Five data barriers to safer small molecule drug development and how to overcome them



CAS
A division of the
American Chemical Society

Table of Contents

3
Introduction

4
Fragmented discovery data, fragmented discovery workflow

8
One-dimensional tools hinder SAR visibility

10
Safety data gaps compromise early-stage go/no-go decisions

13
Limited biological context restricts clinical relevance

15
Disconnected tools slow interdisciplinary workflows

17
Make more informed drug discovery decisions earlier with CAS BioFinder®

18
References



Introduction

Advancing small molecules from idea to candidate hinges on speed, clarity, and early high-stakes decisions. Yet, as discovery workflows grow more complex and data volumes expand, it is harder for medicinal chemists to access the insights needed to accelerate R&D and advance candidates confidently.

The intensifying push toward **AI (artificial intelligence)** and **ML (machine learning)** frameworks adds another layer of complexity to drug discovery. However, these models are only as effective as the data behind them. Without high-quality inputs, their accuracy and value drop, which can skew early-stage decision-making and ripple through downstream development.

In this white paper, we map out **five data challenges slowing your R&D endeavors** to help you identify friction points, explore actionable solutions, and uncover how integrated, purpose-built discovery platforms can break down data barriers to support smarter design decisions in small molecule drug discovery.

5 DATA CHALLENGES

1

Fragmented discovery data, fragmented discovery workflow

Medicinal chemists must review, verify, and interpret a large volume of scientific data to inform development decisions and avoid redundant work. The challenge is that published data spans journal articles, patents, and databases, each with its own formatting and quality standards. This increases the risk of **missed insights, misinterpreted previous findings, and duplicated research efforts.**

The lack of structure across drug discovery-related sources forces researchers to waste valuable time browsing and cleaning data to identify relevant information, instead of advancing high-stakes R&D workflows.



Across industries, data scientists spend up to

80%

of their time cleaning data.[1] In drug discovery, that effort delays progression, treatments, and life-changing pipelines.

Three steps to turn fragmented data into fuel for drug discovery workflows

High-value insights are often buried under layers of disconnected information. Without structure, even the best data fails to support decision-making. What does it take to transform data noise into ready-to-use information?

Consolidate

Integrate disparate publication sources, databases, and patents into a unified, searchable platform to eliminate silos, streamline information retrieval, and accelerate access to relevant data.

Standardize

Harmonize formats, abbreviations, units, and terminologies across chemical, biological, and pharmacological data to ensure consistency and reduce time spent reprocessing information.

Connect

Link related entities (e.g., ligands, targets, diseases) using authoritative identifiers to unify context, reveal hidden relationships, and speed up insight generation across workflows.

CAS BioFinder Discovery Platform™: Your trusted source for discovery data

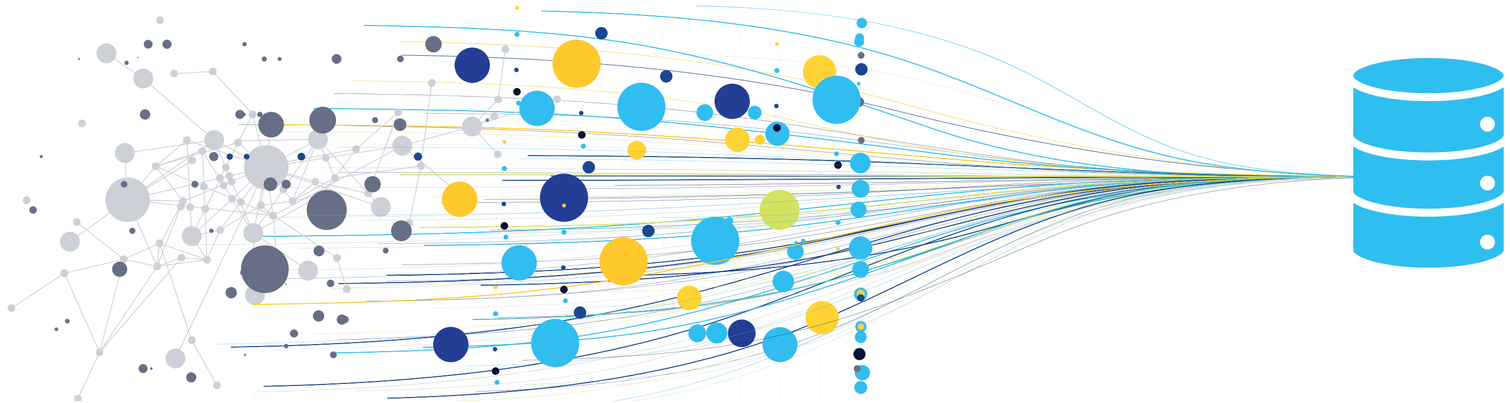
At CAS, we centralize and connect drug discovery-relevant information to help you generate accurate, actionable insights fast to accelerate decisions and workflows.

