SEMINAR REVIEW

Life Science AI Exchange

Agentic AI & Technology Best Practices

24 June 2026 • Pistoia Alliance LSAIE Programme • Open Webinar

Reviewed by Rob Gill

The second seminar of the 2026 Life Science AI Exchange programme opened with a question that the industry has struggled to answer precisely: not whether agentic AI works, but how to make it work reliably, at scale, and in environments where the cost of a silent error is a failed scientific programme. Over two hours, four presentations — from an AI scientist platform builder, two Novo Nordisk research teams, a GxP-validated enterprise platform, and Genentech/Roche — assembled a remarkably coherent picture of the current state of the art, and of the gaps that remain. The through-line connecting all four was trust: how it is earned, how it is preserved under pressure, and what happens when it breaks.

1. Building an AI Scientist: Lessons from Benchling

Nicholas Larus-Stone — VP AI, Benchling

Nicholas Larus-Stone opened by identifying the most persistent misconception about AI scientists: that summarising literature and generating hypotheses constitutes the job. His argument was that this accounts for a small fraction of the scientific journey, and that a genuinely useful AI scientist must operate at the seam between the digital and physical world — including wet lab execution and automation. Twelve months ago, he noted, a credible AI scientist was not buildable. The combination of powerful reasoning models, Anthropic's agent skills standard for repeatable workflows, and the emergence of background agents triggered by events rather than people has changed that calculus.

The DBTL loop as the organising unit

The core architecture Benchling has developed treats the Design–Build–Test–Learn loop as the fundamental unit of scientific work, applicable across discovery biology, preclinical science, and process development alike. Within each loop: the human scientist sets the goal; the AI scientist generates a plan, selects models, suggests and designs experiments, captures results, analyses data, and recommends next steps — surfacing key decisions for human approval at each stage. The scientist remains principal investigator; the AI acts as a highly capable research associate who never sleeps and never loses track of prior work.

A live demonstration showed this in practice: from a prior yield optimisation experiment, the AI scientist performed data analysis, proposed a Design of Experiments (DOE), generated downloadable run sheets and pre-populated notebook entries, and flagged caveats about its own analysis — all within a conversational interface that preserved human control at each decision point.

> *You should be sceptical of demos, because in AI it's easy to make an interesting demo and much harder to put it into practice. Let me show you what our customers are actually doing.*

Key lessons from deployment

The lessons Larus-Stone drew from real customer use are instructive. Structured data remains critical: organisations that care about data structure realise AI benefits faster. LLMs can help structure messy data — CRO reports, batch records, legacy ELN entries — but they cannot compensate for its absence. Multi-LLM architectures are increasingly necessary, both to extract the best performance from different providers' strengths and to avoid single-vendor lock-in. Wet lab connectivity remains the hardest unsolved problem: the gap between proposing an experiment and executing it in automation is where most AI scientist projects currently stall. And the next frontier, Larus-Stone argued, is proactive intelligence — agents that know when to act without being asked, triggered by experimental events, inventory states, or data arriving off a machine at 8pm.

A customer case study — with Prime Medicine on analytical method validation for gene therapy — illustrated the scale of potential impact: what had taken months of manual data aggregation was compressed to days, with the AI scientist guiding follow-up studies to complete validation within weeks. The economic argument Larus-Stone made during the Q&A was characteristically direct: most organisations are underspending on AI. In a field where a wet lab experiment costs tens or hundreds of thousands of pounds, spending ten pounds to ask an AI agent whether that experiment has already been done is an obvious value trade.

2. Benchmarking Biomedical Reasoning and Building Agents Scientists Trust

Frederik Steensgaard Gade & Josefa Stoisser — Novo Nordisk / TU Denmark

The Novo Nordisk contribution was two-part: Frederik Steensgaard Gade presented fundamental research on whether LLMs can reason over biomedical knowledge graphs, and Josefa Stoisser presented a practical design pattern for building bioinformatics agents that laboratory scientists actually trust. Together they formed one of the most technically rigorous contributions of the session.

Does apparent LLM reasoning reflect genuine inference?

