causaly

WHITEPAPER

Scientific Information Retrieval System: The Foundation of Causaly's Agentic AI Platform

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

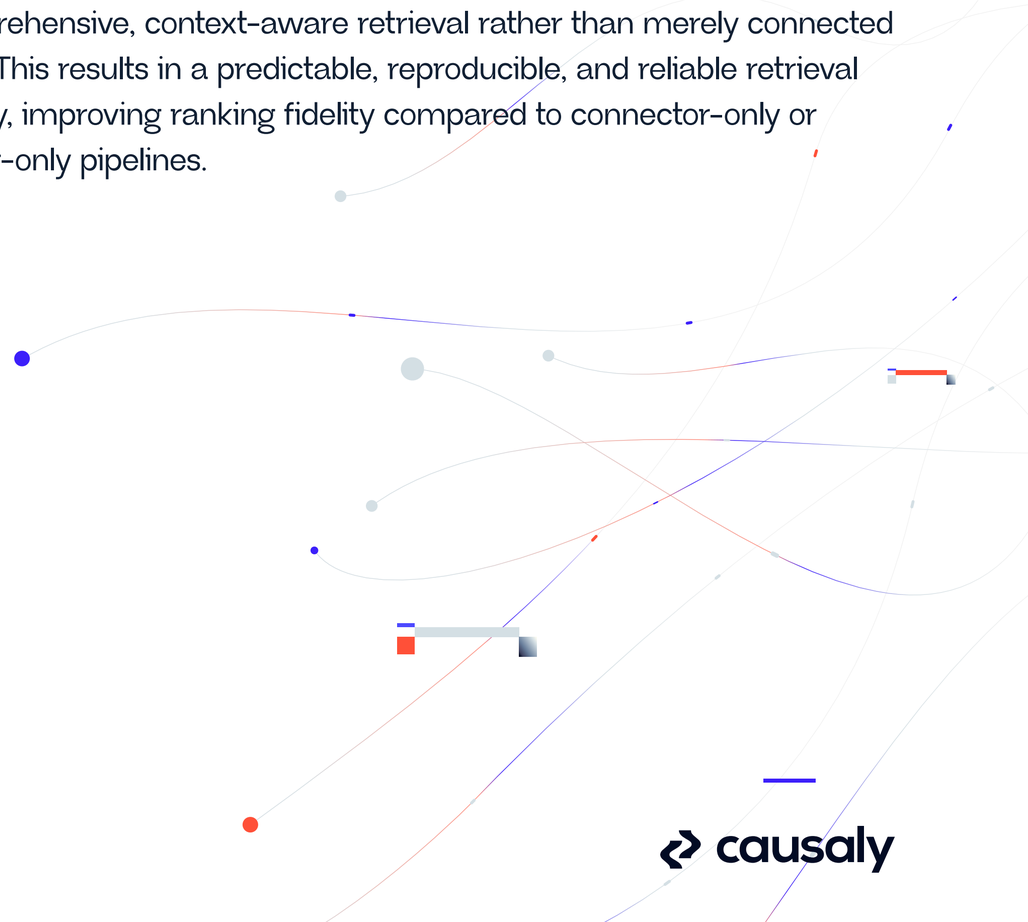
Biomedical R&D depends on finding the right evidence and insight across vast, noisy, heterogeneous, and rapidly evolving scientific corpora.

Conventional approaches such as lexical search, vector-only retrieval, or “connector-first” architectures (e.g., MCP-style integrations of disparate sources) often return a top-k set of results that are usually inconsistent, poorly ranked, and lacking in domain-specific semantics. These limitations become evident in long-tail scientific questions, where similarity search alone cannot reliably uncover mechanistic plausibility, directional relationships, or refuting evidence.

Causaly’s Scientific Information Retrieval System (SIRS) addresses this gap by unifying a production-grade data fabric (document index + vector store) with ontology-driven query understanding and expansion that leverages public and custom ontologies, expert vocabularies, and curated rulesets. This foundation is further enhanced through knowledge–graph–based evidence grounding and ranking, enabling more precise and context-aware retrieval.

By achieving a balanced fusion of these capabilities, SIRS surfaces evidence that is domain-aware, mechanistically connected, directionally consistent, and transparently sourced, delivering clear, cited answers suitable for making decisions.

When deployed alongside an enterprise data fabric, SIRS continuously ingests public scientific data, literature, and internal data, ensuring comprehensive, context-aware retrieval rather than merely connected data. This results in a predictable, reproducible, and reliable retrieval quality, improving ranking fidelity compared to connector-only or vector-only pipelines.