The CAS BioFinder Discovery Platform brings together expertly extracted, structured, and validated data so you gain easy access to curated information from:

- Thousands of top life sciences journals.
- Over 100 global patent offices.
- Trusted chemical, biological, and pharmacological databases—including the CAS Content Collection™ that powers CAS SciFinder®.

Scientific data landscape:
Countless sources of unstructured data

CAS BioFinder Discovery Platform:
A single, integrated source of drug discovery data





How do we make sure you don’t miss anything?

Every day, CAS scientists and data experts manually curate and verify scientific information, so you don’t have to second-guess its accuracy or reliability.

How do we make sure your insights are built on solid ground?

CAS experts assign consistent identifiers to ligands, proteins, and diseases and link related chemical, biological, and pharmacological data, ensuring clarity and alignment across sources.

Atorvastatin

CAS Registry Number: 134523-00-5

20 other names for this Ligand

- Atorvastatin acid
- (3R,5R)-7-[2-(4-Fluorophenyl)-5-isopropyl-3-phenyl-4-phenylcarbamoylpyrrol-1-yl]-3,5-dihydroxyheptanoic acid
- (βR,δR)-2-(4-Fluorophenyl)-β,δ-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino) carbonyl]-1H-pyrrole-1-heptanoic acid
- (βR,δR)-2-(p-Fluorophenyl)-β,δ-dihydroxy-5-isopropyl-3-phenyl-4-(phenylcarbamoyl) pyrrole-1-heptanoic acid
- 1: PN: WO2019129231 SEQID: 2 claimed sequence
- 1H-Pyrrole-1-heptanoic acid, 2-(4-fluorophenyl)-β,δ-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-, [R-(R*,R*)]-
- Alip
- Atavastatin
- Atofast
- Ator
- Atorcor
- Atorfit 20
- Atorin
- Atoris
- Atorvox
- Atrox
- Cardyl
- Lipilou
- Lipinon
- Statofilm

Canonical SMILES

O=C(O)CC(O)CC(O)CCN1C(C=2C=CC(F)=CC2)=C(C=3C=CC=CC3)C(C(=O)NC=4C=CC=CC4)=C1C(C)C

Isomeric SMILES

C(NC1=CC=CC=C1)(=O)C=2C(=C(N(CC[C@H](C[C@H](CC(O)=O)O)O)C2C(C)C)C3=CC=C(F)C=C3)C4=CC=CC=C4

InChI

InChI=1S/C33H35FN2O5/c1-21(2)31-30(33(41)35-25-11-7-4-8-12-25)29(22-9-5-3-6-10-22)32(23-13-15-24(34)16-14-23)36(31)18-17-26(37)19-27(38)20-28(39)40/h3-16,21,26-27,37-38H,17-20H2,1-2H3,(H,35,41)(H,39,40)/t26-,27-/m1/s1

