Rube Goldberg workflows" of many single-purpose algorithms) has been fragmented and slow. Deep Genomics's Brendan Frey notes that his company moved from dozens of specialized genetics models to one foundation model and found "synergies between tasks" that made it more powerful (www.deepgenomics.com). Similarly, the success of AlphaFold (essentially a specialized foundation model for protein structure) has inspired analogous efforts for small molecules and genomes.

However, generic foundation models have no innate domain-specific knowledge: if trained mainly on natural language, they only learn token patterns, not chemical physics or biological causality (pubs.acs.org). For example, models like ChemGPT (2022) or MTMolecule (2023) were trained on chemical SMILES strings and made progress in molecule generation, but they fundamentally treat molecules as sequences of characters (pubs.acs.org). SMILES can encode formal charge and specified stereochemistry (Daylight atom specification; isomeric SMILES rules). It represents a molecular graph with optional chiral indications. Thus, foundation models typically excel at broad reasoning but lack a built-in grasp of molecular structure or cellular pathways. Efforts like BioGPT (a GPT-like model pre-trained on PubMed) have shown that domain-specific pre-training helps – BioGPT outperformed previous models on biomedical relation extraction and QA (academic.oup.com) – yet even these need further fine-tuning for drug-specific tasks.

Text-based foundation models also lack verifiable factual grounding. A model might invent a plausible-sounding metabolic pathway or cite a "source" that doesn't exist. This is a concern for safety-critical domains like medicine; reported "hallucinations" of LLMs (fabricated but fluent information) have drawn scrutiny (pubs.acs.org). Regulators require auditable evidence for medical claims, which opaque general models cannot provide (pubs.acs.org) (www.nature.com). Thus, any deployment in pharma will likely need careful retraining on verified data and mechanisms for traceability.

In summary, foundation models offer unprecedented scale and breadth, but applying them to pharmaceutical problems demands intense domain adaptation. This can come in two forms: (1) better *initial knowledge integration* (pre-training on pharma data) and (2) *post-training fine-tuning*. Many AI startups and pharma labs are trying both. For instance, BIGCHEM is a distributed computing framework for solving quantum-chemistry problems in parallel (Hicks and Martinez, 2024). Others, like Insilico's MMAI Gym, take existing general models (even closed-source ones) and fine-tune them with scientific "curricula."

Previous Efforts and Domain-Specific LLMs

Even before MMAI Gym, several teams created specialized LLMs for chemistry and biology:

- **BioGPT (Luo et al., 2022)** – A domain-specific generative Transformer pre-trained on biomedical literature. The paper reports **44.98% F1 on BC5CDR end-to-end relation extraction**. Separately, it reports **78.2% accuracy on PubMedQA**, a question-answering task (BioGPT paper). These scores

refer to different tasks, not a single entity-extraction evaluation.

- **DrugGPT (Zhou et al.)** – A *collaborative* LLM for drug analysis, published in *Nature Biomedical Engineering*. DrugGPT incorporates clinical knowledge bases (Drugs.com, NHS, PubMed) and a triaged architecture: one LLM analyzes queries and selects relevant knowledge, another retrieves information, and a third generates answers. It was benchmarked on five tasks (drug recommendation, dosage, adverse reaction ID, drug–drug interaction, pharmacology QA) and outperformed GPT-4, ChatGPT (GPT-3.5), and Med-PaLM2 on all, achieving state-of-the-art with fewer parameters (www.nature.com). Importantly, DrugGPT is *knowledge-grounded*: it only generates answers based on explicit medical sources, addressing the hallucination problem by design.
- **Domain-specific Finetuning (Zhang et al., 2024)** – A study in *Chemical Science* demonstrated the power of fine-tuning for chemistry text mining. The authors curated five tasks (compound entity recognition, reaction role labeling, NMR data extraction, MOF synthesis info, action sequence generation) and compared LLMs with/without fine-tuning. They found that fine-tuned GPT models achieved 69–95% exact accuracy across tasks with minimal data (pubs.rsc.org), *substantially higher* than prompt-only or larger in-domain models. For instance, a fine-tuned GPT-3.5-turbo outperformed multiple bespoke models despite using far less data. This study underscores that "fine-tuning LLMs can revolutionize complex knowledge extraction with versatility and low code requirements" (pubs.rsc.org).
- **Other GPT-3.5/Mistral experiments** – Beyond publications, many internal efforts have likely been underway. Stanford's Chemistry AI group has explored GPT-based retrosynthesis, and pharma companies have run pilots with GPT-like models for target ID and literature review. Some reports (e.g., a Cambridge Consultants blog) mention prototype LLMs being fine-tuned on proprietary discovery data (though these are often not public). The field is moving fast, but the MMAI Gym is among the first to formalize and commercialize such training at scale.

Table 1 highlights representative LLM projects in pharmaceutical R&D:

Project (Model)	Org / Authors	Base Model & Size	Fine-Tuning / Data	Key Achievements	Reference
Insilico MMAI Gym – Chemical Superintelligence (CSI) (Qwen3-14B-MMAI)	Insilico Medicine	Qwen3-14B (14B params)	2-week SFT+RFT with 4M medchem chains, 100M synth descriptions, etc.	Company-reported SOTA or near-SOTA on 4 of 22 TDC ADMET tasks; beat TxGemma-27B-Predict on 12 of 22.	Insilico launch release
Insilico MMAI Gym – Biology/Clinical (BSI) (Qwen3-4B-MMAI)	Insilico Medicine	Qwen3-4B (4B)	Finetuned on clinical trial-updates data; reward via DeepSeek's GRPO	Company-reported ClinBench F1 ≈0.82 → 0.94; accuracy 0.92, recall 1.00.	Insilico launch release
Insilico MMAI Gym – Target Discovery (BSI) (Qwen3-1.7B-MMAI)	Insilico Medicine	Qwen3-1.7B (1.7B)	Finetuned on genomics/pathway data; reward via GRPO	Company-reported top composite ranking for novel target identification on TargetBench.	Insilico launch release

