

Scaling AI In Chemistry R&D: The Data Governance Imperative

Maintaining the scientific accuracy R&D decisions demand

Table of Contents

3

Executive summary

4

AI adoption in chemistry is expanding

5

Why AI adoption is uniquely challenging
in chemistry

7

Developing governance-ready scientific AI
infrastructure

9

Scientifically defensible AI as a competitive
advantage

10

References

Executive summary

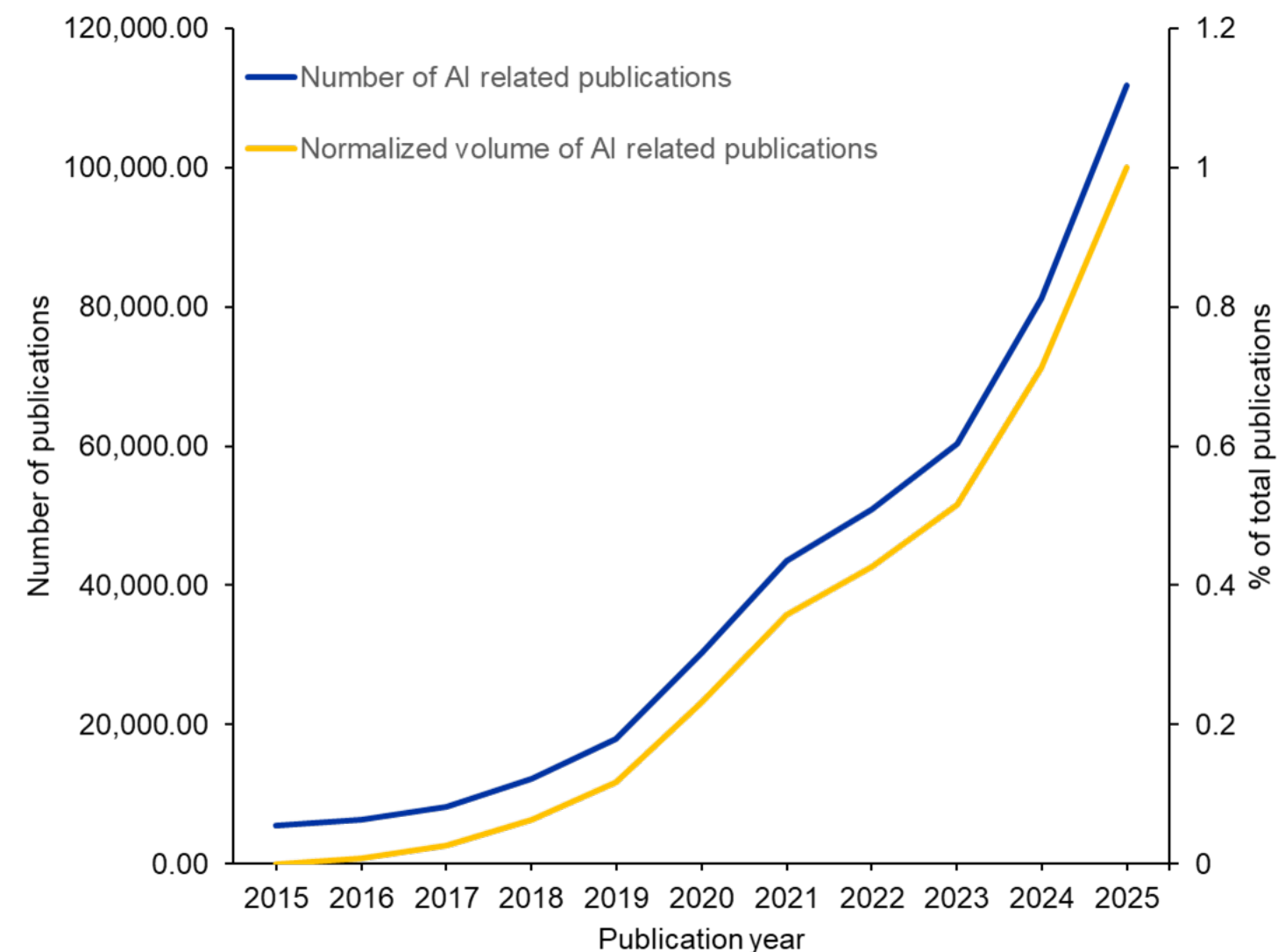
AI is rapidly entering R&D organizations, but most efforts stall before delivering enterprise value. The root cause is structural: chemistry's accuracy demands are exceptionally high, while the underlying scientific data is complex and inconsistent. Without curated data, human validation, and clear governance, AI outputs cannot be trusted or defended. Organizations that address these gaps can avoid pilot stagnation and scale AI initiatives.