causaly

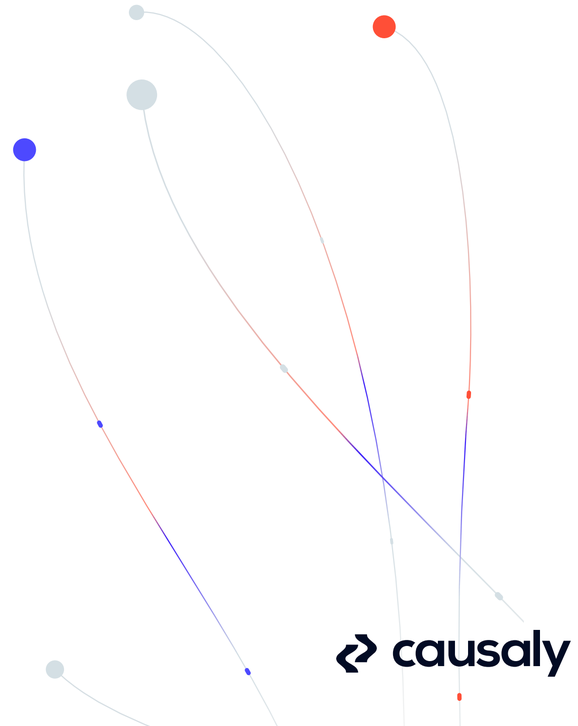
waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Introduction to Causaly's Scientific Information Retrieval System (SIRS)

The Scientific Information Retrieval System (SIRS) is Causaly's scientific search framework for life sciences.

It is designed to transform how scientific information is extracted with greater accuracy, relevance, transparency, and contextual understanding. By combining curated rules, ontologies, machine learning, and knowledge graphs into a unified search and summarization pipeline, SIRS redefines how scientists discover and interpret biomedical knowledge. In addition to supporting human researchers, SIRS also supplies agentic systems and LLMs with highly relevant, grounded data, reducing the risk of hallucinations and incorrect answers.

SIRS achieves higher levels of precision and semantic understanding by aligning retrieval with scientific intent and grounding evidence within a unified Data Fabric and Biology Knowledge Graph (Causaly BioGraph). This enables researchers to obtain not only relevant information but also verifiable, mechanistically grounded insights that support rigorous scientific reasoning.



causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

At its core, SIRS is designed to excel in three key areas beyond the capabilities of general-purpose retrieval systems:

01

Understand scientific queries with precision

SIRS interprets the user's question in a scientifically grounded way by:

- Expanding queries using ontologies to include synonyms, related terms, and hierarchical relationships.
- Disambiguating entities, biological mechanisms, and relationship directionality (for example, distinguishing *inhibits* from *activates*).
- Interpreting complex, niche, or "long-tail" biomedical questions through intent modelling trained on scientific language.

02

Retrieve and rank evidence transparently

To ensure the retrieval process is both accurate and explainable, SIRS:

- Combines dense (vector-based) and lexical, or sparse (keyword-based) and knowledge graph retrieval methods, guided by domain ontologies to maintain scientific relevance.
- Applies graph-aware re-ranking to prioritize evidence that reflect mechanistic plausibility and correct causal direction.
- Fuse all retrieved entities and relationships in the Scientific Data Fabric and Knowledge Graph before ranking.

