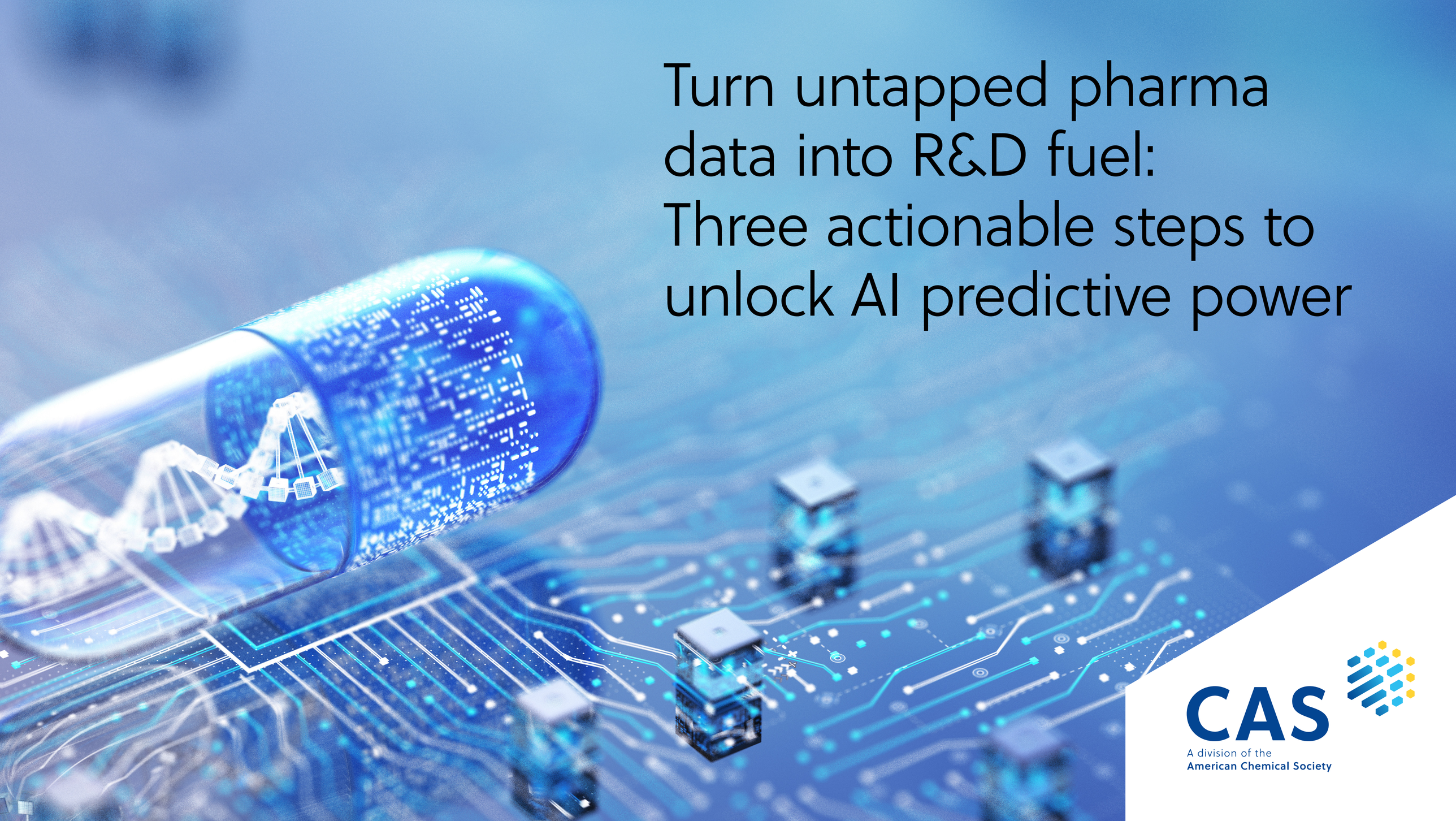
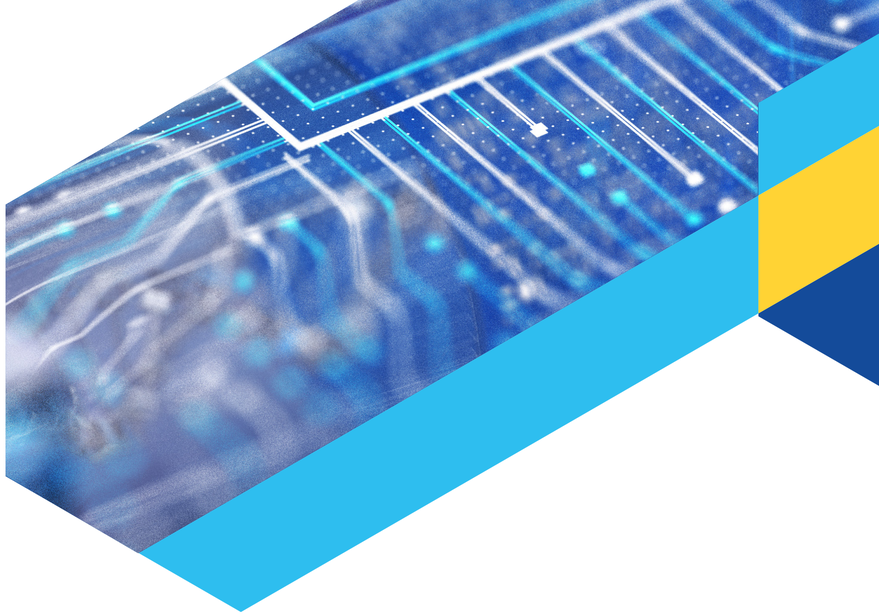