Project (Model)	Org / Authors	Base Model & Size	Fine-Tuning / Data	Key Achievements	Reference
BioGPT (BC5CDR relation-extraction checkpoint)	Luo *et al.* / Microsoft (2022)	BioGPT	Biomedical-literature pre-training; task-specific fine-tuning	44.98% F1 on BC5CDR end-to-end relation extraction	BioGPT paper; official checkpoints
DrugGPT	Zhou *et al.* (published online 2025) – Nature Biomed. Eng.	Knowledge-grounded collaborative LLM system	Integrated clinical KBs; chain-of-thought prompts	Achieved SOTA on 5 drug tasks (recommendations, DDIs, etc.), surpassing GPT-4 with fewer params (www.nature.com).	DrugGPT study
Fine-Tuned LLM for Chem Text (GPT-3.5-turbo)	Zhang *et al.* (2024) – *Chem. Sci.*	GPT-3.5-turbo	Supervised fine-tuning on curated chemical corpora	Exact accuracy 69–95% on 5 chem text tasks (entities, NMR, retrosynthesis) – beating specialized models trained on much more data (pubs.rsc.org).	pubs.rsc.org
Insilico MMAI Gym – Liquid Model (LFM2-2.6B)	Insilico & Liquid AI (2026)	Custom "Liquid" 2.6B	Extensive Chem42/PandaOmics data; multi-stage SFT+RFT	Outperformed a 27B "TxGemma" model on TDC ADMET tasks; SOTA in MuMO-Instruct molecular optimization; expert-level retrosynthesis.	insilico.com

Table 1. Examples of fine-tuned foundation models in pharmaceutical R&D. Performance improvements are benchmarked relative to generalist baselines or prior art. (MuMO-Instruct: multi-objective molecular optimization tasks; TDC: Therapeutics Data Commons ADMET tasks.)

These and other emerging models illustrate a central point: **with targeted training, even mid-sized LLMs can achieve or surpass state-of-the-art for drug discovery challenges**. Notably, the Insilico/Liquid "LFM2-2.6B" (2.6B parameters) achieved better performance on safety/pharmacokinetics benchmarks than a much larger 27B parameter model (insilico.com), highlighting that data and training regimen often matter more than sheer scale. It also shows that an "AI Gym" concept need not rely on using the largest models – smaller "scientific specialist" models can excel if properly trained (insilico.com).

Applications of Foundation Models in R&D

Drug discovery tasks amenable to LLMs include:

- **Molecule Design and Optimization:** Prompting an LLM to propose chemical structures or SMILES for desired properties (affinity, solubility, etc.). Fine-tuned models can integrate ADMET filters and multi-step optimization (as in the MuMO-Instruct benchmark (www.genengnews.com)).
- **Retrosynthesis and Synthetic Planning:** Converting target molecules into reaction sequences. Insilico reports that its Gym-trained Liquid model matched top-tier proprietary and specialist chemistry models on retrosynthesis (company report).

- **Property Prediction (ADMET):** Predicting absorption, distribution, metabolism, excretion, toxicity endpoints from structure. Standard QSAR models exist, but LLMs can capture more context by reasoning on similar molecules. Insilico's benchmarks use TDC to evaluate LLMs on these tasks (www.eurekalert.org).
- **Target Identification:** Interpreting omics data (gene expression, pathways) to suggest novel drug targets. GPTs, when fine-tuned with omics reasoning tasks, have been shown to excel at multi-objective target scoring (www.eurekalert.org).
- **Biomarker and Genomics Insights:** Analyzing transcriptomic data, disease mechanisms, or protein interactions. Fine-tuned LLMs can reason over gene lists and pathways, an ability Insilico's Gym explicitly trains through "omics-aware reasoning" (insilico.com) (www.eurekalert.org).
- **Clinical Trial Planning and Safety:** Summarizing trial designs, predicting phase success rates, identifying safety signals. Insilico uses ClinBench to evaluate prediction of Phase 2 trial success or failure (Insilico launch release). Early studies (and Insilico's own models) have shown LLMs can outperform older methods when fine-tuned on clinical data (www.genengnews.com) (www.eurekalert.org).
- **Regulatory and Text Mining Tasks:** Extracting structured data from literature (patents, publications). The chemical text-mining study (pubs.rsc.org) demonstrates LLMs can automate extraction of NMR shifts, reaction roles, etc., crucial for building proprietary databases.
- **Medical & Scientific Writing:** Automating drafting of reports, summarizing findings, or even writing code for analysis (as Genentech's primer notes, ChatGPT can produce initial model scripts, though error-prone) (pmc.ncbi.nlm.nih.gov). Such tasks free up scientists for creative work.

Across these areas, the unifying strategy is domain adaptation: present the model with high-quality samples of how experts would solve each problem (long chain-of-thought answers, annotated diagrams, benchmark datasets) and use that to refine its latent knowledge. The MMAI Gym implements this systematically with millions of examples and specialized benchmarks.

Challenges of General LLMs and Need for Fine-Tuning

Despite their general intelligence, flagship LLMs face *systematic* shortcomings in pharmaceutical contexts. The Insilico team and others have identified several failure modes:

- **Hallucinations and Implausibility:** In absence of relevant data, LLMs may confidently hallucinate drug facts. For example, an un-tuned GPT might predict toxicities by incorrectly generalizing from text (e.g. "Because NSAIDs cause gastric upset, all molecules with an aromatic ring cause liver injury"), resulting in "vague or chemically implausible" output (insilico.com) (www.genengnews.com). Zhavoronkov notes that general LLMs often fail crucial endpoints like hERG inhibition or LD_{50}, making them unreliable for real decision-making (www.genengnews.com). The ACS *Central Science* outlook also emphasizes this: LLMs lack the deep biochemical grounding and often produce fluent but factually incorrect text (pubs.acs.org).