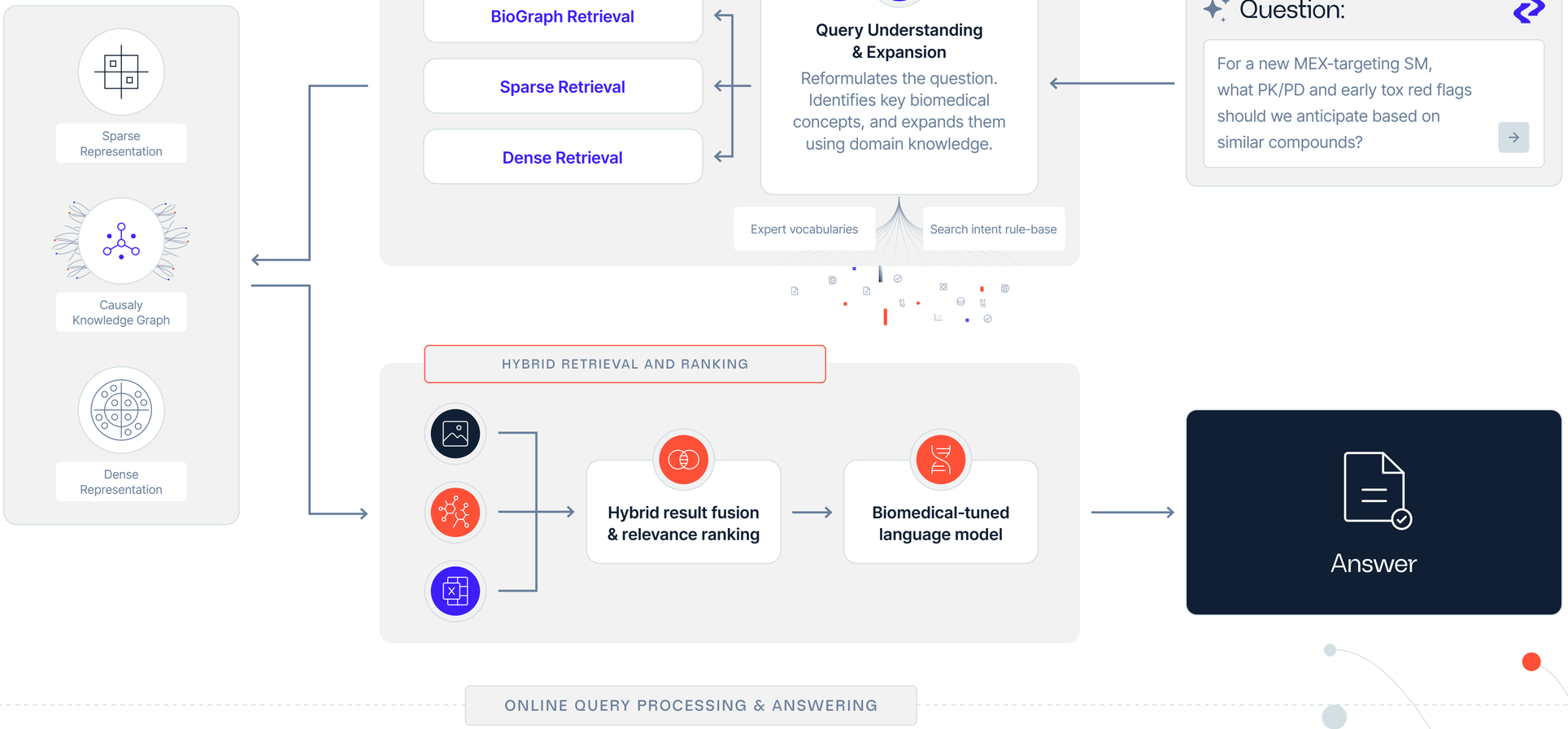
03

Deliver answers grounded in verifiable data

SIRS provides scientifically trustworthy outputs by:

- Presenting relationship evidence, where applicable, as evidence cards that include provenance, confidence levels, and any conflicting findings.
- Provides direct links to the original source text and metadata for full transparency.
- Ensures retrieval is reproducible and consistent, supporting high-stakes decision-making.

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem



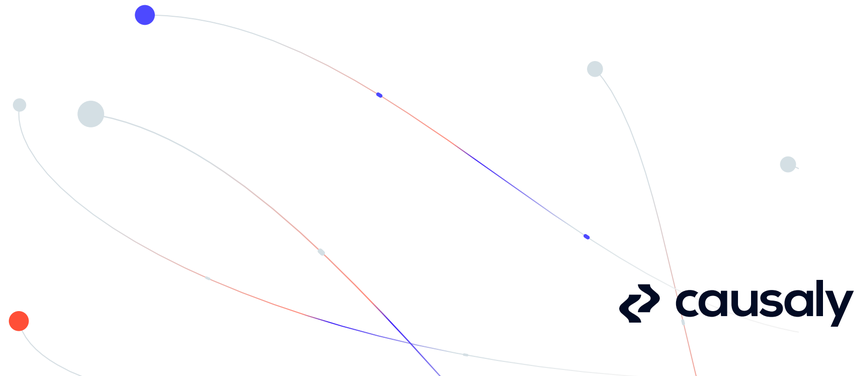
causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Distinctive Capabilities of Causaly's SIRS

Several core capabilities differentiate Causaly's SIRS from general-purpose retrieval systems, each contributing to its scientific relevance, accuracy, reproducibility, and transparency. The below provides an overview;

- **Life-sciences, domain-specific engine:** Causaly is built specifically for biomedical research and drug development. The system encodes years of domain work in **custom ontologies**, two large-scale **knowledge graphs** (biology and therapeutic landscapes), expert vocabularies, and **tens of thousands of hand-crafted rules** for information extraction, retrieval, and ranking. Together with fine-tuned AI models for search translation and intent understanding, Causaly SIRS offers the most advanced and complete scientific retrieval and ranking system for biomedical research.
- **Scientific Data Fabric as the foundation:** Underneath SIRS is a scientific data fabric that organizes heterogeneous scientific content — literature, structured scientific databases, clinical trials, IP data, and internal, multi-modal proprietary data — into a coherent evidence layer. With a domain-aware document index, vector store, and biomedical chunking strategy combined with graph representations to ensure that retrieval and summarization operate over **structured, normalized scientific concepts,** not only raw text.

