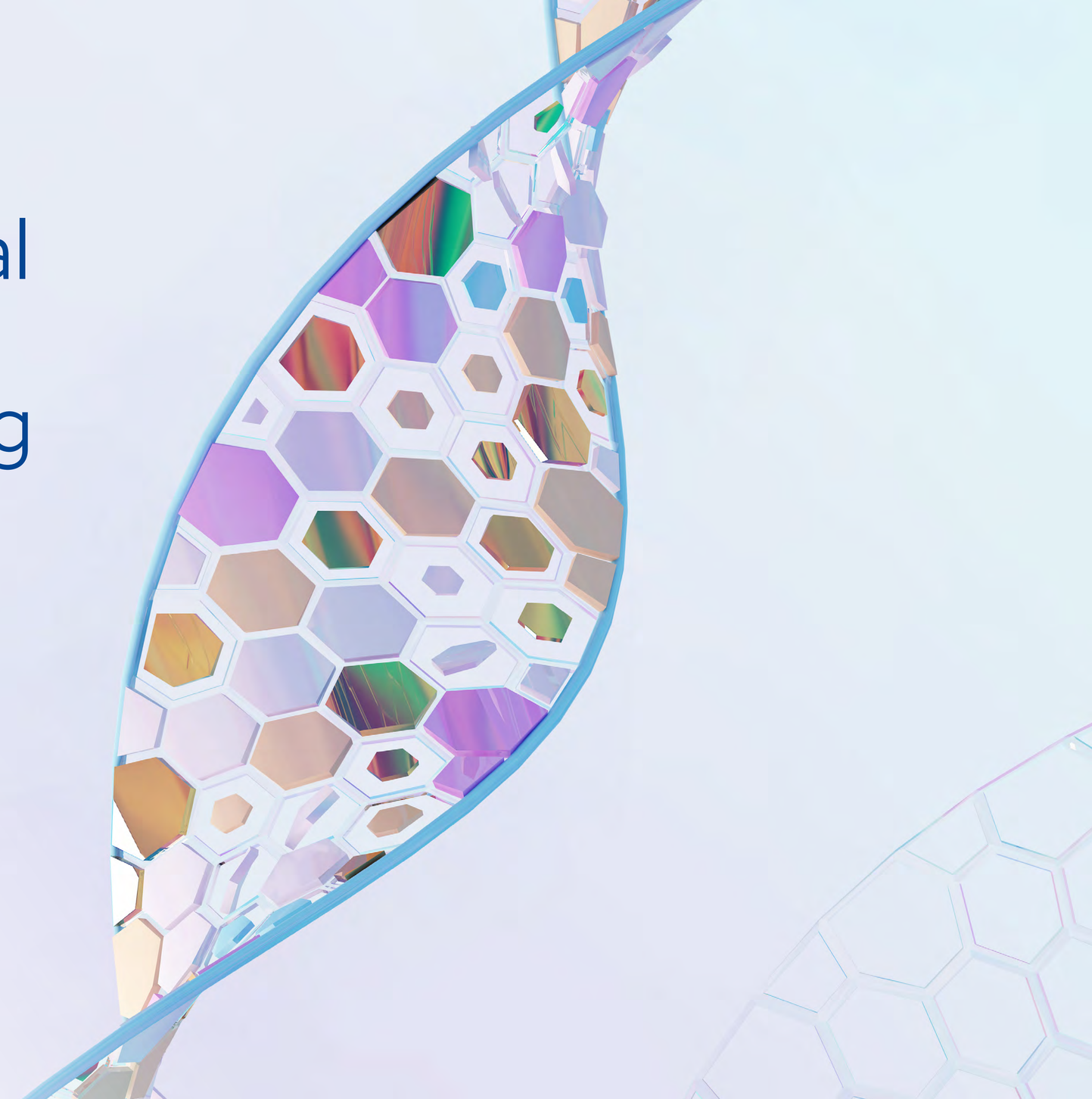
How unified biological and chemical data strengthens early drug discovery decisions



CAS
A division of the
American Chemical Society

Table of Contents

The problem with disconnected scientific data

Drug discovery relies on a tight collaboration between biologists, medicinal chemists, and other scientists. To align on early discovery decisions and move programs quickly, interdisciplinary teams must work from a shared scientific foundation. However, biological and chemical data rarely overlap. Instead, each function gathers and organizes information through its own scientific and operational lens, creating disconnected data ecosystems.

Over time, data fragmentation blurs the full picture of complex discovery frameworks, preventing teams from turning early insights into decisive action. When teams try to build on mismatched data foundations, collaboration stalls, decisions take longer, and the early stages of discovery lose critical momentum. These disconnects compromise the integrity of high-stakes workflows, reducing pipeline efficiency and return on R&D investment.