- **Inadequate Representations:** SMILES supports specified stereochemistry, including double-bond configuration and tetrahedral chirality. If chirality is not specified for a chiral atom, it remains unspecified (Daylight isomerism specification). As one analysis notes, “the tokenization of SMILES or SELFIES simplifies chemical representations and loses information” (pubs.acs.org). Thus LLMs may treat one string as “different words” instead of the same molecule. MMAI Gym mitigates this by multi-format training (SMILES ↔ IUPAC ↔ SELFIES) (insilico.com), so the model learns the underlying concept of a particular compound.
- **Lack of Ground Truth and Traceability:** In regulated healthcare, it is essential to know *why* a model made a recommendation. “LLMs often function as opaque black boxes,” making it difficult to verify a given prediction (pubs.acs.org). For tasks like mechanism-of-action or patient stratification, regulators will demand evidence of training data and logic. Without fine-tuning, general LLMs cannot easily cite their knowledge sources. DrugGPT tackles this by building in evidence-traceable prompting, but typical LLMs do not. Insilico’s approach of supervised reasoning traces also improves traceability: each answer comes with a “reasoning chain” that can be evaluated for plausibility (insilico.com).
- **Scope Mismatch:** Foundational LLMs are trained for linguistic fluency, not multi-modal scientific reasoning. Tasks like protein binding affinity or pharmacokinetics involve simulation and quantitative models beyond word prediction. Therefore, even advanced prompting rarely achieves “science-grade” accuracy. As the*ACS Perspectives* paper observes, LLMs “excel at symbolic reasoning and task coordination, but they lack deep biochemical and structural grounding,” necessitating integration with physics-based modules (molecular dynamics, docking) for credible drug insights (pubs.acs.org). This explains why simply scaling an LLM (to 100B parameters) has not solved drug-specific problems; without domain tuning, the intelligence remains general but superficial.
- **Evolution of Data:** Drug discovery relies on up-to-date data (new targets, molecules, trial results). Many LLMs are static with cut-off dates and may not incorporate the latest science. Fine-tuning can address this by retraining on new patents, assay outcomes, and literature. Insilico’s Gym can continually train on fresh in-house molecules and public benchmarks, keeping the model aligned with cutting-edge knowledge.

These challenges motivate domain adaptation; fine-tuning and knowledge-grounded systems such as DrugGPT are approaches to evaluate (DrugGPT study). Insilico’s positioning of MMAI Gym explicitly addresses the “LLM performance gap” in drug discovery (insilico.com). Alex Zhavoronkov emphasizes that neither prompting nor “light fine-tuning” solves this gap; what is needed is systematic, iterative training akin to schooling a scientist (www.genengnews.com) (insilico.com). In essence, a foundation model must be taught the *language of molecules and biology*, complete with context and reasoning patterns used by experts (insilico.com) (www.eurekalert.org).

Insilico MMAI Gym: Architecture and Curriculum

Insilico's **Science MMAI Gym** is a domain-specific AI training environment. Launched in early 2026, it transforms any eligible causal/frontier LLM into a specialized drug-discovery engine. Its core concept is a structured, multi-modal curriculum that immerses the model in pharmaceutical tasks. Fig. 1 below outlines the Gym's high-level architecture:

(academic.oup.com) *Figure 1. Overview of Insilico's MMAI Gym training framework. (Source: Insilico Medicine (insilico.com), annotated by author.)*

The Gym offers two main **tracks** reflecting Insilico's long-term vision: **Chemical Superintelligence (CSI)** and **Biology/Clinical Superintelligence (BSI)**. There is also a combined **Pharmaceutical Superintelligence (PSI)** stream. Conceptually:

- **CSI Track (Chemistry focus):** Concentrates on small-molecule and organic chemistry reasoning. Key tasks include multi-step reaction planning, property optimization (ADMET), retrosynthesis, docking and binding analysis, and generative design of novel compounds. Training data features Insilico's *Chemical SuperIntelligence* dataset: millions of medicinal chemistry "optimization chains" (i.e., sequences of compound modifications and associated property changes), 100M descriptions of organic reactions, large libraries of 3D molecular conformations and simulations, etc. (insilico.com). The model learns to reason about structure-property relationships, synthetic feasibility, and the chemical logic behind pharmacology.
- **BSI Track (Biology/Clinical focus):** Focuses on genomics, proteomics, disease biology, and clinical trial reasoning. Tasks include target identification from omics data, pathway analysis, multi-objective scoring of potential targets against disease profiles, biomarker selection, and trial outcome prediction. Key data includes biological reasoning datasets: gene expression profiles, pathway/network databases, protein interaction data, annotated disease mechanisms, and historical clinical trial outcomes. Insilico uses its *Biological SuperIntelligence* resources (e.g., PandaOmics for target generation). The Gym also incorporates proprietary benchmarks like ClinBench to measure understanding of clinical protocols and endpoints (insilico.com) (www.eurekalert.org).
- **PSI Pathway:** Partners who need an integrated solution (chemistry + bio) can pursue the full PSI (Pharmaceutical SuperIntelligence) curriculum, which covers both CSI and BSI domains. This reflects Insilico's vision of an AI that bridges drug design through development.

The **curriculum** itself is multi-stage. According to the launch announcement (insilico.com) and technical blog (www.eurekalert.org), the Gym's modules include:

- *Medicinal/Organic Chemistry:* Multi-step optimization chains (progressive modification of a lead molecule through successive design iterations), reaction reasoning (interpreting reaction mechanisms and catalysts), retrosynthesis (breaking target into precursors), and 3D structure-property tasks

(predicting molecular binding or conformation). The Gym provides sequenced tasks of increasing complexity, e.g. “given this scaffold, propose a functional group modification to improve solubility while maintaining potency” or “write a plausible synthetic route from starting materials.”

- *Biology and Target Discovery:* Omics analysis tasks (ranking gene targets given expression data), pathway reasoning (explaining how target modulation affects disease), disease mechanism interpretation, and multi-objective target scoring (balancing efficacy, safety, druggability). For example, the model might be trained on tasks such as “Given this RNA-seq fold-change and pathway diagram, identify which protein is a promising target” or “score these four targets by the likelihood of successful drug intervention.”
- *Clinical Development:* Understanding and predicting clinical trial design and outcomes. Tasks include interpreting trial protocols, endpoint definitions (e.g. progression-free survival), and predicting trial success probabilities. Insilico uses proprietary benchmarks (e.g. **ClinBench**) where F1 metrics are measured for answering clinical questions. The Gym might have simulated tasks like “Given this Phase II trial design document, predict whether it will succeed or fail.”