AI adoption in chemistry is expanding

Chemistry R&D is seeing unprecedented interest and investment in generative AI and predictive modeling, driving massive growth. More specifically, AI in the chemicals market is predicted to have a compound annual growth rate (CAGR) of up to 36.9% over the next five-to-eight years, with current areas of interest spanning molecular design, reaction prediction, materials design, drug design, and quantum chemistry.¹⁻⁴ The substantial increase in journal publications relating to AI in chemistry over the last decade exemplifies the scale of current scientific interest in the field.

Figure 1. Annual publication volume of AI-related chemistry journal articles from 2015-2025



That scientific momentum intersects with broader enterprise mandates to operationalize AI. McKinsey's 2024 global survey reported that, across industries, 65% of organizations were regularly using generative AI in at least one business function, while overall AI adoption had risen to 72%.⁵ If growth rates continue as predicted, AI adoption in the chemistry industry is expected to contribute \$3.5 trillion to the global economy by 2030.⁶

Emerging use cases in the sector are also becoming clearer as the role of autonomous or agent-based scientific workflows grows.⁷ Industry analysis points to the growing use of AI across knowledge management, discovery and design, and development, scale-up and deployment (Figure 2).⁸

Figure 2. AI use cases across chemistry organizations.

Knowledge Management	Discovery and Design	Development, Scale-up, and Deployment
Literature-based insight generation	Molecule discovery and design	Formulation screening and optimization
Patent and prior art analysis	De novo synthesis design	Predictive scale-up insights
Competitive intelligence	Retrosynthesis and route planning	Manufacturing and supply chain intelligence
Scientific knowledge extraction	AI-assisted ideation	Regulatory, safety validation, and quality assurance
Semantic research		Decision support
Foundational layer Expert-curated data Ontologies Structured scientific knowledge		

For deeper insights into opportunities for AI within chemistry, download our white paper:

[Artificial Intelligence in Chemistry: Current Landscape and Future Opportunities](#)

The industry's explosive growth is driven by real promise. AI has the potential to speed up the identification of reaction precedents, automate the extraction of experimental data from patents and literature, accelerate hypothesis generation, and broaden access to internal and external chemical knowledge.^{9,10} The impact is already being felt across the lifecycle, with 81% of healthcare organizations reporting revenue growth from AI adoption in knowledge management, AI-driven drug discovery reducing preclinical development time by an average of 35%, and AI interventions improving pharmaceutical manufacturing yield by an average of 20-30%.^{11,12}

However, realizing these benefits depends on whether AI outputs can meet the standards that chemistry demands. This white paper explores the unique challenges AI faces in chemistry R&D and recommends a strategic approach for building governance-ready scientific AI infrastructure.

Why AI adoption is uniquely challenging in chemistry

The difficulty of adopting AI in chemistry without compromising scientific integrity can be distilled into three interrelated factors: the complexity of the underlying data, the accuracy demands of chemistry, and the challenge of sustaining both at an enterprise scale.