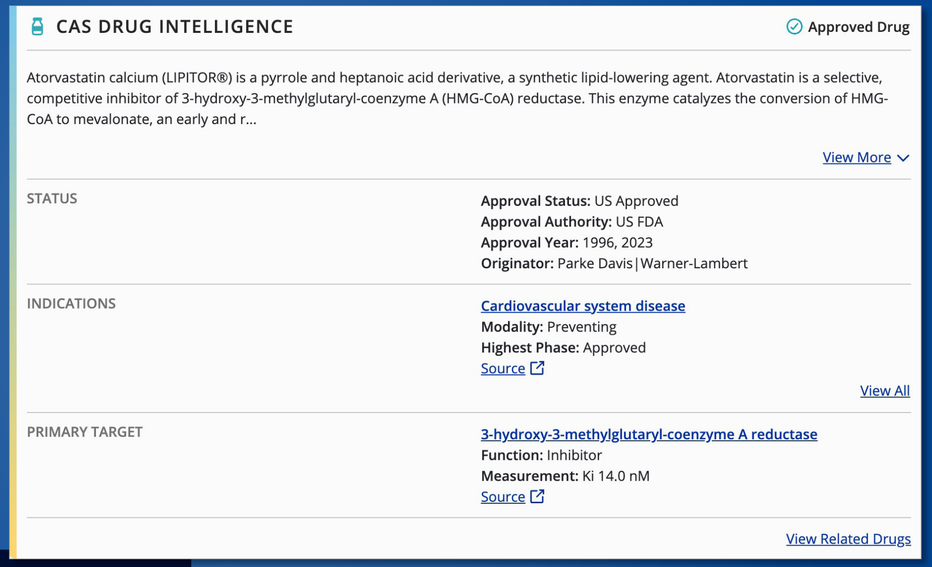
InChI Key

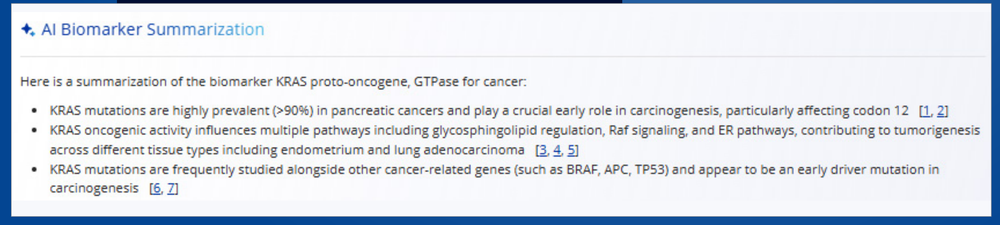
XUKUURHRXDUEBC-KAYWLYCHSA-N

How do we make sure you don't get overloaded with data?

Drug intelligence summaries in CAS BioFinder® filter and prioritize key insights, like clinical-stage compounds, approval status, and known metabolites, so you can quickly identify what's relevant and stay focused on high-impact decisions.

AI biomarker summarization helps you identify key trends and extract the most relevant insights, enabling faster data navigation and result interpretation.





Which integrated data supports your discovery workflow?

Workflow step	Integrated data
Target identification	– Bioactivity data – Pharmacology profiles – Disease relevance – Target class/family information – Literature and patent references
Hit evaluation	– Assay results (including IC50, EC50) – Structure-activity relationship (SAR) data – Functional group substitutions – Physicochemical properties (including logP) – Patent claims
Lead optimization	– Absorption, distribution, metabolism, and excretion (ADME) and toxicity data – Metabolite profiles – Selectivity/off-target data – Matched molecular pairs
Candidate prioritization	– Clinical relevance and druglikeness metrics – Approval status – Patent landscape and freedom-to-operate information

Each stage of small molecule discovery depends on timely access to the right data. The CAS BioFinder Discovery Platform combines curated chemical, biological, pharmacological, and clinical information in one user-friendly interface so you can easily access the information you need to support confident decision-making across your workflow.

2

One-dimensional tools hinder SAR visibility

Medicinal chemists rely on SAR information to understand how molecular changes influence potency, selectivity, and safety. Without easy access to high-quality SAR data, researchers risk overlooking critical relationships and prioritizing compounds that will fail later down the pipeline. Yet, that information alone is not enough. To make SAR data actionable, medicinal chemists need functional tools to explore chemical space with clarity and context.

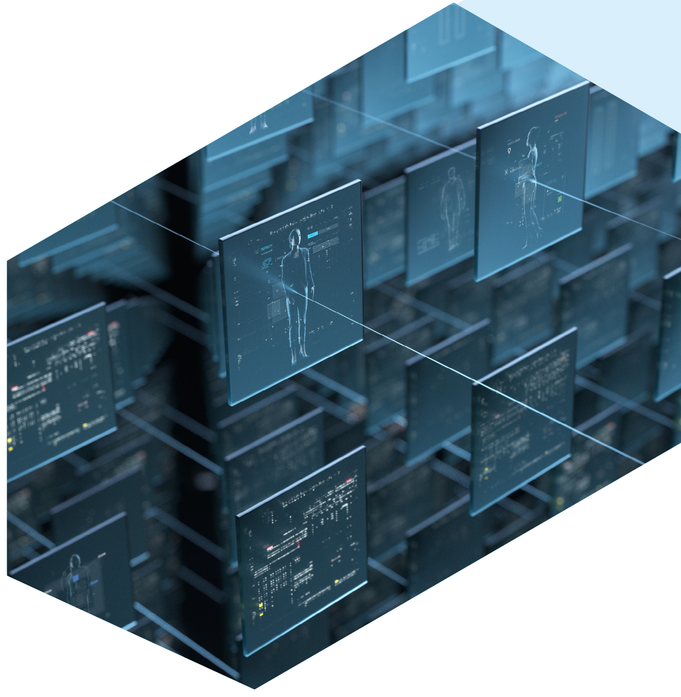
What's needed for actionable SAR analysis

Structural visualization
Reveal how specific molecular changes affect bioactivity and to identify trends in potency, selectivity, and safety earlier in the discovery process.