In addition, an **“On-Demand Reasoning Data Generator”** is a key innovation (www.eurekalert.org). Insilico’s Chemistry42 and PandaOmics platforms can automatically generate synthetic reasoning traces – essentially additional training examples – to continuously supply fresh examples for the Gym. For instance, Chemistry42 might algorithmically generate a new medicinal chemistry optimization case (structure, modifications, outcomes), creating an “infinite curriculum” of high-quality examples.

The **technical workflow** for model training involves:

1. **Curated Datasets and Curation:** The Gym ingests *validated domain data* (internal and public) for training. Insilico cites “millions” of data points: e.g. *“4 million medicinal chemistry optimization chains, 100 million organic synthesis descriptions, hundreds of thousands of molecular dynamics trajectories”* (insilico.com). Data quality is paramount, so the Gym team pre-processes and decontaminates text (removing irrelevant or incorrectly labeled data) to avoid spurious learnings (www.genengnews.com).
2. **Multi-Task Supervised Fine-Tuning (SFT):** The base LLM undergoes standard supervised learning on a mixture of tasks. For each curriculum module, numerous input-output examples (often in chain-of-thought format) are provided. For example, a training item might be a multi-turn chat or a single prompt: “Q: *Suggest the next step in this multi-step synthesis optimization chain.* A: *Step 3: replace NH2 with N-CH3 to improve solubility – explanation follows…*”. The model weights are adjusted to predict these high-quality expert answers. SFT handles tasks like entity recognition, property prediction, and initial strategy generation.
3. **Reinforcement Fine-Tuning (RFT with Domain Rewards):** After SFT, models are further refined via reinforcement learning using specialized reward models. These rewards evaluate the scientific validity of outputs. For instance, a reward model might give high score if the predicted logP value is plausible, or if a retrosynthesis route is syntactically correct. Insilico mentions use of DeepSeek’s GRPO (Group Relative Policy Optimization) algorithm for this phase (www.genengnews.com). This step aligns the model towards achieving measurable objectives (e.g. success rate on a benchmark) rather than just mimicking training text.

4. **Reasoning Trace Curriculum:** A key pedagogical trick is training the model to *simulate its reasoning chain*. During SFT/RFT, examples include intermediate steps ("chain-of-thought") so that the model learns to "think out loud" rather than output an answer directly. As Insilico describes: *"models are taught to generate 'thinking' chains… forcing the AI to work through chemical plausibility step-by-step"* (insilico.com). These traces provide generated explanations for review, but they may not faithfully reflect the model's actual reasoning (Anthropic research). Reviewers should check conclusions against source data, experiments, or validated computational methods.
5. **Chemical Format Augmentation:** To avoid overfitting to chemical naming, the Gym trains the model to translate molecules between different representations. It will see that a concept "aspirin" can be given as SMILES `CC(=O)OC1=CC=CC=C1C(=O)O`, a SELFIES string, or the name "acetylsalicylic acid." By learning to convert between them, *"the model learns the underlying concept of a molecule rather than just its name"* (insilico.com). This multi-format exposure prevents the model from relying on superficial cues.
6. **Benchmarking and Iteration:** After each training phase, the model is rigorously evaluated on held-out benchmarks. Insilico uses both public (e.g. TDC for pharmacokinetics/toxicity (www.eurekalert.org), TargetBench for target ID) and proprietary tests. Performance metrics (F1, accuracy, error results) are compared to pharma-grade thresholds. If performance is insufficient, the curriculum is adjusted (more data collection, reward tuning, etc.). Only when benchmarks meet criteria does a model "complete" that Gym cycle.
7. **Deliverable:** Upon completion, the partner receives an upgraded LLM with **CSI/BSI/PSI enhancements**, along with a comprehensive benchmarking report and, optionally, custom wet-lab validation. Insilico's staff reviews the outputs and can translate model suggestions into experimental plans via their automated lab facilities, closing the loop from *in silico* to *in vitro*.

Figure 2 below conceptualizes the Gym process:

Figure 2. Schematic of Insilico MMAI Gym fine-tuning pipeline. Curated domain data (med-chem chains, omics, etc.) is used in supervised and reinforcement tuning phases. Performance is validated on benchmarks (TDC, ClinBench, etc.) before producing a "CS I / BSI / PSI"–enhanced model. (Adapted from Insilico materials (insilico.com) (www.eurekalert.org).)

Key components of the MMAI Gym training are summarized in Table 2:

Component / Technique	Description	Illustration / Source
Domain-Specific Datasets	Massive curated datasets: ~4M med-chem optimization steps, ~100M organic reactions, protein/omics data, clinical trial records.	Insilico claims thousands of internal data points (insilico.com).
Supervised Fine-Tuning (SFT)	Multi-task training on labeled examples across chemistry, biology, and clinical tasks. Models see expert Q&A or chain-of-thought.	Training examples (curated by domain experts).
Reinforcement Fine-Tuning (RFT)	Continued training with reward models (e.g. DeepSeek's GRPO) that score outputs by scientific correctness or benchmark success.	Rewards align outputs with experimental outcomes (insilico.com).