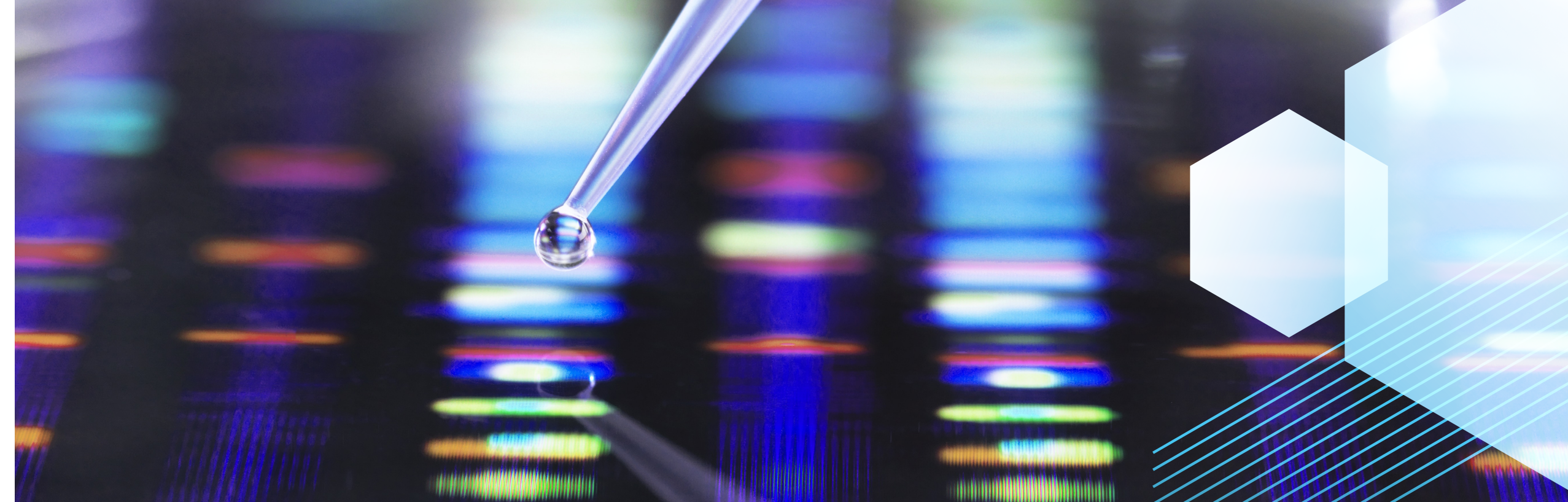
Complex data makes accuracy hard to achieve

Chemistry data is messy, inconsistent, and context-dependent. Chemical structures are frequently represented in non-standardized ways across identifier systems, with CAS Registry Numbers®, synonyms, and database entries for the same substance often failing to align. One 2023 study comparing chemical identifiers across ECHA, PubChem, and the CompTox Chemicals Dashboard found 736 inconsistent entries and 48 additional cases where substance identity was unclear among more than 8,500 mono-constituent organic substances.¹³ As more complex chemistry data, such as tautomeric forms, stereochemical assignments, and contextual reaction data, are layered on top, errors compound through these details.

“The complexity in chemistry data isn’t visible from the outside. It’s stereochemistry, tautomers, conflicting identifiers across systems — the kind of detail that takes a trained chemist to reconcile, and the kind of detail AI gets wrong confidently and silently.”

[Angela Zhou](#), Ph.D., Executive Editor and Lead Scientist in Life Sciences, CAS.

Data assembly methods can also influence accuracy, with a 2024 study highlighting that reaction datasets used for machine learning are often assembled through text mining or manual recording, both of which can introduce errors such as missing reactants, incorrect atom mapping, and even erroneous reactions.¹⁴ As such, when organizations attempt to reconcile what they find across literature, patents, and internal experimental records, they regularly encounter conflicting information that requires expert judgment to resolve.