Turn untapped pharma data into R&D fuel: Three actionable steps to unlock AI predictive power

CAS
A division of the
American Chemical Society

Table of Contents



In the era of AI-driven R&D, pharmaceutical companies can no longer afford to treat data management as optional. Without the right foundation, even the most advanced algorithms struggle to deliver reliable predictions, weakening confidence in insights that guide critical R&D decisions and competitive edge.

As agentic and generative AI models gain traction, they bring heightened demands for pharma data that is not only structured and scalable but also ready to support automation and advanced analytics. Yet, most pharmaceutical organizations lack the combined competencies to unlock AI's real predictive power.

Domain expertise to ensure the science is correct.	**Content expertise** to ensure the data foundation is robust.	**Technical/algorithm expertise** to ensure the model is reliable.
Deep knowledge of therapeutic areas, segments, and workflows unique to your organization and your pipeline.	Knowledge of scientific data modeling, harmonization, and knowledge management to structure your R&D data and bridge gaps with high-quality scientific content.	Knowledge of algorithm development, computational models for prediction, and infrastructure to mine, analyze, and model complex R&D data.

Pharma's strength in domain knowledge is clear, but many organizations fall short in the content and computational expertise required to reach AI's full potential. Too often, internal R&D data remain fragmented or inconsistent, while licensed external content proves difficult to integrate. The result? AI models fueled by limited inputs inevitably create blind spots, which introduce bias and generate predictions that lack the reliability needed to guide confident R&D decisions.

This white paper outlines a practical roadmap to advance your AI and data strategy. By connecting siloed systems, bringing structure to research data, and enriching internal knowledge with authoritative external content, pharma leaders can power up their predictive AI initiatives.

1

Unlock organization-wide AI potential by connecting siloed data

The cost of data silos for AI in pharma

Data silos prevent AI from accessing the full scope of knowledge available in your data landscape, forcing models to train on incomplete datasets and operate at a fraction of their full potential. How does that reflect on your business strategy?