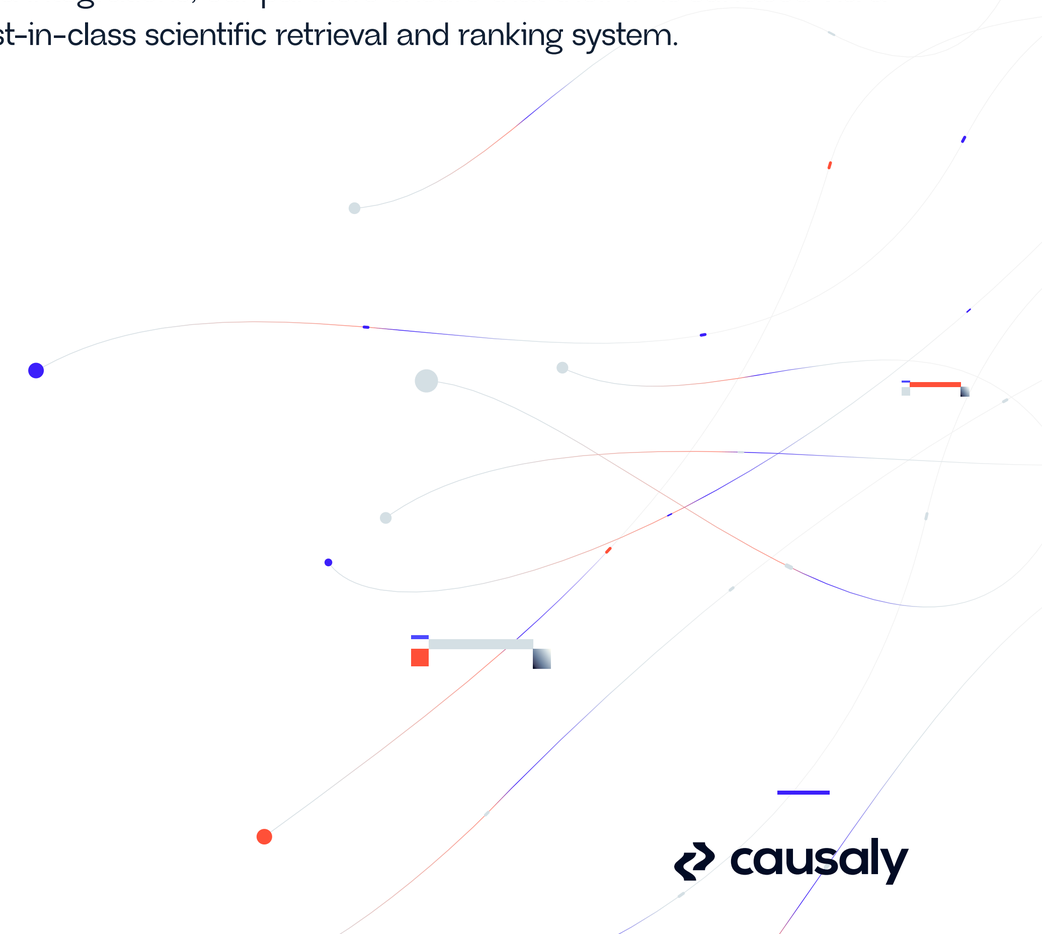


causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Distinctive Capabilities of Causaly's SIRS

- **Multi-stage Scientific Information Retrieval System (SIRS):** SIRS is a staged architecture, not a single RAG call. It detects search intent, performs query understanding, ontology mapping and expansion, applies rule-based routing for optimal retrieval, and combines semantic, lexical, and relationship search with GraphRAG and graph-aware relevance ranking. On top, a knowledge-synthesis layer performs per-document insight extraction, grounding, and synthesis.
- **Accuracy, grounding, and traceability by design:** Answers are generated only from retrieved, **source-grounded insights**. Each statement is tied back to the primary text, classified as a *direct answer*, *related evidence*, or *not relevant*, and regularly evaluated against internal benchmarks by Causaly scientists. This architecture delivers the **factuality, provenance, and reproducibility** required for scientific trust and adoption.
- **Platform integration:** SIRS is a scalable solution that powers the broader Causaly platform. Through hybrid search and graph private data integrations, our partners ensure that their IP is surfaced via a best-in-class scientific retrieval and ranking system.



causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

How Causaly's SIRS works

SIRS structures every query into a five-step pipeline that ensures precision, transparency, and scientific rigor by:

Query Understanding and Expansion

Query understanding combines rule-based matching, ontology-driven interpretation and expansion, and fine-tuned biomedical large language model reformulation to precisely capture the user's scientific intent. This process consists of four steps that work together to interpret, enrich, and operationalize the question, and they are:

01

Query reformulation

The system refines the user's question into a clear, domain-aligned expression that preserves scientific intent and improves downstream interpretability.

02

Keyword extraction

Key biomedical entities and concepts are identified to anchor the search space and ensure the retrieval process focuses on the most relevant signals.

03

Ontology expansion

The query is enriched using curated biomedical ontologies, capturing synonyms and related concepts to enable comprehensive yet precise evidence of discovery.

04

Intent recognition

Encoded domain expertise is applied to infer the underlying intent of the question, guiding the system toward the most suitable retrieval pathway.

causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

These steps produce a structured, domain-aware instruction set that informs how the system searches, prioritizes, and assembles evidence. The enriched query then enters the hybrid retrieval stage.

User query:

What are the key validated targets currently implicated in CKD progression?

Quality Assurance

All AI-generated summaries include direct links to the original sources for verification. In addition, our team of PhD scientists and machine learning experts performs regular reviews to uphold accuracy.