Scaffold-based exploration
Map functional group substitutions, support side-by-side comparisons of analogs, and help identify promising modifications.

Context-rich insight
Substitution history and experimental results to clarify what's been tried, what worked, and what didn't for informed discovery decisions.

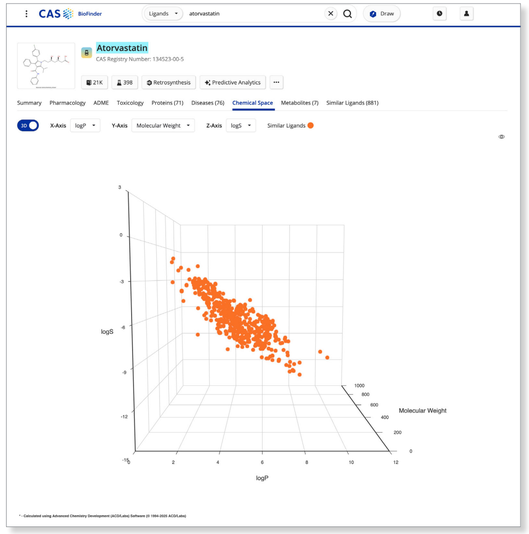
That's where small molecule discovery often falls short. Many common tools still present SAR data in fragmented views and lack assay context, requiring users to piece together substitution outcomes manually. This recurrent gap not only **limits opportunities** to optimize leads but also **increases the risk of wasting time and resources** revisiting analogs with known or poor performance.



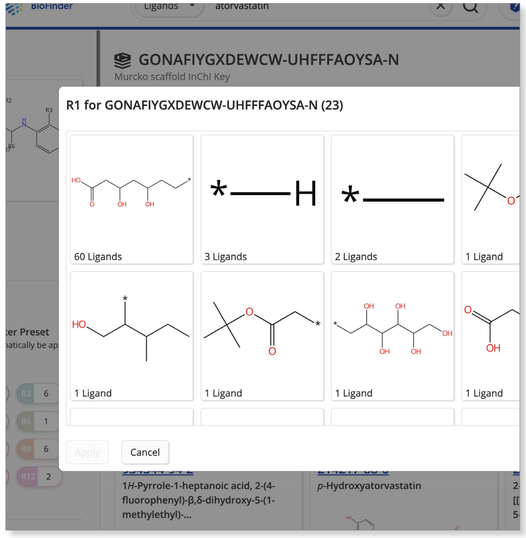
How CAS BioFinder can help: Four functions to bring SAR clarity

Designed with discovery workflows in mind, the functionality of CAS BioFinder gives you a deeper understanding of how structures influence bioactivity.

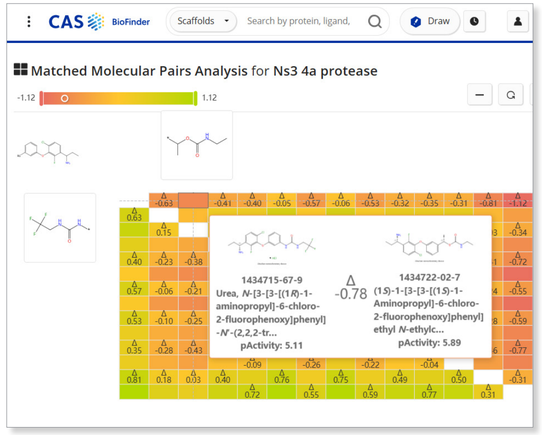
Chemical space visualizations
Visualize relationships in 2D or 3D between molecular properties and bioactivity outcomes, such as potency and selectivity, across compound series.



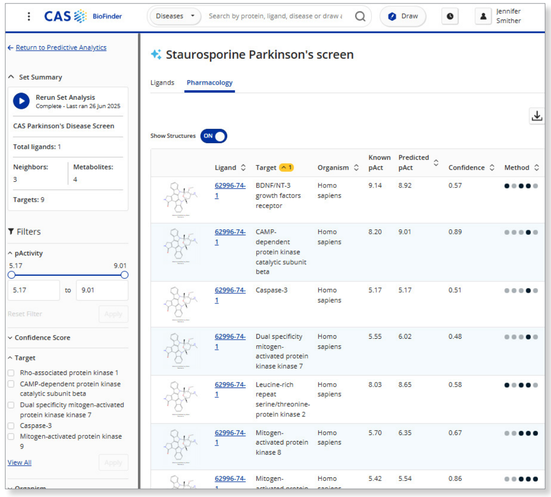
Comprehensive scaffold search
Build upon earlier research by scaffold hopping with defined functional groups at each position to better estimate safety and efficacy properties.