Steensgaard Gade's work addresses a question that sits at the foundation of AI-assisted drug discovery: when a language model correctly identifies a relationship between two biological entities, is it reasoning, or is it retrieving memorised information? The distinction matters enormously. In early-stage drug discovery, where value lies precisely in identifying relationships that are not yet well-known, a model that performs well only on high-degree, widely-cited biology is of limited utility.

Using the iCraft knowledge graph — 1.3 million nodes covering genes, diseases, and chemicals, with 30 million edges derived from biomedical literature — Steensgaard Gade constructed a rigorous link-prediction benchmark with temporal splitting. Training data was drawn from pre-2000 literature; test data from relationships first observed from 2002 onward. This eliminates the leakage risk that bedevils many LLM biomedical benchmarks, where frontier models may have been trained on the very test material they are evaluated against.

The experimental design isolated the contribution of three information sources: random baseline (no information), name prior (the entity name alone, encoding whatever the LLM has memorised during pre-training), and graph context (topological neighbourhood, without the name). Results were striking. In the

easy random-negative setting, graph context roughly doubled baseline performance. But in the harder setting — where negative candidates are topologically similar to the true link — the name prior substantially outperformed context, and adding context to the name actually decreased performance. The conclusion is uncomfortable: apparent LLM discovery capability in biomedical knowledge graphs is largely name-prior and availability-driven. It reflects retrieval from memorisation, not reasoning over graph topology. Performance concentrates in high-degree, well-studied biology and decays towards the sparse, novel connections that drug discovery most needs.

> *Name-based performance is concentrated in high-degree biology. But we actually care more about low-degree biology, where there is less information. Performance does not transfer to sparse or novel biology, where discovery actually matters.*

A Knowledge Graph Foundation Model (Ultra), given only topological information, performed comparably to LLMs given the same context — suggesting that for link prediction, architectural choices matter as much as scale, and that leakage-controlled evaluation is essential for making fair comparisons.

Silent misframing: the danger that doesn't announce itself

Josefa Stoisser's contribution addressed a problem specific to scientific agents used by non-computational biologists: the silent misframing. A silent misframing occurs when a user's underspecified request leads an agent to fill in an unstated assumption from its pre-training prior, produce a technically valid-looking result, and return it without flagging the assumption. The user, unable to inspect the underlying code, acts on the result — and introduces a scientific error that may not be detected for months.

The context is DREX-seq, an RNA sequencing readout for perturbation experiments used extensively at Novo Nordisk for target discovery. A biologist might ask the agent to 'compare the effect of Gene X knockdown' — intending comparison against a non-targeting control — and the agent might interpret this as comparison against an untreated baseline. The volcano plot looks plausible either way. Stoisser's team published a dataset quantifying this: in the most extreme scenario, frontier agents optimise for the completely wrong column up to 63% of the time.

> *A silent misframing happens when a user does not specify part of the task framing, the agent fills it in with an unintended choice, and still produces a valid-looking result. For a non-coding biologist who will not check the code, this is a severe problem.*

The design pattern Stoisser's team developed has four elements. First, build agents on workflows that scientists already trust — the DREX-seq agent cannot write arbitrary bioinformatics code; it invokes vetted pipelines and selects hyperparameters for them, reducing degrees of freedom. Second, put agency in the right places: allow the agent freedom in the task-framing stage before workflow execution, but force it to surface all framing choices to the user and require approval before proceeding; after execution, allow free interpretation of results, but clearly separate trusted analysis from agent interpretation in the UI. Third, convert biologist feedback into an evaluation dataset by collecting annotated conversations from internal users, extracting expected ground-truth answers, and encoding domain expertise into a scoring function. Fourth, use that dataset to drive controlled, measured improvements: generalise the failure mode, implement a change (new tool, new prompt, new constraint), and keep the change only if the eval score improves.

3. Agentic AI in the Enterprise: Governance and GxP Compliance

Sean Blake — Head of Regulatory Affairs and Quality, Sapio Sciences

Sean Blake's contribution shifted focus from the research frontier to the governance infrastructure required to deploy agentic AI in regulated pharmaceutical environments. Sapio's position — as a validated LIMS/ELN platform used across GxP workflows — means that the governance questions are not theoretical; they determine whether customers can use agents in their regulated processes at all.