Chemistry has a uniquely high bar for accuracy

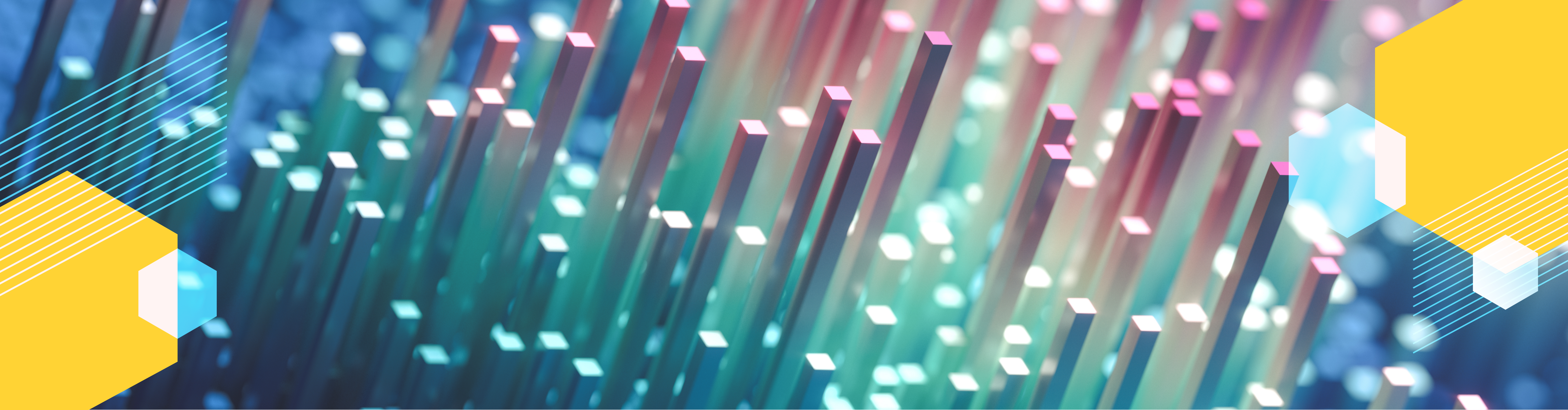
In most enterprise AI applications, small inaccuracies are tolerable, but in chemistry, the margin for error is exceptionally narrow. When AI systems are trained or prompted on flawed data, the result is plausible-but-unsupported outputs, including:

- Hallucinated or misattributed citations
- Invented reaction yields that appear chemically reasonable
- Incorrect catalyst or reagent pairings
- Decontextualized extraction of reaction conditions from patents
- Confusion between laboratory-scale precedent and industrial scalability

These outputs are especially dangerous because they may not be immediately detectable, allowing them to easily propagate into downstream predictions and create layered risk.

In R&D, teams may commit time and resources to synthetic pathways that have no real-world precedent, only discovering the problem late in the process. Downstream, flawed data can lead to incorrect assumptions about scalability or impurity formation, resulting in costly process-level failures. The consequences deepen further when safety and regulatory decisions rest on the same compromised outputs, and hazard classifications or compliance requirements are missed. At the governance level, the cumulative effect is an organization that cannot credibly defend its AI-assisted decisions under audit.

Consequently, to be usable in practice, AI outputs in chemistry must remain structurally accurate, grounded in experimental context, traceable to verified sources, and defensible under audit and regulatory review.^{15,16}



Enterprise-scale requirements outstrip pilot capabilities

Many organizations have run promising AI pilots. However, pilots tend to succeed because the work required to counteract accuracy issues is manageable in that context, with a limited project scope, one-off data preparation, and heavy expert oversight.

At enterprise scale, the volume, complexity, and coordination demands grow dramatically, meaning organizations must commit to:

- Continuous reaction data normalization.
- More intensive expert review for structure and condition accuracy.
- Ongoing ontology alignment across chemical entities.
- Monitoring for drift as new chemistry emerges.
- Cross-functional coordination between R&D, IT, and compliance.