Component / Technique	Description	Illustration / Source
Generative Data Augmentation	Use of Insilico's generative engines (Chemistry42, PandaOmics) to simulate new training examples (reasoning traces) on demand.	AI tools generate synthetic med-chem and omics cases (www.eurekalert.org).
Reasoning Traces (Chain-of-Thought)	Training prompts are augmented with step-by-step reasoning chains, so the model learns to "think" before answering.	Generated explanations may be unfaithful; independently verify scientific claims (insilico.com); faithfulness research.
Chemical Format Conversion	Tasks include translating molecules between SMILES, SELFIES, IUPAC, etc., teaching the model molecular concepts beyond token strings.	Enhances generalization across input formats (insilico.com).
Benchmarking and OOD Validation	Use of public (TDC) and private (ClinBench, TargetBench) out-of-distribution tests to evaluate real-world performance.	Ensures robustness to novel compounds/targets (insilico.com) (www.eurekalert.org).
Data Decontamination	Removing overlap between training and test sets, and filtering noise, to prevent overfitting or data leaks.	Prevents inflated performance (www.genengnews.com).
Liquid Foundation Model (LFM)	Custom architecture (LFM-2.6B) used in research, with non-standard attention for efficiency.	Enables faster training, beats larger models (insilico.com).
CSI/BSI/PSI Membership Tracks	Tiered programs for chemical-only, biology-only, or integrated pipelines, allowing clients to choose the Gym scope fitting their needs.	See membership model discussion (www.genengnews.com).

Table 2. Key components of the Insilico MMAI Gym training program. Each technique is applied systematically to build scientific reasoning in base LLMs.

Thus, the MMAI Gym transforms the broad knowledge of foundation models into domain expertise. It explicitly teaches the *language of chemistry and biology*, not merely leveraging plain-text correlations. As Insilico puts it, the Gym "teaches LLMs domain-specific scientific reasoning – the language, formats, and conceptual chains that chemists and biologists actually use" (insilico.com). By recasting drug discovery problems as rich multi-turn tasks (rather than single-prompt queries), the Gym aims to achieve **"pharmaceutical-grade"** accuracy.

Performance and Case Studies

Insilico's Internal Benchmarks

From the launch and media reports, Insilico has published internal benchmark results illustrating the Gym's impact (insilico.com) (www.genengnews.com). Highlights include:

- **Reported Performance Gains**: Insilico advertises up-to-10× improvement compared with baseline. Its Qwen3-14B-MMAI evaluation reports SOTA or near-SOTA performance on 4 of 22 TDC ADMET tasks and better performance than TxGemma-27B-Predict on 12 of 22 tasks (Insilico launch release). The up-to-10× claim should be read as an attributed company claim, with each benchmark assessed using its own metric and comparator.

- **State-of-the-Art (SOTA) Success**: The Qwen3-14B-MMAI model (after Gym tuning) achieved SOTA or near-SOTA performance on multiple absorption, distribution, metabolism, excretion, and toxicity (ADMET) tasks from TDC (www.genengnews.com). It also matched or exceeded best-in-class "category-specific" models on 5 molecule-optimization tasks in the MuMO-Instruct benchmark (www.genengnews.com). Notably, all these were done by a *single* fine-tuned LLM, whereas previously separate models might have been needed for each task.
- **Clinical/Biology Gains**: Insilico reports that Qwen3-4B's ClinBench F1 increased from approximately **0.82 to 0.94**, with post-training accuracy **0.92** and recall **1.00**. It also reports a top composite TargetBench ranking for Qwen3-1.7B after SFT and GRPO training (Insilico launch release). These are company-reported benchmark results.
- **Site-of-Action (CSI vs BSI)**: Insilico allows partners to choose a chemically-focused or biologically-focused training. These tracks show targeted improvements: CSI yields experts in ADMET and med-chem workflows, while BSI yields experts in target discovery and clinical reasoning. A full PSI curriculum produces both (but presumably at greater cost/time).

The reported results concern the tested models and benchmarks, including the Qwen3-14B-MMAI comparisons on 22 TDC ADMET tasks (Insilico launch release). They should not be interpreted as establishing reliable performance across an end-to-end drug-discovery workflow.

Case Study: Qwen3 Evaluation

A notable example is Insilico's fine-tuning of **Qwen3-14B** (an open-source Chinese LLM by Alibaba Cloud). This evaluation was described in industry media (www.genengnews.com). Key metrics:

- *Baseline (pre-Gym):* Insilico reports that Qwen3-14B, with five-shot in-context learning, failed completely on roughly 70% of medicinal-chemistry benchmarks.
- *Training:* The company reports a two-week Gym training period.
- *Post-Gym Performance:* Insilico reports that Qwen3-14B-MMAI solved over 95% of medicinal-chemistry benchmarks. This describes benchmark coverage, not example-level accuracy or the recovery rate among previously failed individual tasks. It reports SOTA or near-SOTA results on **4 of 22 TDC ADMET tasks**, and better results than TxGemma-27B-Predict on **12 of 22** (Insilico launch release).

These comparisons do not establish a tenfold increase in a universal success rate or reliable performance in every medicinal-chemistry workflow.

This confirms observations from academic R&D: size alone does not solve chemistry. Instead, data and training pedagogy do. Notably, the Qwen3-14B-MMAI had 14B parameters, but a custom-trained 2.6B LFM (Liquid model, see next section) outperformed an untrained 27B model on these tasks (insilico.com). These results underscore that *"with the right data and training procedures, a smaller, efficient model can outthink a giant"* (insilico.com).

Case Study: Clinical Reasoning (Target & Trial Benchmarks)

In the biology/clinical domain, Insilico reports similar success. The Qwen3-4B (4B parameter) and Qwen3-1.7B (1.7B parameter) were tuned on clinical tasks:

- **Qwen3-4B (4B):** Insilico reports that SFT and GRPO-based reinforcement training raised ClinBench F1 from approximately **0.82 to 0.94**, with accuracy **0.92** and recall **1.00**. The benchmark concerns Phase 2 trial outcomes published after January 1, 2025 (Insilico launch release). These are company-reported results, not prospective evidence of clinical reliability.
- **Qwen3-1.7B (1.7B):** Fine-tuned on target discovery tasks using TargetBench evaluations. After supervised fine-tuning and GRPO, it jumped to top rank on TargetBench, excelling at identifying novel therapeutic targets across diseases (www.genengnews.com). This shows even small models can be trained to become cutting-edge in biology reasoning.

In summary, **fine-tuning overcame the size advantage**: 4B and 1.7B models, properly taught, could match or beat generic Titans on focused tasks.