Governance architecture

Sapio's approach begins with the observation that moving from a tool to an agent fundamentally changes the risk profile. Removing human intervention at each step increases risk, and governance must be redesigned accordingly — not bolted on. The AI management system they have implemented is aligned to ISO 42001 and integrated with their existing ISO 27001 and risk management framework, explicitly to avoid AI governance becoming a silo. Regulatory alignment spans OECD AI principles, the NIST AI Risk Management Framework, the EU AI Act, GAMP 5, FDA Computer Software Assurance guidance, GCP, GLP, GMP, 21 CFR Part 11, and EU Annex 11. ISO 42001 certification is in progress, expected by end of Q3 2026.

Validation of AI tools follows the same programme as the rest of the platform: GAMP 5, risk-based, with the key adaptation that acceptance criteria are defined as performance thresholds rather than deterministic outputs. Every agent action is logged, attributable, and exportable for FDA, EMA, and other regulatory inspection. There is no black box: inputs, outputs, model versions, prompts, and agent reasoning traces are all captured.

Autonomy levels and GxP applicability

Blake presented a four-level autonomy framework that defines the governance requirements for each tier:

Autonomy Level	Agent Behaviour	Human Role	GxP Applicability
Assistive	Suggests actions	Review required before any GxP use	Applicable to all GxP contexts
Recommended	Recommends decisions	Must actively approve each recommendation; passive acceptance insufficient	Applicable to all GxP contexts
Semi-autonomous	Initiates workflows on authorisation	Reviews outputs at defined checkpoints; cannot proceed past gates without sign-off	Selective workflows only
Fully autonomous	Executes within defined boundaries	Monitors; intervenes on exceptions; boundaries and escalation triggers formally validated	Restricted; requires risk justification

The key principle — stated explicitly — is that capability must never outpace accountability. If Sapio cannot validate an agentic capability, they will not deploy it.

Security and IP isolation

Infrastructure runs on AWS Bedrock with dedicated per-customer endpoints: no shared context, prompts, or data across tenants. Customer data is never used to train or fine-tune models. Encryption is TLS 1.3 in transit and AES-256 at rest. Customers may bring their own LLM, retaining full control over AI provider and data sovereignty. AI-generated outputs within the customer environment are treated as customer work product under standard IP rules. The acceptable use policy prohibits use of third-party AI tools with proprietary data; all approved agents are governed by IP-safe terms.

4. AI Data Readiness: The Foundation for Trustworthy Agentic AI

Arindam Sett, Edaeni Hamid & Kelly Mewes — Genentech / Roche

The Genentech/Roche team closed the formal programme with what is arguably the most strategically significant contribution: a systematic account of what actually blocks agentic AI from delivering value, and a framework for addressing it before projects begin rather than after they stall. The initiative — ADAR, AI Data Readiness — emerged from a pattern the team observed repeatedly: build an agent harness, evaluate, find poor performance, and discover that the root cause is not the AI but the data layer it is operating on.

The problem pattern

Arindam Sett described the consequences: erosion of trust in agentic systems, fragmented adoption, rework as ML engineers return to data owners for remediation (typically 4–6 months per project), and loss of predictability in AI outcomes. Edaeni Hamid framed three structural drivers. Standardisation gaps mean there is no validated approach to assess whether data from diverse sources — internal R&D systems, CRO outputs, real-world data — is actually ready for agentic use. Integration latency creates a compounding loop as issues surface late in multi-agent orchestration. And strategic blind spots result from organisations not knowing the readiness state of their own data assets, making prioritisation impossible.

> *The real value of AI is still not being seen in many organisations due to the state of the data. Because we are not aware of the readiness state, we are not even in a position to prioritise which data sources to remediate or which new ones to acquire.*

The ADAR evaluation framework

The five-dimension evaluation framework the team has developed covers: data quality (completeness, consistency, freshness — weighted by use case); semantic and metadata contextual intelligence (ontology alignment, versioning, unit consistency — described as critical across nearly all use cases); access and governance (usage rights, secure reproducibility, regulatory constraints for clinical and real-world data); generalizability (distribution balance, bias transparency, representativeness); and structural readiness (engineering compatibility, label quality, sample size, diversity). Each dimension is scored independently, allowing data asset owners to see precisely where gaps lie and what remediation is required.