As such, the manual oversight that was appropriate in a pilot becomes unsustainable, placing a significant additional burden on teams, particularly chemistry domain experts, and diverting their focus from critical innovation work. As AI adoption scales, governance scrutiny also increases. Boards and regulators expecting explainability and documented data lineage further burden teams that are already stretched thin.

This challenge translates to a growing trend of AI initiatives becoming stuck in pilots. According to a McKinsey 2025 global survey, nearly two-thirds of organizations have not yet begun scaling AI across the organization.¹⁷ In the chemical industry specifically, AI opportunities often depend on third-party or internally sourced unstructured data, deeper IT and digital infrastructure, downstream process redesign, and significant change management, which can lead to program delays and even termination.¹⁸

These barriers point to a common underlying gap: the infrastructure needed to sustain AI quality, governing, and oversight at scale has not been built.

Developing governance-ready scientific AI infrastructure

For AI-driven decisions to demonstrate traceability, transparency, and human oversight at scale, organizations need a fundamentally different foundation for AI deployment. This means AI systems influencing chemistry decisions must be built on curated, verifiable data foundations and governed with the same rigor as other scientific processes, meeting expectations for documentation, auditability, accountability, and reproducibility of both data and model outputs.

“In chemistry, an AI output should be held to the same standard as a reproducible, peer-reviewed scientific result. If it can’t be traced, validated, and defended, it isn’t ready to drive a decision.”

[Angela Zhou](#), Ph.D., Executive Editor and Lead Scientist in Life Sciences, CAS.



Clear accountability for AI-assisted outputs

When AI contributes to a chemistry decision, it must be clear who is responsible for that output. Without defined ownership, errors default to being everyone’s problem and no one’s responsibility. In practice, this means assigning named accountability for review and sign-off on AI-assisted recommendations, so that when a synthesis route, hazard classification, or regulatory submission draws on AI-generated data, there is an identifiable expert who has validated the basis for that decision.

Defined expert validation checkpoints

AI can accelerate chemistry workflows, but it cannot replace the expert judgment needed to catch the kinds of errors described earlier in this paper. Expert human review checkpoints must be embedded at defined stages, including:

- Validation of reaction extraction from patents and publications
- Resolution of ambiguous structure representations
- Ongoing updates to reflect new synthetic methods and safety data
- Quality control to prevent propagation of outdated or unsupported information

Structured data provenance and lineage

If an AI-assisted decision is questioned, the organization must be able to trace the output back through the model to the underlying data. That requires a documented record of where the data originated, how it was processed, what transformations or curation steps were applied, and which version of a model produced the output. Without this lineage, defensibility breaks down, which is insufficient under audit or regulatory review.

Integration with compliance and regulatory workflows

AI systems that operate outside existing quality and regulatory frameworks create a parallel, ungoverned decision pathway. For chemistry organizations subject to regulatory oversight, AI outputs must feed into the same documentation, validation, and approval processes that govern other scientific work. This means designing AI workflows so that their outputs are captured within existing compliance systems rather than requiring manual re-entry or post-hoc reconciliation.



Sustained stewardship of curated chemical data foundations

The data that underpins chemistry AI is not static. New compounds are synthesized, reaction conditions are refined, safety profiles are updated, and regulatory requirements evolve. Because of this, organizations must actively maintain the curated data foundations on which AI systems depend to prevent model degradation over time. This sustained stewardship requires dedicated ownership, defined update cycles, and quality standards that treat chemical data as a living asset.

“Curated chemical data isn’t infrastructure you build once; it’s a living asset. The organizations that treat it that way are the ones whose AI initiatives will still be working five years from now.”

[Angela Zhou](#), Ph.D., Executive Editor and Lead Scientist in Life Sciences, CAS.