Comparison with Other Advances

These Insilico-reported results align with independent findings. For instance, the *Chem. Sci.* study by Zhang *et al.* fine-tuned GPT-3.5 on chemical text and noted *"fine-tuned ChatGPT models excelled in all tasks… with minimal data… outperforming models trained on far more in-domain data"* (pubs.rsc.org). This resonates with the MMAI Gym story: a generally-trained model, after exposure to expert reasoning sequences, can outperform specialized pipelines.

In AI-driven chemistry, others have shown that smaller novel architectures can challenge giants. Insilico's partner Liquid AI developed the 2.6B *Liquid Foundation Model* (LFM2-2.6B) which, after Gym training, **outperformed a 27B parameter "TxGemma-27B" model on multiple ADMET benchmarks** (insilico.com). This "David vs Goliath" result – smaller specialist surpassing a huge generalist – echoes the view that intelligent training beats brute force scaling. It suggests the future may involve a mix of foundation LLMs and lean domain-specific models, coexisting in pharma pipelines.

In the broader industry, large pharma (Roche, Novartis, etc.) and tech giants (Google, Amazon) are also investing in similar capabilities, though details are often proprietary. Reports of collaboration between cloud providers and biotech on LLMs abound. For example, Novartis partnered with Microsoft to build AI tools; NVIDIA is developing Clara Discovery platform to accelerate drug design with LLMs. OpenAI and DeepMind have also indicated interest in chemistry (AlphaFold's success already demonstrated AI's utility in life sciences). While these projects lack public benchmarks yet, the consensus is clear: **the ability to fine-tune foundation models for drug domains is considered a key strategic advantage**.

Data Sources and Benchmarks

The effectiveness of any fine-tuning regime hinges on data. Insilico leverages both internal proprietary data (from its own AI-driven drug pipeline) and large public corpora:

- **Proprietary Medicinal Chemistry Data:** Insilico reports “millions” of optimization chains – basically historical records of optimizing lead compounds for potency, selectivity, ADMET, etc. Such data inherently encodes medicinal chemists’ decision-making. Similarly, its Pharma.AI products (e.g. Chemistry42) have generated extensive synthetic chemistry descriptions. The Gym uses this high-quality corporate data to train realistic problem examples.
- **Public Chemistry/Pharmacology Data:** Open databases like ChEMBL, PubChem, DrugBank, PDB, and literature mining (e.g. patents, papers) provide additional cases. LLM fine-tuning projects (BioGPT, ChemSci study) often use such public corpora. The Gym likely includes curated subsets (e.g. known inhibitors, clinical trial abstracts) while avoiding direct overlap with benchmarks.
- **Biological/Ontological Data:** Pathway databases (KEGG, Reactome), expression atlases (TCGA, GEO), and target databases (DisGeNET, OMIM) are sources for biology tasks. Insilico’s PandaOmics platform has integrated multi-omics datasets from thousands of diseases, which feed into target scoring modules.
- **Clinical Data:** ClinicalTrials.gov, FDA labels, and real-world evidence sources might be used. For ClinBench, Insilico has assembled or licensed a dataset of past Phase II trials with outcome labels. The Gym can train models to link trial design features to success probability.

Benchmark Suites: To measure progress, established benchmarks are employed:

- **Therapeutics Data Commons (TDC):** A public suite of ADMET tasks (e.g., Caco-2 permeability, plasma protein binding, microsomal clearance, hERG inhibition). Insilico found that generic LLMs have very high mean absolute error or low AUROC on many TDC tasks (www.eurekalert.org). MMAI Gym training aims to drastically reduce these errors. (Insilico’s reported gains imply MAEs dropping an order of magnitude closer to experimental noise.)
- **MuMO-Instruct (Multi-Objective Molecular Optimization):** An open benchmark for optimizing multiple properties of a molecule while preserving its core structure. The Gym-trained models achieved SOTA success rates on five MuMO tasks (www.genengnews.com), demonstrating multi-property reasoning.
- **ClinBench:** A proprietary benchmark by Insilico (not publicly detailed) for clinical trial reasoning. It contains statements or Q&A about Phase II designs and outcomes. Role: to test if an LLM can interpret trial objectives and predict success. The fine-tuned Qwen-4B excelled here.
- **TargetBench:** An open benchmark (by the research community or Insilico) for novel target identification. It likely involves multi-disease scoring of targets. The Gym-tuned Qwen-1.7B secured the top rank on this.
- **Additional Academic Benchmarks:**
- The RSC study used specific test sets for entity recognition and reaction conversion (pubs.rsc.org).
- Other community benchmarks (e.g., MoleculeNet tasks, CHEMBIO tasks) could be used for cross-validation, though not mentioned explicitly.

Insilico reports systematically evaluating “out-of-distribution” (OOD) performance. That means the final models are judged on compounds, targets, and trials **not seen during training**, ensuring true predictive power. OOD validation is critical to avoid overfitting – a known risk when fine-tuning on narrow tasks. Insilico’s iterative training process with OOD checks was described in industry press (www.genengnews.com).

Business Model: AI Gym Memberships

The Science MMAI Gym is offered as a **membership-style program**. Insilico describes it as flexible and partner-friendly (insilico.com) (www.genengnews.com). Key elements:

- **Custom Engagements:** Clients (pharma companies, biotech, AI labs) can opt for short sprints (e.g. 2-week intensive) or longer programs (several months), depending on their pipeline needs. Each engagement focuses on the client’s chosen track: CSI for chemistry-heavy programs, BSI for biology/clinical, or combined PSI.
- **Bring Your Model:** Partners typically supply the base model they want to improve (own GPT-4 chat instance, open LLM, or cloud-native AI). The Gym team then applies the fine-tuning regimen to that model. For partners without their own model, Insilico offers its internal CSI/BSI models as starting points.
- **Deliverables:** At the end of the membership, the client receives the enhanced LLM (CSI/BSI/PSI model) and a detailed benchmarking report. Insilico also offers **wet-lab validation**: the model’s suggestions can be experimentally tested in Insilico’s automated labs (e.g. high-throughput assays, synthesis robots) for critical leads (insilico.com). This integration of *in silico* and *in vitro* closes the feedback loop, as promised in the “AI-Ground truth.”
- **Reported Performance Claims:** Insilico advertises “up to 10x performance improvement compared to baseline” (Insilico launch release). This advertised potential gain should not be described as a contractual performance guarantee. Success criteria, pricing, and other commitments should be established from the actual agreement.
- **Partnerships Invited:** Insilico invites pharma and biotech companies, AI labs, and cloud providers to bring their models to the Gym for CSI, BSI, or PSI training (Insilico launch release).
- **Value Proposition:** For a pharma, the value is converting general AI assets into domain experts rapidly. Rather than hiring hundreds of chemists to build specialized models, a firm can leverage Insilico’s data and curricula to instantaneously upgrade an LLM. The membership model mirrors cloud services: pay for compute and know-how in exchange for a refined model.