Kelly Mewes outlined the deployment roadmap: from data auditing and standardised readiness scoring (current phase), through targeted improvement roadmaps with data asset owners, to a future state of automated data certification — a 'seal of approval' that only assets meeting minimum quality thresholds can enter production-grade model pipelines. The longer-term vision is real-time trust layer integration: agentic orchestration that continuously monitors data quality as a first-class pipeline component.

> *The organisations that will win the AI race will not necessarily be the ones with the best algorithms. They will be the ones with the best data pipelines fuelling high-quality data assets.*

A practical ROI framing emerged from the Q&A: one measurable impact of AI data readiness is simply the reduction in data remediation work during agentic projects. If that remediation currently takes 4–6 months per project, the savings from pre-qualifying data assets are immediately quantifiable. The ADAR CLI tool and Python library currently in development will make baseline readiness assessment systematic rather than ad hoc.

5. Panel Discussion & Q&A

Vladimir Makarov facilitated a 25-minute panel discussion that drew on audience questions submitted throughout the session. Four themes emerged with enough depth to warrant reporting separately from the formal presentations.

5.1 The Economics of Agentic AI

Vladimir framed the opening question around a tension running through several audience submissions: how do you measure the value of an AI agent, and what drives the decision to deploy one rather than a cheaper deterministic alternative?

Nicholas Larus-Stone offered the most direct answer, and the most deliberately provocative one: most life sciences organisations are vastly underspending on AI. His argument rested on the cost asymmetry between inference and experimentation. A wet lab experiment costs tens or hundreds of thousands of pounds; asking an AI agent to check whether it has already been done costs pennies. He cited a real case where an agent identified a redundant experiment and saved £100,000 — a figure that would justify thousands of agent queries at current inference costs. The practical implication: organisations that are reluctant to authorise AI agent usage on cost grounds are applying the wrong cost comparator.

> *If you are unwilling to spend £10 on having an AI agent do prep work before you go spend £50,000 to run an experiment, that cost-reward maths makes no sense to me.*

Edaeni Hamid offered an important counterweight. The cost of using an agent in production is often not the inference cost — it is the data remediation cost incurred before the agent can be deployed effectively. If making a data asset AI-ready requires six months of multiple FTE time, that cost needs to appear in the business case from the outset. The practical implication drawn from the Genentech/Roche experience: the first question to ask before starting an agentic project is not 'can we build this?' but 'is the data in a state where building this is worth it?' Arindam Sett reinforced this: one of the most immediately measurable ROI signals for the ADAR initiative is the reduction in mid-project data remediation time across the portfolio.

5.2 Deterministic vs. Non-Deterministic Systems

An audience question from Mishrek Koposki raised a productive tension: if agentic systems have some degree of creative freedom when unconstrained, can that non-determinism be deliberately exploited for drug discovery — specifically for hypothesis generation — rather than simply managed as a risk?

Nicholas Larus-Stone drew a clean distinction: non-determinism is valuable for one-off, exploratory, ad hoc analyses where the goal is to surface novel possibilities. Deterministic pipelines are correct for processes that are run repeatedly and where consistency and auditability matter. The decision about which mode to use should be made explicitly and documented — it is a design choice, not a default.

Josefa Stoisser agreed and added a critical UI/UX dimension: whatever the underlying mode, users must be able to see clearly whether an output is deterministic, probabilistic, or in the more speculative creative register. Conflating these in a single interface — presenting a potentially hallucinated hypothesis with the same visual weight as a result from a validated pipeline — is itself a governance failure. The obligation to be transparent about the nature of an output applies regardless of how useful the output turns out to be.