From experimental pilots to governance-ready AI

At its core, the shift from experimentation to governance-ready AI is a progression of capability maturity across data, workflows, and oversight (Table 1).

Table 1. Maturity model for building governance-ready AI in chemistry

Core Capability	Ad-hoc/ Experimental AI	→ To reach Operational AI	Operational AI	→ To reach Governance-Ready AI	Governance-Ready AI
Data Provenance and Lineage	Data sources fragmented, lineage unclear + inconsistently documented	Identify critical datasets and implement basic metadata + lineage tracking	Key datasets tracked with partial lineage visibility	Extend lineage to full workflows (data → model training → prediction output) with version control and enforce standards	End-to-end lineage with full traceability across systems, model versioning, and regulatory-compliant documentation
Human Validation	Informal or inconsistent expert review with limited documentation or reproducibility	Define risk-based review criteria and introduce structured checkpoint workflows + documented rationale	Validation checkpoints in place for high-impact use cases with refined reviewer qualifications	Formalize sign-off authority and standardize validation across all workflows	Embedded, mandatory validation with clear accountability and audit trails
Compliance Integration	AI workflows operate outside regulatory processes	Map AI use cases to existing QA/compliance frameworks and address major gaps	Partial alignment with QA/compliance processes, including document risk assessments	Fully embed AI into validation, documentation, and approval processes	Fully integrated with good practice standards (GMP), quality, and regulatory systems
Data Stewardship	Project-based ownership, inconsistent data quality	Assign data owners and define quality standards for priority datasets	Centralized datasets with emerging governance practices and defined quality metrics	Establish formal stewardship roles, processes, lifecycle management, and chemical-specific standards	Sustained stewardship with enterprise-wide governance and continuous quality monitoring
Model Governance	Models developed ad hoc, with limited documentation of architecture/assumptions	Document model architectures, training approaches, and performance tracking	Model registry established with version control and performance tracking	Implement formal model validation protocols, bias testing, and performance monitoring	Complete model lifecycle management with approval workflows, continuous monitoring, and regulatory-grade documentation

Organizations that deliberately invest in moving across these maturity stages, rather than stopping at initial experimentation, build the foundation for AI-assisted decisions that are trusted, traceable, and aligned with both scientific and regulatory expectations.



Scientifically defensible AI as a competitive advantage

Meeting chemistry's bar for accuracy is feasible in controlled pilots, but sustaining it at enterprise scale demands purpose-built infrastructure: curated scientific data, expert validation, and structured traceability. When that foundation is in place, built-in defensibility increases audit readiness, cross-functional confidence grows, and adoption can move more easily from tentative experimentation to embedded practice across R&D and operations.

"Chemistry organizations have always competed on scientific rigor. AI raises the stakes. The winners will be the ones who hold the line on integrity while everyone else is racing on speed."

[Angela Zhou](#), Ph.D., Executive Editor and Lead Scientist in Life Sciences, CAS.

Organizations that invest in this foundation will spend less time correcting flawed outputs, less time reconstructing the rationale behind prior decisions, and more time advancing meaningful R&D. Ultimately, in a sector where the cost of being wrong compounds faster than the cost of being slow, scientifically defensible AI becomes a defining competitive advantage.

Access science-smart AI with the CAS SciFinder Discovery Platform™

The CAS SciFinder Discovery Platform combines trusted scientific data with science-smart AI to help researchers uncover connections, accelerate discovery, and make more confident decisions. Powered by the CAS Content Collection™, it enables scientists to explore chemical knowledge, reaction insights, and scientific literature more efficiently across the research workflow. CAS is a division of the American Chemical Society.