Interactive matched molecular pair analysis
Explore substitution effects and uncover activity-driving structural changes.



Predicted bioactivity data for novel and existing compounds
Go beyond published data. Leverage predicted activity insights to anticipate interactions, support confident lead optimization, and uncover drug repurposing opportunities.



Through CAS BioFinder drug discovery-specific features, we help you **uncover structure-activity patterns faster, prioritize high-potential modifications, and accelerate lead optimization.**

3

Safety data gaps compromise early-stage go/no-go decisions

While early discovery decisions have long-term consequences, medicinal chemists continue to lack access to early-stage safety data and predictive tools needed to evaluate and advance compounds confidently. Without visibility into off-target effects, metabolic liabilities, or toxicity risks, assessing safety is compromised until much later in development.

This lack of foresight **slows workflows and increases the risk of advancing leads with hidden safety issues,** which may not be flagged until *in vivo* studies or clinical trials. This is especially risky when timelines are tight, costs are high, and course correction is limited.

How to close the safety gap in early discovery

To avoid late-stage setbacks, medicinal chemists must bring safety profiling into the design phase, when there is still time to make informed adjustments. This requires:

ADME and toxicity data
Incorporating and harnessing ADME and toxicity data earlier in the workflow.

Metabolic transformations
Exploring metabolic transformations in depth to identify reactive intermediates.

Off-target risks
Leveraging predictive models to evaluate off-target risks and potential liabilities beyond existing literature.

Yet, fragmented safety data and generic prediction models limit the ability to evaluate risk with the clarity that early-stage design demands.

One in four trials fail in Phase II due to safety concerns[2]

Late-stage safety surprises are costly. Without early insight into off-target effects or metabolic risks, high-risk compounds can slip through—only to fail later with disastrous consequences.

How CAS BioFinder can help: Build confidence into early safety decisions

CAS BioFinder combines small molecule drug discovery-relevant data and specialized models in a single discovery environment to help you identify and eliminate high-risk compounds earlier in the design process.

ADME and toxicity data

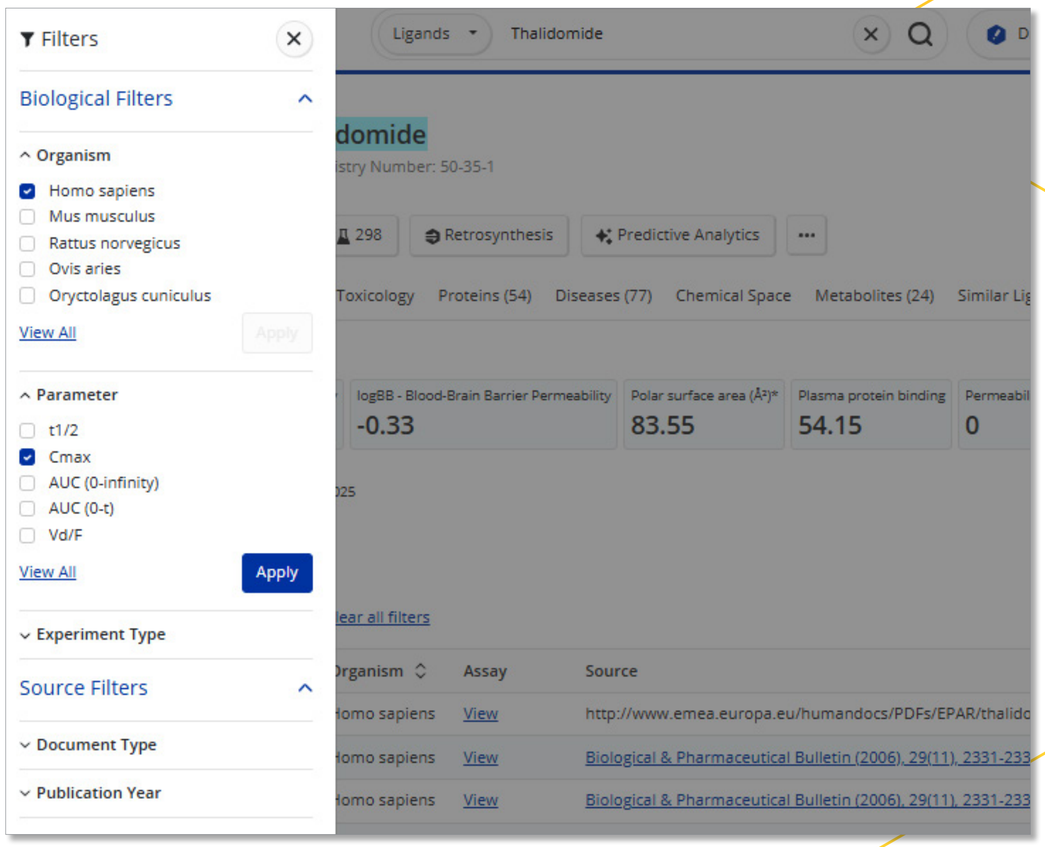
Integrate ADME and toxicity data from literature and curated sources into your workflow to evaluate risk early and guide safer compound prioritization.