How Our AI generates Search Summaries

Our AI Overviews provide concise summaries of search results by following these steps:

01 **Query reformulation:**
"validated targets implicated in chronic kidney disease progression"

02 **Extracted Keywords:**
"molecular targets" "chronic kidney disease" "disease progression"

03 **Ontology expansion:**
Concept-id: ST1513403 (internal ontology)
Concept-semantic_type: Chemical Viewed Structurally
Concept-name: Molecular Target
Concept-expansion: Biological target, Therapeutic target

Concept-id: C1561643 (UMLS id)
Concept-semantic_type: Disease or Syndrome
Concept-name: Chronic Kidney Diseases
Concept-expansion: Chronic renal disease, CKD, Loss of renal function

Concept-id: C0242656 (UMLS 1d)
Concept-semantic_type: Pathologic Function
Concept-name: Disease Progression
Concept-expansion: Disease advancement, Pathologic progression (custom synonyms)

04 **Intent recognition:**
ruleId: 45
ruleIntent: Question that links a specific disorder to its relevant molecular targets.

causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

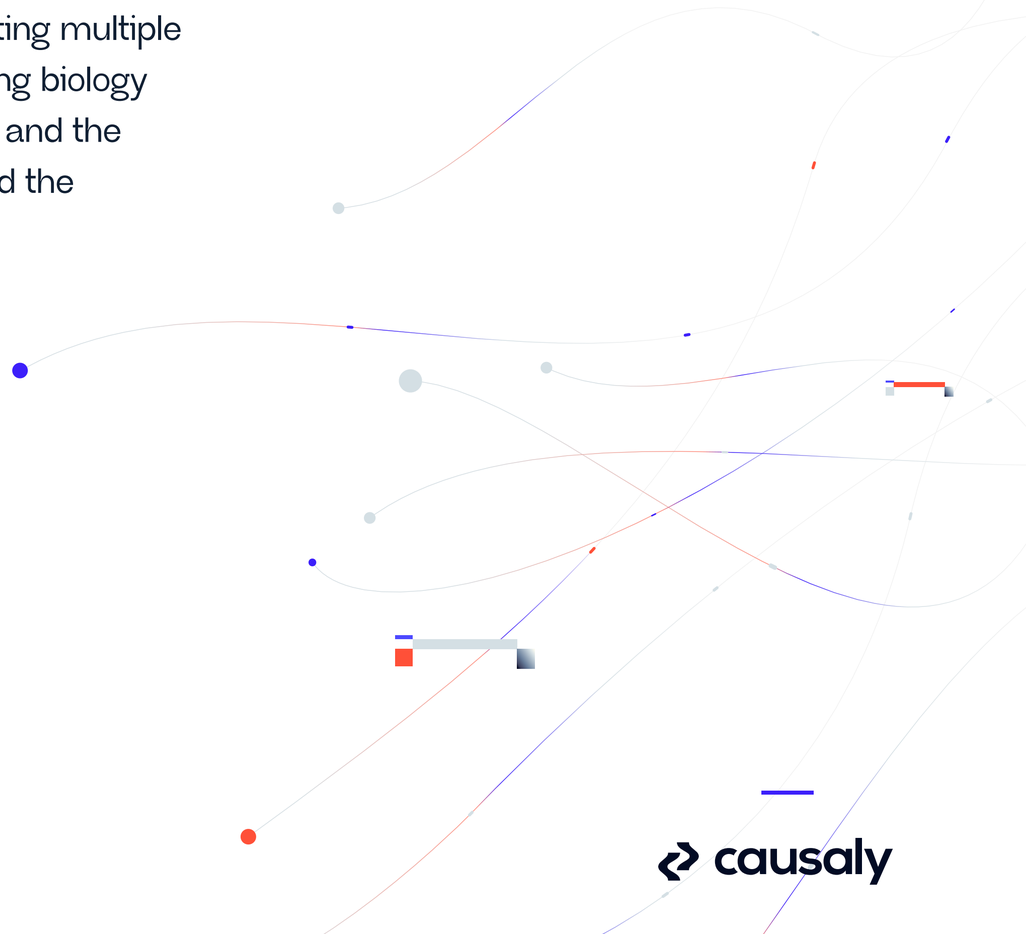
Hybrid and Multimodal Document Retrieval

Hybrid retrieval ensures that no relevant evidence is overlooked by integrating multiple retrieval modes that capture both the scientific language and the underlying biology of a query. This stage acts as the bridge between the interpreted question and the scientific knowledge represented across text, tables, figures, metadata, and the BioGraph. It comprises of three components:

Semantic (dense) retrieval

To capture meaning beyond literal phrasing, grasp a broader context and scientific nuance, SIRS:

- Retrieves relevant content from the dense embedding space for both the user question and source data.
- Enables the system to surface evidence even when terminology varies across studies or data sources.