In essence, the MMAI Gym is positioned as an R&D **infrastructure service**. It encapsulates Insilico’s data and expertise as a platform: similar to cloud ML training but bespoke for pharma. Members benefit from proprietary data that no standalone LLM could easily access. This could shift how drug companies procure AI: from internal ML teams to “AI training as a service.”

Wider Industry Context and Perspectives

Organizational and Cultural Shifts

Experts highlight that realizing LLM benefits in pharma is as much cultural as technical (www.deepgenomics.com). Deep Genomics' Brendan Frey, for example, emphasizes that using foundation models requires a *"multilingual"* organization where biologists, chemists, and AI researchers collaborate seamlessly (www.deepgenomics.com). Rather than handoffs between siloed specialists, a foundation model approach thrives when wet-lab scientists and AI developers co-design workflows. Insilico's open Gym aims to foster this by making AI training a collaborative process: companies send their models and data, review benchmarks, and potentially adjust domain priorities.

Similarly, ACS's "Prompt to Drug" perspective (Zhavoronkov *et al.*, 2026) outlines a "system-of-systems" where LLMs orchestrate specialized AI agents and physics-based simulators (www.linkedin.com) (pubs.acs.org). Their long-term vision of **Pharmaceutical Superintelligence (PSI)** is not a single monolithic model but a coordinated platform. Insilico's CSI/BSI tracks can be seen as initial steps toward PSI. The emphasis is on integration: LLM fine-tuning is only one piece. For a full PSI, you need data infrastructure, automated labs, regulatory alignment, and human oversight, all pointed out by Zhavoronkov's review (www.linkedin.com) (pubs.acs.org). In that sense, the MMAI Gym is a specialized training environment *within* the larger PSI concept.

Regulatory and Safety Considerations

Bringing LLMs into pharma pipelines invites regulatory scrutiny. Key challenges include model explainability, data provenance, and reproducibility. Insilico's approach addresses some by design: requiring chain-of-thought reasoning and benchmarking improves transparency. But open issues remain: for example, if an LLM suggests a novel compound, how do we prove it meets safety standards? FDA's January 2025 draft guidance for drug and biological products contains non-binding recommendations and is not for implementation. Future requirements should not be presented as settled. Insilico's emphasis on rigorous testing and "human-in-the-loop" seems aligned with a precautionary approach.

Market and Technological Trends

The creation of the MMAI Gym reflects two major trends:

1. **AI as an R&D Utility:** Just as biotechs outsourced informatics to cloud and contract labs, more companies are now treating AI capabilities as a service. Insilico is effectively renting AI knowledge. Large tech firms see opportunities here: Microsoft's cloud offers diffusion of OpenAI's models for pharma, Google Cloud markets Vertex AI for life sciences, and smaller players (AbCellera, Exscientia) build domain pipelines. McKinsey estimates that generative AI could create **$60–110 billion in annual economic value across the pharmaceutical and medical-product industries** (McKinsey report). This is a potential-value estimate, rather than a forecast that the value would be realized by 2025.

2. **Data and Model Democratization:** The open LLM era (e.g. Meta's LLaMA, Mistral, Qwen) means companies no longer must rely on closed APIs. They can now fine-tune open-source models internally, or send them to a service like MMAI Gym. This lowers barriers to entry. Conversely, domain-specific models like Insilico's internally developed "CSI model" (a proprietary chemistry LLM) show pharma also wants in-house models. The Gym bridges the gap: it can turn your chosen model into a specialist, whether that model is community-driven or proprietary.

Perspectives from Experts and Press

Pharma AI expert Alex Zhavoronkov (Insilico's CEO) frames the Gym as analogous to human training: *"AI models train in the Gym… emerge in better shape"* (www.genengnews.com). News outlets have largely covered MMAI Gym as a breakthrough: Genetic Engineering & Biotechnology News (GEN) headlined "No Pain, No Gain: Insilico 'Gym' Gets AI Models Into Shape" (www.genengnews.com), noting the novelty of treating LLMs like gym trainees. Industry commentary emphasizes that the Gym does more than augment semantics – it injects domain *intelligence*.

Yet some caution may be advised. The results reported are impressive, but come from Insilico's benchmarks; independent validation will be needed. A public manuscript is available; the OpenReview PDF identifies it as "Published as a workshop paper at ICLR 2026". Critics also warn of hype: one-language-model-does-it-all risk, overconfidence in a single AI, or neglecting hard labs work. The LinkedIn discussion around the ACS article reflects this tension: the idea of "prompt-to-drug" is intriguing but also raises red flags about over-reliance on AI without safeguards (www.linkedin.com).

In summary, the MMAI Gym has been welcomed as an innovative platform by industry media and some AI experts, but the true test will be real-world deployments. Partnerships (with biotechnology companies, CROs, or cloud vendors) will likely accelerate validation and reveal both strengths and limitations of this approach.