Metabolite insights

Explore known and predicted metabolites to anticipate reactive functionality or metabolic hot-spots and reduce downstream safety risks.

Predictive models

Predict off-target interactions using advanced predictive models and machine learning to highlight safety liabilities before they become costly setbacks.





Predictive analytics: Quality in, confidence out

Predictive tools are only as reliable as the data behind them. Without high-quality, reliable data, AI and ML risk generating misleading results that skew decisions and delay progress.

Why you can trust CAS BioFinder predictions?

Grounded in curated, high-quality data

Predictive models in CAS BioFinder are trained on the CAS Content Collection™ — one of the most comprehensive, manually curated scientific databases in the world — ensuring predictions are based on accurate, well-structured data from peer-reviewed research and the latest patents.

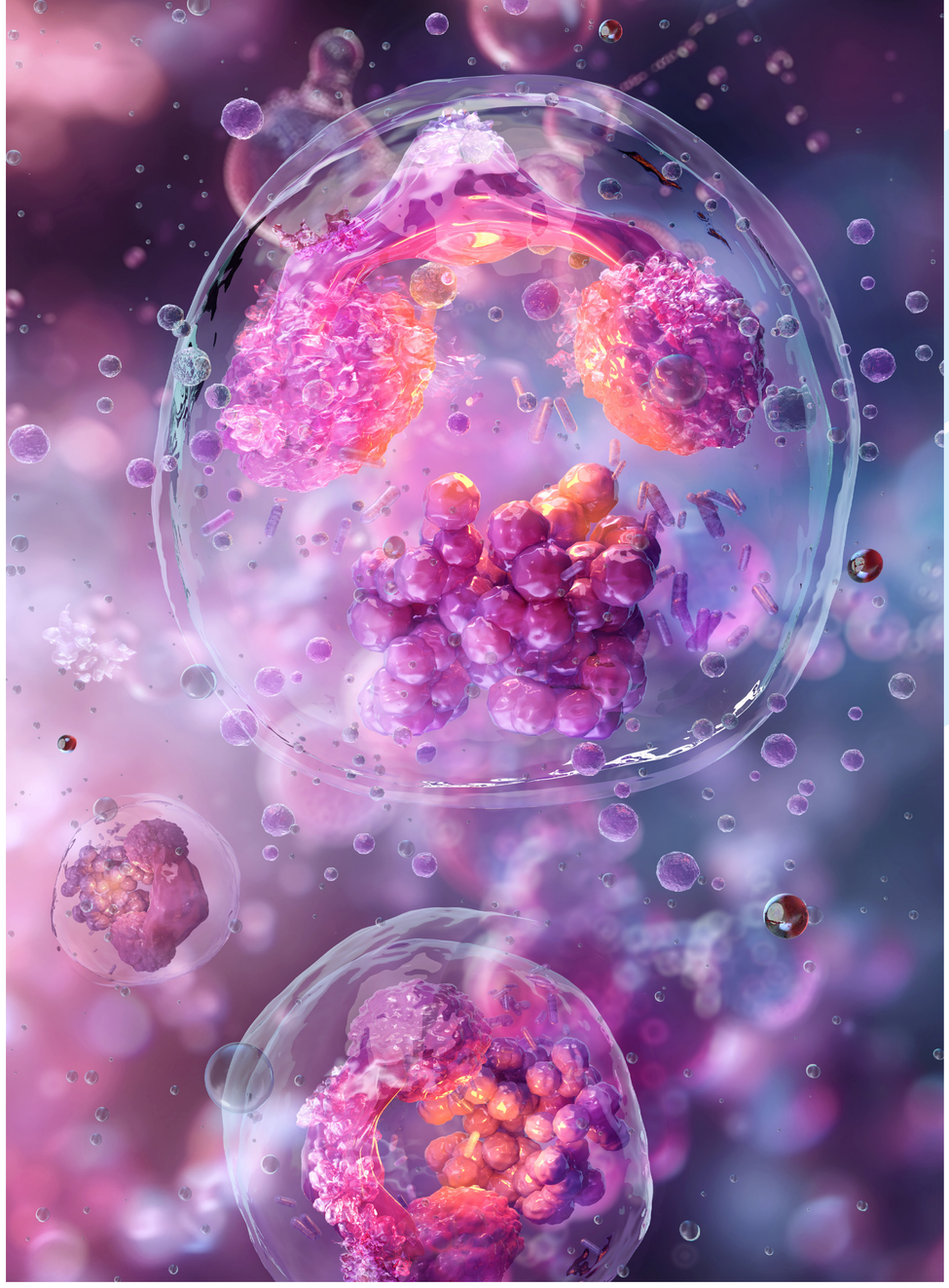
Visibility into source and context

Unlike many AI tools that produce results without context, CAS BioFinder allows you to trace predictions back to supporting experimental data, publications, or related compounds, so you gain full visibility into the rationale behind each prediction.

Smart combination of experimental and predicted data

CAS BioFinder combines published experimental bioactivity and metabolite data with predictive insights into off-target effects enabling you to uncover connections and make more informed, confident decisions faster.

By making safety data accessible and grounding predictions in curated, trustworthy sources, CAS BioFinder empowers you to evaluate risk earlier so you can focus on advancing high-potential candidates.



4

Limited biological context restricts clinical relevance