causaly

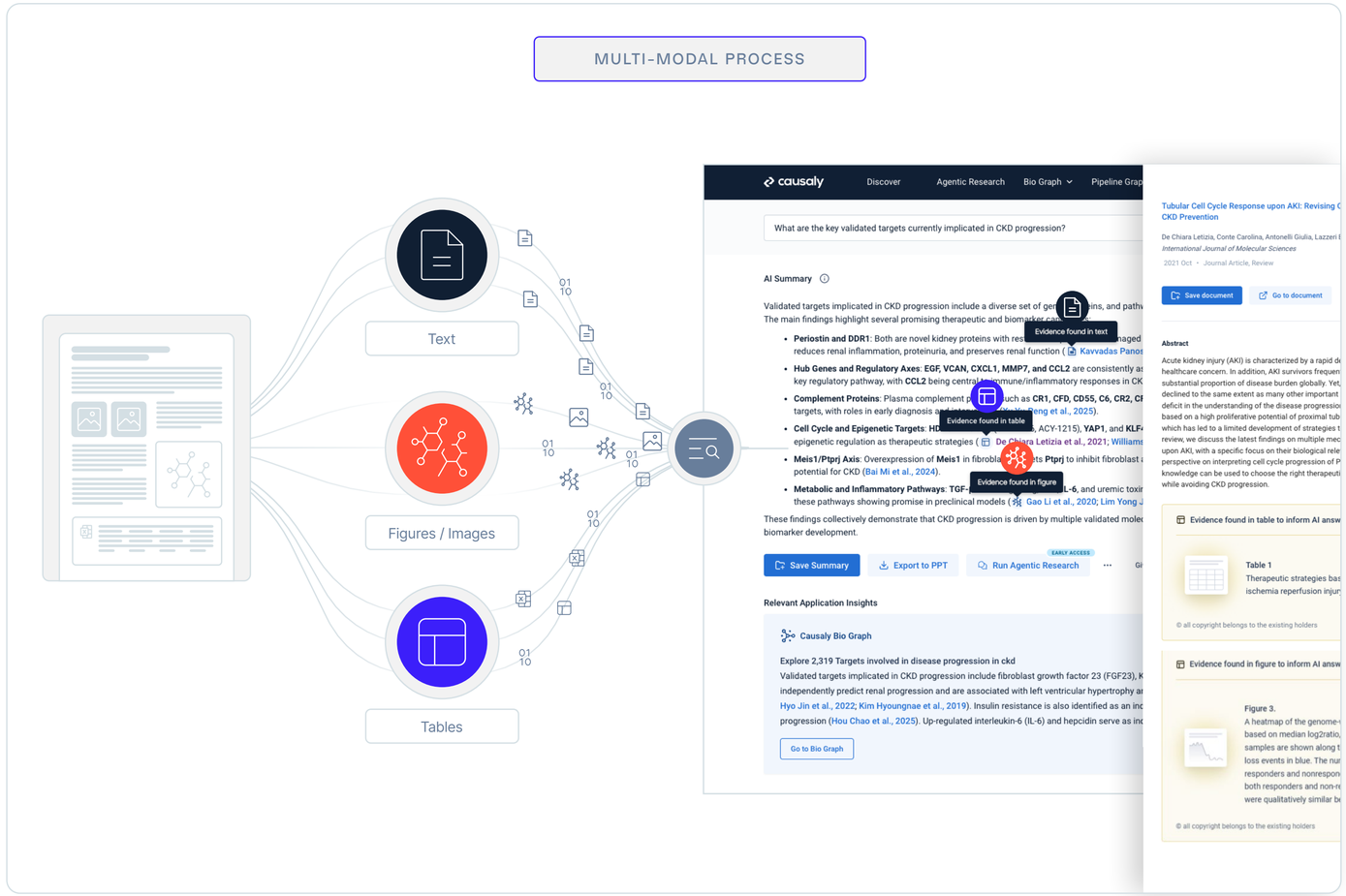
waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Hybrid and Multimodal Document Retrieval

Multimodal retrieval

Biomedical evidence is not limited to text. Key findings are often encoded in tables, figures, charts, microscopy images, and structured datasets. To ensure these diverse evidence formats are fully captured and searchable, Causaly's multimodal retrieval pipeline:

- Captures these evidence types during the parsing and indexing stage
- Ensures research insights are not lost simply because they are expressed visually or numerically rather than in text



causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Lexical (sparse) retrieval

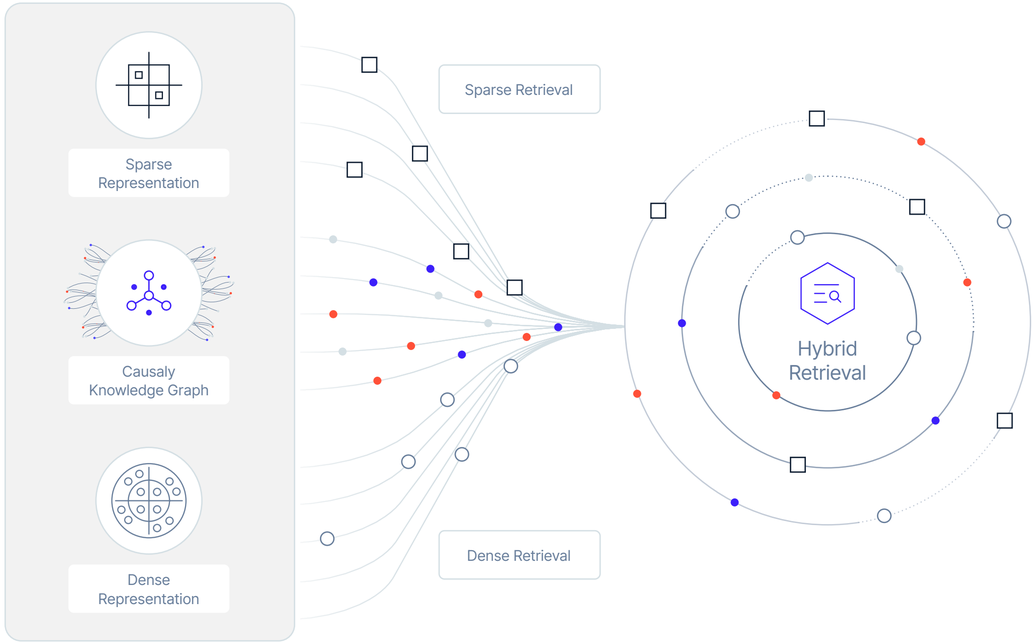
To maintain precision, factual grounding and expand the conceptual scope with curated ontologies, SIRS:

- Performs precise text, tables, figures, and metadata matching to identify exact terms, identifiers, and study-specific metadata.
- Supports high-specificity retrieval and domain-controlled filtering where accuracy and factual grounding are essential.

BioGraph retrieval

To uncover mechanistic explanations, SIRS:

- Queries curated biomedical causal graph to identify mechanistic relationships between entities.
- Provides structured insight and relevance boosting using concept-concept relationship-based evidence, aware of relationship directionality.



Together, these retrieval modes deliver domain-faithful, explainable, and comprehensive evidence sets that reflect both the language and the biology embedded in the user's question.

causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Retrieval Fusion and Ranking

After retrieval, evidence is aggregated across dense, sparse, and BioGraph search. The ranking stage evaluates these multi-signal candidates and fuses them into a unified, scientifically grounded, and ranked list of evidence. A key part of this process is how ranked evidence is organized and evaluated as outlined below:

Evidence and Document-level prioritization

Rather than ranking passages in isolation, the system evaluates evidence at the **results (e.g., chunk, excerpt, relationship, table) and document level** (e.g., PMC article, PDF, preprint), ensuring that each ranked item represents a coherent and contextually valid body of work. This reduces noise and elevates documents containing the most relevant evidence across all retrieval modes.across all retrieval modes.

Adaptive scoring for scientific intent

Ranking logic applies a range of boosting factors and prioritizes factual accuracy and contextual relevance by adjusting weights according to our ranking framework and the context of the query. This produces a focused, evidence-rich set of top-ranked documents that surface the most scientifically relevant material.

Examples of evidence scoring:

EVIDENCE	SCORE
In conclusion, this study shows that PD-L1 expression is not a prognostic factor in early-stage, surgically resected NSCLC.	-5
In addition to being related to the efficacy of ICI drug therapy, PD-L1 expression may also be related to the prognosis of NSCLC patients, according to recent reports.	0.5
We next showed that PD-L1 regulated NSCLC cell proliferation via Gas6 and the MerTK signaling pathway in vitro and in vivo.	5
Another study showed that PD-1, CTLA-4, and PD-L1 promoter hypomethylation has a significant impact on the course of the NSCLC progression.	11.25

In the illustrative table above, you can see that the evidence claiming no factor has a negative score. Meanwhile, evidence showing “significant impact” has a positive score.

Ranking is important because it determines which documents and evidence users see first, and it shapes the AI summary or report that Causaly provides.

causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Knowledge Synthesis Layer

On top of SIRS, there's a dedicated knowledge synthesis layer. This layer applies additional LLM-based relevance filter on a per-document basis.

It extracts insights from each document, classifying them as direct answer, related evidence, or not relevant, and then builds the final response from per-document, source-grounded insights rather than a bulk of raw chunks. This enhanced context engineering and grounding approach further improves SIRS's retrieval precision and relevance, delivering highly accurate, traceable answers with a stronger scientific argument structure. The knowledge synthesis layer achieves this through four processes:

01

Per-document insight extraction

For each document in the ranked top-k, SIRS extracts question-specific insights. The model further adds an LLM-based relevance filter step by distinguishing between:

- Direct answers to the query.
- Related but indirect findings that could support scientific reasoning.
- No answer cases where the document is only tangentially relevant.

02

Strict evidence grounding

Each insight generated on a per-document level is grounded in the source data. Supported claims are kept along with the cited excerpts, while unsupported statements are dropped. This grounding step enables UI highlighting and gives scientists an immediate way to inspect and verify every claim.