AI model training

- Models trained on incomplete datasets inherit blind spots and poor pattern recognition, weakening their ability to find meaning in complex R&D data.
- Reduce the overall return on digital and R&D investments.

AI model output

- Fragmented data creates bias and gaps, which lead to missed connections and unreliable insights.
- Missed therapeutic opportunities and slower innovation, reducing competitive advantage.

Why pharma R&D struggles with data fragmentation

1. Multi-domain R&D workflows create silos

In pharma organizations, chemistry, biology, preclinical research teams, and other business units operate in parallel, generating unique datasets often stored in disconnected systems such as:

- Electronic lab notebooks (ELNs).
- Dashboards and internal databases.
- Data lakes and other specialized infrastructure.

The sheer complexity of pharma R&D workflows makes it difficult to keep track of data generation, storage, and integration, which slows knowledge sharing and retrieval across the board. In addition to internal silos, mergers and acquisitions can quickly introduce another layer of complexity. Post-acquisition integration often brings together organizations with different data silos, standards, and systems, making it even harder to consolidate historical information and unify scientific knowledge.

This fragmentation prevents AI models from accessing the full scope of your institutional knowledge, limiting their ability to operate to their full potential. However, with so many moving parts, consolidating information into a usable foundation not only requires intensive manual work but also advanced knowledge management expertise.

2. R&D data diversity slows consolidation

Modern drug discovery and development require an unmatched spectrum of scientific data, including:

- DNA and RNA sequences.
- Chemical structures and reactions.
- Formulations and ingredients.
- Antibodies and proteins information.
- Biomarkers and toxicity results.
- Preclinical assay data.

While each dataset is valuable on its own, the real power of scientific data comes from linking knowledge across labs and business units. Only a connected view gives AI models the full scientific context for reliable predictions and decision-making. However, creating a single, shared view of pharma R&D data requires researchers to manually clean and consolidate information, diverting time and focus from high-value scientific work.

How CAS Custom Services℠ can help:

CAS Custom Services helps organizations unify information across labs, business units, and systems by consolidating fragmented data into a connected, centralized platform. Our knowledge management and domain experts design custom data systems with advanced search capabilities to allow your teams to access and leverage critical information faster. By lifting the burden of manual, resource-intensive data consolidation, CAS helps researchers to focus less on data cleanup and more on advancing high-value scientific work.

From hours to minutes: Accelerating data retrieval with CAS custom solutions

A large health-tech organization partnered with CAS Custom Services to convert decades of research data into searchable assets. Thanks to a custom knowledge management system designed with advanced search capabilities, researchers can now quickly tap into the organization's full breadth of R&D data, accelerating the innovation of new health-tech solutions.[1]

"One search took a senior scientist ~8 hours; now it takes minutes with scientifically aware search from CAS."

Company estimate

Beyond internal data: Bridging the gaps for stronger insights

Breaking down internal silos is only part of the challenge. To build a complete and reliable foundation for AI, pharmaceutical R&D teams must also integrate licensed and third-party datasets with their internal sources. When external data remains disconnected, AI models are left with an incomplete view of the scientific landscape. These gaps introduce blind spots that weaken predictive accuracy and limit the model's ability to support high-stakes R&D decisions.

CAS knowledge management experts help bridge these gaps by applying unique identifiers, curated connections, and unmatched scientific expertise to harmonize disparate sources. By linking regulatory status, sourcing options, formulations, reactions, structural information, and more, CAS Custom Services helps you integrate external data seamlessly so you can enrich your AI foundation for stronger, reliable predictive insights.

2

Fix the hidden data problems weakening AI prediction

Unstructured R&D data undermines AI predictive power

Even when data is accessible, structural inconsistencies across teams and business units make it difficult to align R&D information to boost AI predictive models. Without standardization, AI models struggle to generate evidence-based insights that streamline decisions and pipelines.