When medicinal chemists lack access to tools that connect chemical structures to disease biology, biological pathways, or therapeutic indications, they are forced to make critical design decisions without a clear view of clinical relevance. This disconnect in biological data can **misdirect lead optimization** and prioritization, increasing the risk of advancing compounds that are out of sync with clinical needs.

Bridging the gap between the lab and real-world relevance

To align compound design with therapeutic relevance, medicinal chemists need tools that bring clinical and biological context into the discovery phase. This involves:

Deepening visibility
Deepening visibility into biological pathways and disease associations.

Connecting ligands
Connecting ligands and targets to real-world indications and clinical relevance.

When biological and pharmacological data are connected, design decisions become more strategic and better aligned with downstream priorities.



How CAS BioFinder can help:
Design with clinical context in mind

Understanding how a compound fits into the broader therapeutic landscape is essential to conduct informed small drug discovery. CAS BioFinder links small molecules and targets to disease biology, pathways, and indications, so you don't have to waste time manually piecing information together.

With CAS BioFinder, you can:

- **Review integrated biological contexts,** including detailed breakdowns of diseases, pathways, targets, and ligands.
- **Understand therapeutic pathways** and explore the downstream effects of molecular interventions.
- **Identify relevant compounds** already tested for similar or adjacent disease areas.
- **Leverage built-in drug intelligence** to explore clinical-stage molecules and review approval status, indications, and associated metabolite data.

These capabilities help you design confidently with therapeutic relevance in mind for smarter prioritization and better alignment with translational goals.