Vladimir extended the thread with a research interest of his own: is it possible to build agents that assess the quality of other agents' outputs — distinguishing likely hallucinations from well-grounded inferences? Nicholas noted that stacking LLMs can increase confidence (if one model is 99% accurate, cross-checking with a second raises effective reliability), but no combination of LLMs achieves certainty. The moment any LLM enters the loop, some non-determinism is introduced. The implication for governance is that confidence quantification and output classification — flagging outputs by their epistemic status — is a first-class design requirement rather than an afterthought.

5.3 Statistical Validity of Agent-Generated Code

Darren Green raised a question that cuts to the heart of agentic AI in research settings: when an agent generates statistical or bioinformatics code, who validates that the approach is methodologically sound? The concern is not primarily about bugs — it is about agents confidently recommending the wrong statistical test to a scientist who lacks the expertise to recognise the error.

Josefa Stoisser described Novo Nordisk's current solution: any code produced by the DREX-seq agent is routed to internal statisticians for review before it can be promoted to an approved workflow or skill. Biologists create the questions; statisticians validate the methods; only validated methods become reusable agent capabilities. This separates the creative exploration phase (where agent-generated code is provisional) from the production phase (where only statistician-approved pipelines run). It also creates a natural audit trail: the statistician's approval is the accountability record.

Nicholas Larus-Stone extended this to a broader cultural point. The value of having AI systems that are interpretable and traceable — showing the code they produce rather than hiding it — is precisely that it enables this kind of expert review. An opaque agent that returns results without exposing its reasoning cannot be validated in this way. Interpretability is not just a technical property; it is a prerequisite for the human oversight structures that make agents trustworthy in regulated environments. He also noted that LLMs are well-suited to one-off statistical explorations but that the natural next step — once a method is proven — is to encode it as a deterministic pipeline, which is cheaper, faster, and more auditable.

5.4 Agent-to-Agent Ecosystems and Interoperability

Vladimir closed the Q&A with a question about the longer-term architecture of agentic AI: how do agents become discoverable by other agents, communicate between themselves, and interoperate across organisational boundaries?

Nicholas Larus-Stone was candid about the current state: the problem is largely unsolved. MCP (Model Context Protocol) provides a partial answer — Benchling's AI scientist operates as both an MCP client and an MCP server, enabling interaction with other agents behind their own MCP servers. But MCP was not designed for autonomous agent-to-agent communication; it is a tool-calling protocol that can be pressed into service for this purpose. Google's agent-to-agent (A2A) protocol exists but has not achieved significant adoption. Autonomous agent discovery — where an agent can find and commission another agent without human orchestration — remains genuinely open.

Arindam Sett added the human-in-the-loop dimension, which he argued is the more pressing problem. In a single-agent system, the human sits at the end of the workflow and validates the output — manageable. In a multi-agent system where agents are communicating with each other, the number of potential breakpoints multiplies with each additional agent, and tracing failures becomes substantially harder. He used the analogy of a tumour board in oncology: specialist agents representing different domains (genomics, pathology, pharmacology) converging on a recommendation — a genuinely useful model, but one that requires the ecosystem to mature considerably before the governance and oversight mechanisms are fit for purpose. His conclusion: the field is moving toward this architecture, but peripheral decisions about frameworks, standards, and human oversight placement in agent networks need to be resolved before it is safe to deploy at scale.

> *Even with agent-to-agent communication, where does the human sit? We multiply the number of breakpoints when agents talk to each other. The ecosystem has to mature before we get to the point where we fire off a question and specialist agent networks converge on an answer.*

6. Key Themes and Conclusions

A programme that spanned AI scientist architecture, biomedical reasoning evaluation, enterprise governance, and data readiness converged on a surprisingly unified set of conclusions. The following themes run through all four presentations.