AI model training

- Models trained on unstructured and unverified data introduce errors and bias. Without retrieval augmented generation (RAG) in place, finite training datasets can limit long-term performance and efficacy as new findings/literature become available.
- Reduce the return on digital and R&D investments.

AI model output

- Unstructured data and insufficient context degrade prediction accuracy.
- Skewed portfolio prioritization decisions, missed opportunities, and threatened market share.

Why are pharmaceutical companies prone to data chaos?

1. Pharma R&D data lacks consistent structure

R&D data comes in many shapes and sizes, and some of them are simply incompatible. From raw instrument outputs to spreadsheets and images, pharma teams work in silos with different data types, file formats, and naming conventions, reducing organizational knowledge's value and weakening AI's predictive power across the pipeline.

Different structure, including raw instrument outputs, sequence formats, spreadsheets, and image files.	**Varied terminologies,** including abbreviations, chemical identifiers, and protein and gene names across labs and systems.	**Inconsistent reference standards,** including Fahrenheit vs. Celsius, pounds vs. kilograms, and other measurement metrics.

2. Pharma R&D data often lacks integrity

Capturing the depth of pharma R&D operations requires complete, detailed metadata. Metadata provides the context that turns raw results into usable information for your teams and AI models.

Critical metadata includes:
- Experimental parameters, including temperature, pressure, and reaction time.
- Provenance details, including who generated the data and when.
- Data lineage with detailed information about raw data processing.

While progress has been made in technology's ability to harness unstructured data, scientifically harmonized metadata concepts and frameworks that connect multi-model data are still critical for success.

When experimental parameters or provenance are missing, your AI cannot determine what is relevant or actionable. This critical context is often incomplete, making datasets hard to interpret, difficult to compare, and nearly impossible to reuse effectively. Without advanced knowledge management expertise, efforts to structure and harmonize pharma data often drain valuable time and resources while delivering little meaningful impact.

How CAS Custom Services can help:

CAS Custom Services helps organizations build a reliable, AI-ready data foundation sourced from organizational knowledge by:

- **Standardizing** disparate formats, terminologies, and languages to ensure accuracy and consistency across data sources.
- **Structuring** information into retrievable formats to boost accessibility within knowledge management systems.
- **Curating** your data and filling in the gaps with authoritative information from the CAS Content Collection™.

By reducing format discrepancies, eliminating knowledge data inconsistencies, and validating records against authoritative scientific sources, CAS helps organizations turn data chaos into AI-readiness, allowing teams to spend less time fixing data and more time advancing research.

Boosting data integrity with CAS Custom Services

A large diversified chemical company turned to CAS Custom Services to resolve issues with inconsistent and inaccurate substance data spread across multiple systems. CAS scientists harmonized records, aligned substance representations with CAS REGISTRY®, and validated entries to ensure accuracy across platforms.[2]

The collaboration eliminated manual data verification and correction, freeing 3,300 hours for researchers to focus on higher-value work.

Beyond existing data: Capturing accuracy in every new entry

Structuring institutional data is just the start. To maintain long-term integrity, your organization needs consistent, controlled data entry at the source. By deploying a unified data collection system with clear entry guidelines and quality check protocols, you can eliminate the need for manual consolidation and reduce the time your teams spend on extensive verification or AI preprocessing.

Our CAS Custom Services experts offer tailored guidance on data management processes and collection so you can standardize data entry from the source. This allows your AI models to leverage new information confidently without intensive preprocessing, accelerating the time to insights and informed decisions.

3

Enrich your data foundation to overcome AI-limiting gaps

Data homogeneity weakens AI predictive capabilities

Even well-organized internal data is often too narrow to support complex AI models. Without diverse, representative datasets, your models are flying blind. Training AI only on internal information limits the scope of learning.

In other words: models are more likely to introduce bias, miss critical variables, and generate insights that cannot support high-stakes R&D workflows.