[Explore more](#)

References

1. Yang L, Guo Q, Zhang L. AI-assisted chemistry research: a comprehensive analysis of evolutionary paths and hotspots through knowledge graphs. Chem Commun. 2024;60(55):6977-6987. doi:10.1039/D4CC01892C
2. Ltd CMIP. Artificial Intelligence (AI) in Chemical Market Trends, 2032. Coherent Market Insights. March 4, 2025. Accessed April 15, 2026. <https://www.coherentmarketinsights.com/industry-reports/artificial-intelligence-ai-in-chemical-market>
3. Research S. AI in Chemicals Market Size, Share, Growth & Trends by 2033. Accessed April 15, 2026. <https://straitsresearch.com/report/artificial-intelligence-in-chemicals-market>
4. AI In Chemicals Market Size, Share & Trends Report, 2030. Accessed April 15, 2026. <https://www.grandviewresearch.com/industry-analysis/ai-chemicals-market-report>
5. The state of AI in early 2024 | McKinsey. Accessed April 15, 2026. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024>
6. AI In The Chemistry Industry Statistics: Market Data Report 2026. Accessed April 15, 2026. <https://gitnux.org/ai-in-the-chemistry-industry-statistics/>
7. A review of large language models and autonomous agents in chemistry - Chemical Science (RSC Publishing) DOI:10.1039/D4SC03921A. Accessed April 15, 2026. <https://pubs.rsc.org/en/content/articlehtml/2025/sc/d4sc03921a>
8. From molecule and material discovery to customer acquisition, using AI in the chemical industry could create several promising new use cases. | McKinsey. Accessed April 15, 2026. <https://www.mckinsey.com/industries/chemicals/our-insights/how-ai-enables-new-possibilities-in-chemicals>
9. Vangala SR, Krishnan SR, Bung N, et al. Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature. J Cheminformatics. 2024;16(1):131. doi:10.1186/s13321-024-00928-8
10. Schilling-Wilhelmi M, Ríos-García M, Shabih S, et al. From text to insight: large language models for chemical data extraction. Chem Soc Rev. 2025;54(3):1125-1150. doi:10.1039/D4CS00913D
11. AI Pharmaceutical Industry Statistics | Fact-Checked 2026. Accessed June 5, 2026. <https://worldmetrics.org/ai-pharmaceutical-industry-statistics/>
12. AI in Healthcare 2025 — NVIDIA Report Highlights Adoption Rates, Revenue Growth, and Key Challenges | Navalink AI. Accessed June 5, 2026. <https://www.navalink.ai/article/ai-in-healthcare-2025-nvidia-report-analysis>
13. Getting the SMILES right: identifying inconsistent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard. Environ Sci Adv. 2023;2(4):612-621. doi:10.1039/d2va00225f
14. Chen LY, Li YP. AutoTemplate: enhancing chemical reaction datasets for machine learning applications in organic chemistry. J Cheminformatics. 2024;16(1):74. doi:10.1186/s13321-024-00869-2
15. Rules for trustworthy artificial intelligence in the EU | EUR-Lex. August 2, 2025. Accessed April 15, 2026. <https://eur-lex.europa.eu/EN/legal-content/summary/rules-for-trustworthy-artificial-intelligence-in-the-eu.html>
16. Gant TW, Boxall A, Burgwinkel D, et al. Building trust in the integration of artificial intelligence into chemical risk assessment: findings from the 2024 ECETOC workshop. Arch Toxicol. Published online February 17, 2026. doi:10.1007/s00204-025-04286-8
17. The State of AI: Global Survey 2025 | McKinsey. Accessed April 15, 2026. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

CAS connects the world's scientific knowledge to accelerate breakthroughs that improve lives. We empower global innovators to efficiently navigate today's complex data landscape and make confident decisions in each phase of the innovation journey. As a specialist in scientific knowledge management, our team builds the largest authoritative collection of human-curated scientific data in the world and provides essential information solutions, services, and expertise. Scientists, patent professionals, and business leaders across industries rely on CAS to help them uncover opportunities, mitigate risks, and unlock shared knowledge so they can get from inspiration to innovation faster. CAS is a division of the American Chemical Society.

Connect with us at cas.org