Data integration restores continuity across disciplines, providing researchers with a shared line of sight and R&D leaders with the confidence to prioritize programs grounded in consistent, cross-functional evidence. In a landscape where speed and precision determine competitive advantage, unified data empowers teams to make better-informed decisions, direct resources toward the most qualified leads, and reduce late-stage failure risks.

The same, but different: Why cross-functional data unification matters

Despite approaching problems from different angles, discovery biologists and medicinal chemists rely on the same core data: disease context, pharmacology profiles, and safety information. With this information, each team generates function-specific insights that drive early discovery decisions and guide what gets prioritized or filtered out (Table 1).

Table 1. How different foundational data types inform interdisciplinary workflows.

Foundational data type	How discovery biologists use core data	How medicinal chemists use core data
Disease context	**Identify relevant targets** by understanding biological pathways, phenotypes, and clinical associations.	**Align compound design** with therapeutic goals and disease-specific mechanisms.
Pharmacology profiles	**Prioritize targets** based on known pharmacological relevance in the disease context.	**Prioritize compounds** based on their ability to induce the desired biological response.
Bioactivity data (IC50, EC50, Ki…)	**Validate target relevance** with quantitative biological evidence of ligand responsiveness.	**Optimize compound potency** by analyzing structure-activity relationship (SAR) trends and quantitative activity data.
Selectivity/off-target data	**Refine target selection** by exploring functional overlaps that could impact therapeutic potential.	**Mitigate risk** by deprioritizing structures associated with known off-target and adverse effects.
ADME and tox data	**Anticipate biological risks** associated with target modulation, such as toxicity or poor exposure.	**Refine structures** to enhance bioavailability and mitigate the risk of liabilities and late-stage failures.

Overview of core data types and their critical role in informing biological and chemical decisions across interconnected discovery workflows.

While biologists and chemists rely on the same core information, they rarely work from the same data source. Each team uses different databases and data management systems, which impedes their ability to share knowledge across functions. When foundational data isn't connected, early decisions lack alignment, increasing the risk of operational delays and late-stage failure.

Three ways discovery data integration sets teams up for success

Unifying biological and chemical data is not just a technical improvement; it's a strategic shift that provides teams with a holistic view of the therapeutic space so they can make better-informed decisions that ripple through the pipeline. The value of discovery data integration can be summarized in three core areas.

1. Stronger cross-disciplinary alignment

From scientific literature and patents to structured databases and internal systems, discovery-relevant data originates from many disparate sources. However, these sources are not systematically validated against one another or formatted consistently, making it difficult for teams to access a reliable, consolidated evidence base (Table 2).

Table 2. Typical sources for commonly used discovery data.

Data type	Typical data sources
Disease context	Literature (e.g., PubMed), ClinicalTrials.gov, Reactome, KEGG, DisGeNET
Targets and ligands	PubChem, UniProt, ChEMBL, PDB, BindingDB, DrugBank, GtoPdb
Pharmacology profiles	DrugBank, ChEMBL, PharmGKB, NCBI, regulatory agency databases (e.g., Drugs@FDA)
Bioactivity data	Literature, ChEMBL, PubChem BioAssay, internal assay datasets
Selectivity/off-target data	Literature, PubChem BioAssay, SEA (Similarity Ensemble Approach), GtoPdb
ADME/Tox data	Literature, eTOX database, ADMETlab, HSDB in PubChem, DrugBank, Open TG-GATEs, National Library of Medicine databases, internal toxicology reports

Early discovery decisions are driven by multiple data types that reside in diverse external and internal sources, each adding critical biological and chemical context.

Consolidating information into a unified view reduces disconnects in how researchers interpret key inputs such as disease context, assay readouts, and pharmacological data. When scientific data is structured consistently across domains, discovery teams can evaluate it through a shared lens even when approaching it from different angles. This strategic shift not only helps close interpretation gaps but also accelerates drug discovery pipelines by preventing cross-functional misalignments and inherent operational delays.

Improved research focus and reduced time wasted on non-viable candidates.

Minimized back-and-forth and misinterpretation, handoff delays, and trial-and-error cycles.

Faster and better go/no-go decisions for reduced late-stage failure risks.

2. Uncovering hidden connections and opportunities

In addition to aligning researchers, unified data can reveal evidence-based insights that might otherwise remain hidden. When biology and chemistry data are connected, teams can uncover relationships between targets, compounds, pathways, and disease mechanisms that are often missed in siloed discovery environments. This level of integration helps teams identify targets with unexpected roles in disease, detect previously overlooked off-target interactions, or revisit compound series in a new therapeutic context. Such insights are valuable when exploring under-characterized biology or making decisions in crowded or high-risk therapeutic areas.

With a more comprehensive view of the scientific and therapeutic landscape, discovery teams can maximize the value of cross-functional findings by applying them in the right context at the right time. This integrated perspective de-risks programs earlier, empowers quicker and better-informed decisions, and exposes opportunities that traditional workflows routinely miss.