AI model training

- Models trained on homogeneous or incomplete datasets cannot capture the full range of variables and real-world scenarios.
- Blind spots in R&D strategy; poor return on digital and R&D investments.

AI model output

- Models fueled by homogenous datasets lead to data bias, lower prediction accuracy, and skewed insights.
- Flawed performance of AI agents; poor prioritization decisions; wasted resources; and competitive setbacks.

External data is essential, but difficult to trust

Enriching your data landscape with external content is critical for discovery and development. This requires access to pharmacological databases, existing formulations and patents, clinical trial data, and more. However, not all sources can be trusted.

Open-source, third-party, and even licensed information is often unreliable. Licensed datasets can still have gaps in coverage, hampering efforts to build a unified, AI-ready foundation. Without proper data cleaning and validation, merging internal and external datasets can introduce errors into predictive models and significantly impact AI-driven workflows.

The challenge for pharma organizations

Extracting and cleaning large volumes of external multi-modal data from literature, patents, and regulatory filings is a critical step before integration, but it places a heavy burden on internal teams. This demands domain expertise and rigorous validation to ensure accuracy and reliability. When organizations lack the right support, they often lose time and resources on work that delays R&D progress instead of accelerating it.

How CAS Custom Services can help:

Building high-performing AI models requires more than large datasets; it requires the right data. CAS helps you enrich your internal data ecosystem to build the trusted foundation your AI models need to generate predictive insights that drive confident, informed decisions.

CAS Content Collection: Trusted data foundation	CAS APIs: Continuous data refresh	CAS curated datasets: Custom AI training
The CAS Content Collection combines human expertise and advanced data technology to extract, verify, and connect insights from global scientific literature, delivering a reliable, enriched foundation for R&D and AI.	CAS APIs keep your data ecosystem up to date automatically, enabling your AI models to capture new opportunities from the latest publications, patents, and regulatory filings.	CAS custom-curated datasets are not one-size-fits-all. Our experts connect and curate data in ways that match your specific R&D goals, providing the depth and context you need to make AI insights relevant and actionable.

Boosting AI prediction accuracy with custom-curated training datasets

To optimize synthesis planning for complex reactions, ChemLex partnered with CAS Custom Services to build a high-quality, custom-curated training dataset for its proprietary AI model. Using the CAS Content Collection™, CAS experts delivered structured, validated data tailored to ChemLex's needs. The result? A high-performing proprietary model delivering more accurate regioselectivity predictions than seasoned chemists.[3]

"It is very challenging for humans to collect accurate data from a vast body of literature and transform it into a structured, AI-ready format. That's why we really enjoyed the collaboration, CAS did a great job supporting us."

Co-founder and Vice President, ChemLex

Getting AI right starts with the right data

The real goal of AI enablement in pharma is not about modernizing R&D but avoiding resource-intensive trial-and-error and guiding research in the right direction from the start. Without the right data management strategy, AI can slow progress by producing misleading insights that push research in the wrong direction. To ensure AI drives smarter decisions rather than wasted effort, organizations must build a foundation that is accurate, connected, and aligned with strategic goals.

CAS Custom Services helps pharma organizations achieve that foundation by combining scientific expertise with trusted, structured content and advanced data management capabilities. From unifying internal datasets to enriching AI models with curated external information, CAS enables teams to turn untapped R&D data into the fuel that powers up trustworthy AI and measurable impact.

Reach out to find out how CAS Custom Services can help you get your data ready for successful AI workflows

Learn more at cas.org.

References

1 Health-tech unlocks hidden insights. https://www.cas.org/resources/gated-content/cas-unlocks-potential-dark-data

2 Unified data systems save company thousands of hours and dollars. https://www.cas.org/resources/gated-content/cas-data-integrity

3 Establishing new standards for AI prediction accuracy with custom training data. https://www.cas.org/resources/gated-content/chemlex-ai-model

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

SVCWHP102680ENG

CAS
A division of the
American Chemical Society