#	Theme	Observation
1	Structured data is still the foundation	Companies that prioritise structured data realise AI benefits faster. LLMs can help remediate messy data but cannot compensate for its absence. The Genentech/Roche ADAR framework and Benchling's architecture both rest on this.
2	Silent errors are the primary risk	Josefa Stoisser's silent misframing concept, Steensgaard Gade's name-prior retrieval masquerading as reasoning, and the ADAR data quality gaps all describe the same underlying problem: AI failure that does not announce itself. Designing for detectability is as important as designing for accuracy.
3	Trust must be engineered, not assumed	All four presentations described specific engineering choices for building trust: citations and grounding (Benchling); approved workflow constraints and explicit framing disclosure (Novo Nordisk); full audit trails and performance-

#	Theme	Observation
		threshold validation (Sapio); data quality scoring and certification (Genentech/Roche).
4	Apparent reasoning capability may be retrieval	Steensgaard Gade's results show that LLM performance on biomedical link prediction is largely driven by pre-training memorisation of well-studied biology. In domains where novelty matters — early drug discovery — this is a significant limitation that current benchmarks frequently obscure.
5	Human oversight must be purpose-designed	Human-in-the-loop is not a single control; it is a design variable. Novo Nordisk's plan approval gate, Sapio's four-level autonomy framework, and Benchling's proactive decision surfacing all represent different implementations of the same principle: the right human involvement at the right point in the workflow.
6	Governance must be integrated, not bolted on	Sapio's approach of integrating AI governance within existing ISO and validation frameworks — rather than creating a separate AI compliance layer — reflects hard-won experience. AI governance silos create inconsistency and burden. The same applies to data governance: ADAR is designed as a pre-project activity, not a remediation exercise.
7	Multi-LLM architectures are becoming standard	Both Benchling (for performance and sovereignty) and Novo Nordisk (implicitly, through specialist model comparisons) point toward multi-LLM approaches. Single-provider dependency creates both intelligence and commercial risk.
8	Agentic creative freedom requires explicit governance	The Q&A surfaced a productive question about whether non-determinism can be used deliberately for drug discovery (novel hypothesis generation) rather than managed as a risk. The consensus: yes, but the user must know whether they are in a deterministic or creative mode. Transparency about the nature of an output is a governance requirement, not a nice-to-have.
9	Economics of agents are favourable but underappreciated	The cost of asking an AI agent to review prior art before running a £50,000 wet lab experiment is trivial by comparison. The value loss occurs not in agent cost but in low adoption and poor data quality. Measuring ROI requires tracking adoption rates and remediation time, not just model accuracy.
10	Ecosystem interoperability is unresolved	The Q&A discussion of agent-to-agent protocols (A2A) surfaced an honest assessment: MCP gets organisations part of the way, but autonomous agent discovery and communication is an unsolved problem. Multi-agent systems multiply breakpoints in proportion to complexity, and human oversight placement in agent networks remains an open design question.

6.1 A note on evaluation methodology

A thread connecting Steensgaard Gade's research and the broader seminar discussion is the inadequacy of prevailing evaluation practices. Benchmarks that rely on well-studied entities reward memorisation over reasoning. LLM-as-judge methods can produce high apparent precision for systems that fail under rigorous external evaluation. Model accuracy metrics do not predict task adoption or business impact. The Novo Nordisk approach — building evaluation datasets from real user traces and domain expert feedback — is a methodological advance that deserves wider adoption. For the Pistoia Alliance community, developing shared, rigorous evaluation frameworks for common life sciences AI tasks remains an important pre-competitive opportunity.

6.2 Looking ahead

The next events in the LSAIE programme will continue after the summer break and will include the August round table on "Practical Application: Change Management, Adoption, GxP Compliance & Validation Requirements" — a natural follow-on to the themes raised here — and the Boston AI-Ready Data Workshop in November 2026, which will address the FAIR-to-AI-readiness gap that the Genentech/Roche presentation illuminated from a different angle. The ADAR framework and the Pistoia Alliance's AI-Ready Data initiative are addressing structurally similar problems; there is clear scope for collaboration between the two workstreams.

The overarching message from this seminar is that the agentic AI wave in life sciences is real and accelerating — but that the organisations delivering durable value are the ones that have addressed the unglamorous foundational work: data quality, governance architecture, evaluation rigour, and honest measurement. The gap between those organisations and those still discovering that their data was not ready halfway through a project is widening, and it is a gap the Pistoia Alliance is unusually well-placed to help close through pre-competitive work.

Review prepared by Rob Gill for the Pistoia Alliance Life Science AI Exchange programme, June 2026.