5

Disconnected tools slow interdisciplinary workflows

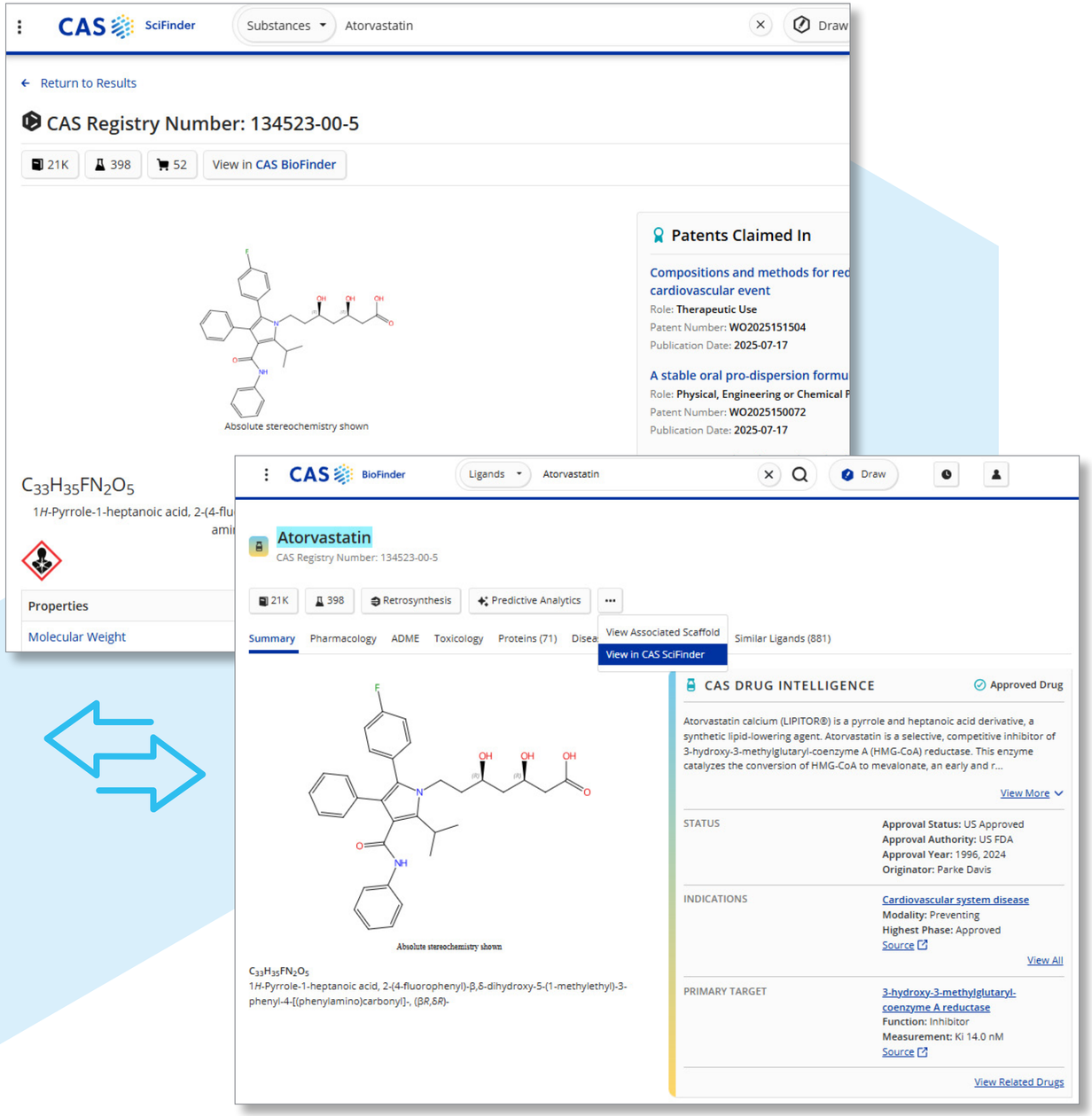
Medicinal chemists often work across a patchwork of disconnected systems like design tools, visualization software, and various databases. These siloed platforms make it difficult to maintain continuity across workflows and hinder collaboration with biologists and toxicologists.

The result? Inefficiencies at every step.

- **Time lost** switching between tools, visualization platforms, and disconnected databases.
- **Limited ability** to share data rapidly and securely across teams, causing misalignment and duplicated efforts.
- **Reduced visibility** into workflows and progress across disciplines.

These shortcomings make it harder for teams to align with the most promising candidates, slowing interdisciplinary workflows and reducing confidence in shared decisions. By bringing molecule design, analysis, and safety data into a unified, shared environment, researchers can work more efficiently and stay aligned even as priorities evolve.





How CAS BioFinder can help: Connect data, disciplines, and discovery priorities

To reduce inefficiencies and streamline workflows, you need drug discovery-relevant context and tools that support continuity in cross-disciplinary projects. CAS BioFinder helps you seamlessly connect chemical and biological workflows so you can make faster, more confident decisions.

What you can do	Why it matters
Unify chemical structures, predicted and eperimental data, and literature references.	Centralizes key discovery data in a single platform to reduce context-switching and rework.
Move seamlessly from CAS BioFinder to CAS SciFinder and back without disconnecting your workflow.	Eliminates the need to switch tools to connect structure-based insights with biological context.
Promote inter-disciplinary work by giving teams a shared, secure platform.	Streamline alignment and collaboration across disciplines with a single source of truth.

Make more informed drug discovery decisions earlier with CAS BioFinder

Small molecule discovery demands speed, confidence, and informed decisions grounded in reliable data. However, persistent information and analysis challenges prevent medicinal chemists from finding critical insights needed to support confident decisions, streamline R&D, and accelerate life-changing therapeutic pipelines.

Solving critical data challenges means going beyond basic information retrieval. It requires a drug discovery platform that not only unifies chemical, biological, and pharmacological information but also integrates the smart tools you need to uncover what's possible, not just what's published.

That's why we designed CAS BioFinder: To give you the necessary tools to make informed decisions earlier in your workflow so you can expedite safer, better, and more effective therapeutics to patients in need.

To learn more, **click here** or visit **link.cas.org/biofinder.**





References

1. CrowdFlower. Data Science Report 2016. CrowdFlower; 2016. https://web.archive.org/web/20170403203026/https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf
2. Harrison RK. Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov. 2016;15(12):817-818. doi:10.1038/nrd.2016.184

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

© 2025 American Chemical Society. All rights reserved.

BIOWHP102624ENG250902

CAS
A division of the
American Chemical Society