Faster investigation of underexplored biology or chemical space.

Earlier detection of target redundancy and potential off-target liabilities.

Enhanced strategic positioning in crowded therapeutic areas.

3. Speed and confidence for high-stakes decisions

High-stakes decisions in discovery should not rest on intuition. They require decision gates supported by consistent, comprehensive evidence that biologists and chemists can trust. When information is fragmented, teams risk missing subtle but critical mechanistic or safety liabilities that could sway decision-making and prioritize the wrong candidates.

A connected data environment minimizes reliance on siloed interpretation and provides a consolidated evidence base that streamlines triage, target selection, and program prioritization. This improves confidence in early go/no-go decisions, enabling leaders to focus on programs with the strongest foundations and operate more effectively in a highly competitive R&D landscape.

Confident program decisions grounded in consistent data.

Improved ability to triage novel programs with limited experimental evidence.

Stronger alignment between scientific evidence and strategic investment.

Unifying discovery data: What slows organizations down?

While data integration reduces rework and misalignment, achieving it is not simple. Efforts to standardize and consolidate discovery data in-house often push researchers beyond their core capabilities

The nature of scientific data makes integration far more intricate than most organizations anticipate. **Why is this?**

Scientific outputs are not designed to work together

Biological and chemical information originates from sources that were rarely built with interoperability in mind. As such, each dataset follows its own rules, structures, units, and conventions, resulting in deep inconsistencies that slow integration. Three of the most consequential inconsistencies are:

- **Inconsistent identifiers:** Proteins, genes, compounds, and diseases are labeled differently across databases, publications, vendors, and internal systems.
- **Incompatible formats:** Chemical structures, sequence data, assay results, pathway annotations, and toxicity readouts are stored in unharmonized formats.
- **Limited metadata alignment:** Including experimental conditions, protocols, endpoints, or measurement units that are unstandardized across fields, labs, and countries.

This lack of consistency results in a fragmented data environment where nothing aligns out of the box.

Researchers need access to as much information as possible so they don't miss key insights and push forward the wrong candidates. However, consolidating and standardizing existing scientific information is a huge effort, costing precious time and diverting discovery teams from selecting drug development candidates and advancing high-value programs.

Scientific data curation never ends

Discovery data is not static. New publications, patent filings, and database updates are added daily, and each new addition can alter how targets, compounds, or pathways are interpreted. As new information continually reshapes the drug discovery landscape, staying current becomes an ongoing responsibility rather than a one-time task.

Transforming a constant stream of publications, assays, and database updates into usable insight demands continuous investment in expert curation, including the involvement of humans to verify relevance, update discovery-relevant information, and interpret biological context as new evidence emerges. Most discovery teams don't have the dedicated bandwidth to manually maintain this level of curation without having to pause core research activities—and slow the entire pipeline.

Integration needs more than scientific expertise

Integrating discovery data at scale is an organizational commitment that requires more than understanding scientific content. It involves sustaining pipelines, managing identifiers, and ensuring systems keep pace with new information, requiring dedicated support across data management and IT domains. Furthermore, dedicated resources find standardizing and indexing taxonomy within a scientific field to be a huge task. Without the right expertise and infrastructure in place, scientists are compelled to divide their time between high-value research and resource-intensive data curation, thereby slowing progress and increasing operational fatigue.

Biological and chemical data unified in one platform

Integrating discovery data should not require researchers to divert significant scientific time toward harmonizing formats, resolving conflicting information, or tracking constant updates across literature and databases. R&D leaders need solutions to alleviate the burden of data integration and enable their teams to focus on advancing high-impact discovery pipelines. CAS BioFinder® was designed to address this need.

Purpose-built for drug discovery, **CAS BioFinder** helps bridge the gaps that slow progress by delivering up-to-date, structured, and expertly curated data drawn from top life sciences journals, global patent offices, and authoritative scientific databases. At the core of the platform is the CAS Content Collection™, the world's largest repository of human-curated scientific data, built from more than 110 years of experience in organizing scientific knowledge.

With biological, chemical, and pharmacological information unified across targets, ligands, diseases, pathways, biomarkers, and assays, discovery teams can work from a shared, reliable source of truth rather than a patchwork of disconnected systems.

THIS UNIFIED FOUNDATION:

- Reduces the time spent reconciling inconsistent or incomplete discovery-relevant information.
- Supports clearer and more consistent interpretation across interdisciplinary teams.
- Drives confident early-stage workflows by grounding decisions in an expert-curated evidence base.
- Improves communication and alignment between biologists and medicinal chemists.

By resolving data fragmentation and connecting evidence across disciplines, **CAS BioFinder** empowers discovery teams to make better-informed decisions, reduce downstream risk, and advance high-potential candidates with confidence.

Connect with us to learn more ›

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

BIOWHP102836ENG260212

CAS
A division of the
American Chemical Society