Future Directions and Implications

The emergence of Insilico's MMAI Gym points toward several future trends:

- **General to Specialized Pipelines:** We may see more hybrid models: foundation LLMs fine-tuned by pharma-specialized curricula. Insilico's CSI/BSI/PSI could become standards (like "checkpoints" for pharma LLMs). The strategy of domain training might extend beyond LLMs to multi-modal models (e.g. combining chemical structure inputs, biomedical texts, and clinical images).
- **Integration with Automation:** A critical path is closing the loop with experimental data. Insilico's mention of *automated assay platforms* hints at "closed-loop AI" where a Gym-optimized model designs compounds, robotic labs synthesize/test them, and results feed back for further training. This is reminiscent of "self-driving laboratories" and is aligned with the "Liquid Intelligence" paradigm. Over time, such systems could vastly shrink cycle times.
- **Regulatory and Ethical Frameworks:** As LLMs become involved in candidate selection, regulators will need to define new categories (AI-made molecule?) and standards for algorithmic audit. Models will need explainability features – for instance, requiring the chain-of-thought output (from the Gym) to be

part of documentation. If Insilico's PSI vision materializes, companies may treat the final model more like an official "expert system" which human scientists consult, rather than a black-box oracle.

- **Open Science and Collaboration:** Interestingly, Insilico is initially offering the Gym mostly through paid partnerships, but the spirit of "AI Gym for Science" suggests community building. For example, Open Targets is a public–private partnership between academic institutes and pharmaceutical companies. It's possible we may see similar coalitions around AI: shared benchmarks (ClinBench is one), open data for target discovery, even open-sourcing parts of trained models. Deep Genomics' blog calls for open toolkits ("GenomeKit") and cultural shifts (www.deepgenomics.com), which will be echoed in pharma's AI initiatives.
- **Economic Impact:** Economic estimates should be distinguished from measured reductions in drug-discovery or development time. Smaller companies and academia could leverage such platforms to compete with big pharma by renting intelligence. However, there will be winners and losers: firms with the ability to integrate AI into operations will surge ahead, while those slow to adapt may be left behind.
- **Technological Evolution:** Just as deep learning progressed from CNNs to transformers, future model architectures will likely evolve. Insilico's use of a *Liquid architecture* hints that novel designs (mixtures of experts, neurosymbolic hybrids) may become the new norm for specialized tasks. Concurrently, techniques like parameter-efficient tuning (LoRA, prefix-tuning) could make Gym-like training cheaper.
- **Ethical and Social Considerations:** The amplification of AI in drug pipelines raises questions about workforce changes (e.g., will intermediary medicinal chemist roles change drastically?), data biases (models trained on certain chemical scaffolds may overlook others), and global health (will generative discovery benefit neglected diseases?). The concept of "superintelligence" itself draws sci-fi fears; in reality, the immediate challenge is more prosaic: responsibly validating AI leads and maintaining human oversight (www.linkedin.com).

Conclusion

Insilico's MMAI Gym exemplifies a critical milestone in AI-driven pharma R&D. By applying rigorous fine-tuning curricula to foundation models, it bridges the gap between general-purpose AI and pharmaceutical expertise. Insilico reports that its specialized models outperform general LLMs on drug-discovery tasks in internal benchmarking (Insilico launch release). These company-reported comparisons should be assessed by model, task and metric.

Yet this approach also highlights broader truths: domain knowledge cannot be magically consolidated by scale alone. It requires meticulous teaching (as the Gym provides), evaluation on real-world problems, and integration into broader research systems. The future of drug discovery will likely be shaped by hybrid teams of AI and experts working in tandem, as Deep Genomics notes: AI models as assistants require *"our people… understand each other"* across disciplines (www.deepgenomics.com).

For practitioners and decision-makers, the MMAI Gym underlines that **fine-tuning foundation models is now a core competency in pharma**. Early adopters who harness these tools and data services can expect accelerated pipelines and novel insights. However, caution and validation are imperative: models must be

thoroughly benchmarked (as the Gym emphasizes) and tied back to empirical evidence.

In sum, the story of Insilico's MMAI Gym is emblematic of the fine-tuning revolution in AI and drug discovery. It shows that with curated data and methodical training (supervised + reinforcement), an LLM's weaknesses can be systematically addressed. As one of the first full-stack "AI gyms" for science, its success or failure will inform the industry's trajectory. If the claimed 10× boosts hold up in practice, we may indeed be entering a new era where an AI, sculpted by a "pharma curriculum," joins the lab bench as a potent collaborator on the quest for new medicines.

Source links

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

1. https://www.linkedin.com/posts/andriibuvailo_a-new-paper-suggests-that-a-plain-language-activity-7432433229489438720-S9A5
2. https://www.deloitte.com/global/en/industries/life-sciences-health-care/research/the-convergence-of-ai-technologies-and-human-expertise-in-pharma-r-and-d.html
3. https://www.deepgenomics.com/blog/getting-most-out-ai-foundation-models-tips-pharmaceutical-companies
4. https://pmc.ncbi.nlm.nih.gov/articles/PMC11984503
5. https://insilico.com/news/l16526c8p1-insilico-medicine-launches-science-mmai?amp=true
6. https://www.genengnews.com/topics/artificial-intelligence/no-pain-no-gain-insilico-gym-gets-ai-models-into-shape
7. https://www.nature.com/articles/s41551-025-01471-z
8. https://www.eurekalert.org/news-releases/1113489
9. https://pubs.rsc.org/en/content/articlehtml/2024/sc/d4sc00924j
10. https://insilico.com/blog/mmai-liquid-ai
11. https://pubs.acs.org/doi/full/10.1021/acscentsci.5c01473
12. https://academic.oup.com/bib/article/23/6/bbac409/6713511
13. https://daylight.com/meetings/summerschool98/course/dave/smiles-atoms.html
14. https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
15. https://www.osti.gov/pages/servlets/purl/2371584
16. https://arxiv.org/abs/2210.10341
17. https://github.com/microsoft/BioGPT
18. https://daylight.com/meetings/summerschool98/course/dave/smiles-isomers.html
19. https://insilico.com/mmai
20. https://www.anthropic.com/research/reasoning-models-dont-say-think
21. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
22. https://www.mckinsey.com/industries/life-sciences/our-insights/generative-ai-in-the-pharmaceutical-industry-moving-from-hype-to-reality
23. https://arxiv.org/abs/2603.03517
24. https://openreview.net/pdf?id=4m2TcQpwkf

25. https://opentargets.org

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.