03

Structured, multi-document synthesis

The final overview is produced from these grounded, per-document insights — not from the full set of raw chunks. The knowledge synthesis layer performs enhanced context engineering across documents by clustering per-document insight-citation data and filtering out the text that is not related to the user's question. The result is an overall summary that is accurate, with easily traceable sources and a better scientific argument structure.

causaly

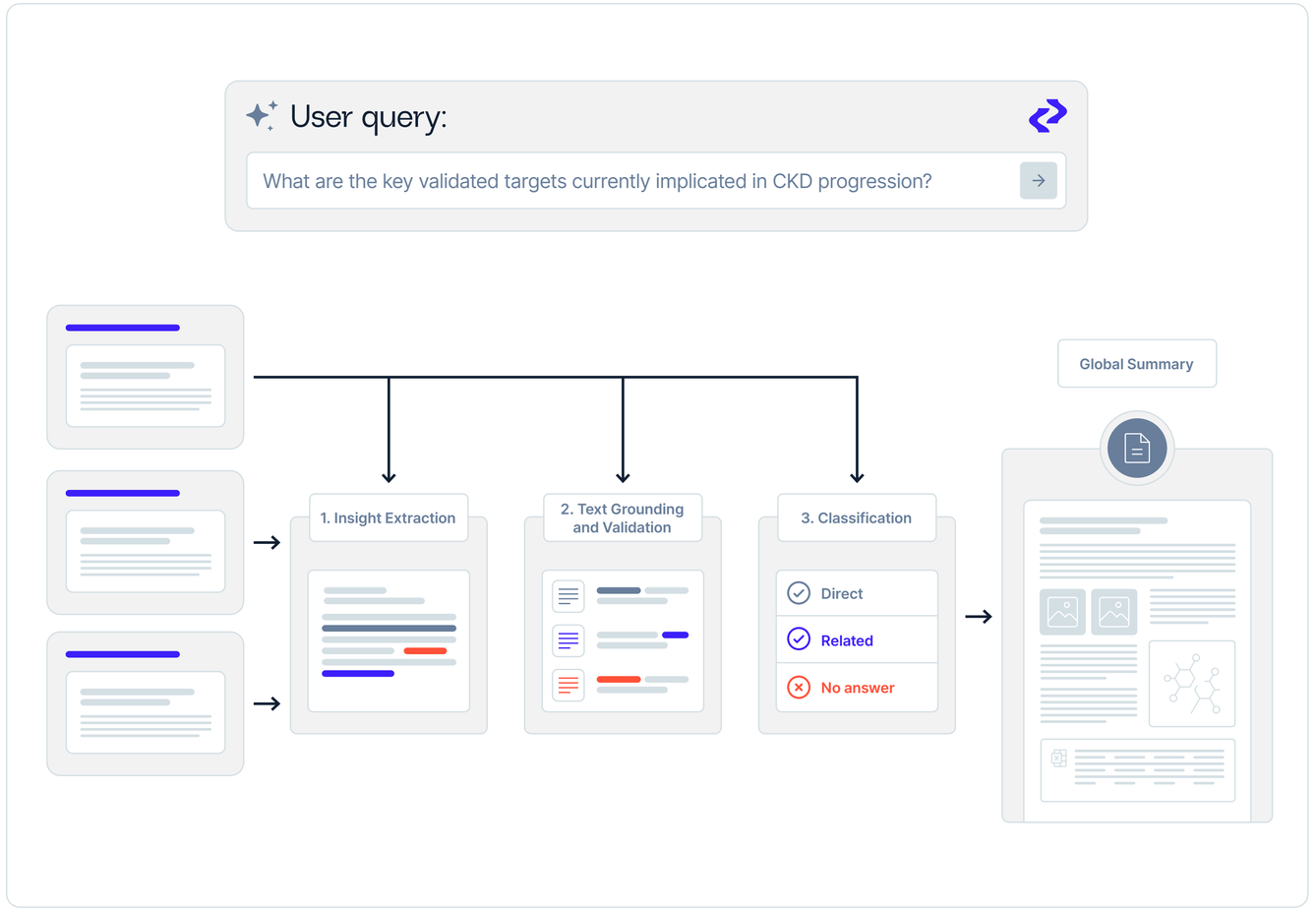
waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Knowledge Synthesis Layer

04

Answer-type aware behavior

It is imperative that scientific GenAI systems are fine-tuned to determine when the source data supports a direct answer, is sharing only related insights, or when the appropriate response is simply “I don’t know.” This guards against overconfident hallucinations when evidence is sparse, while still surfacing adjacent facts that can inform scientific assumptions. It also provides the LLM with a structured framework for synthesis, which improves both precision and usefulness for long-tail scientific questions.



causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

AI applications

Causaly enables scientists to continue their research with an array of AI-powered applications:

Agentic Research

A multi-agent research framework that plans, executes, and synthesizes scientific workflows with transparency and rigor, emulating the scientific process. Agentic Research includes:

- **Deep Research:** Addresses complex, multi-step scientific questions and produces comprehensive, ready-to-share reports by synthesizing insights from internal and external data.
- **Expert Agents:** Purpose-built autonomous agents designed to complete specialized scientific tasks end-to-end.

BioGraph

Analyzes more than 500 million biomedical facts across eight relationship types, enabling users to answer causal questions and explore biological knowledge through multiple visualization modes.

- **BioGraph API:** Provides programmatic access to biological data at scale, supporting rigorous and reproducible bioinformatics analyses.

causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

AI applications

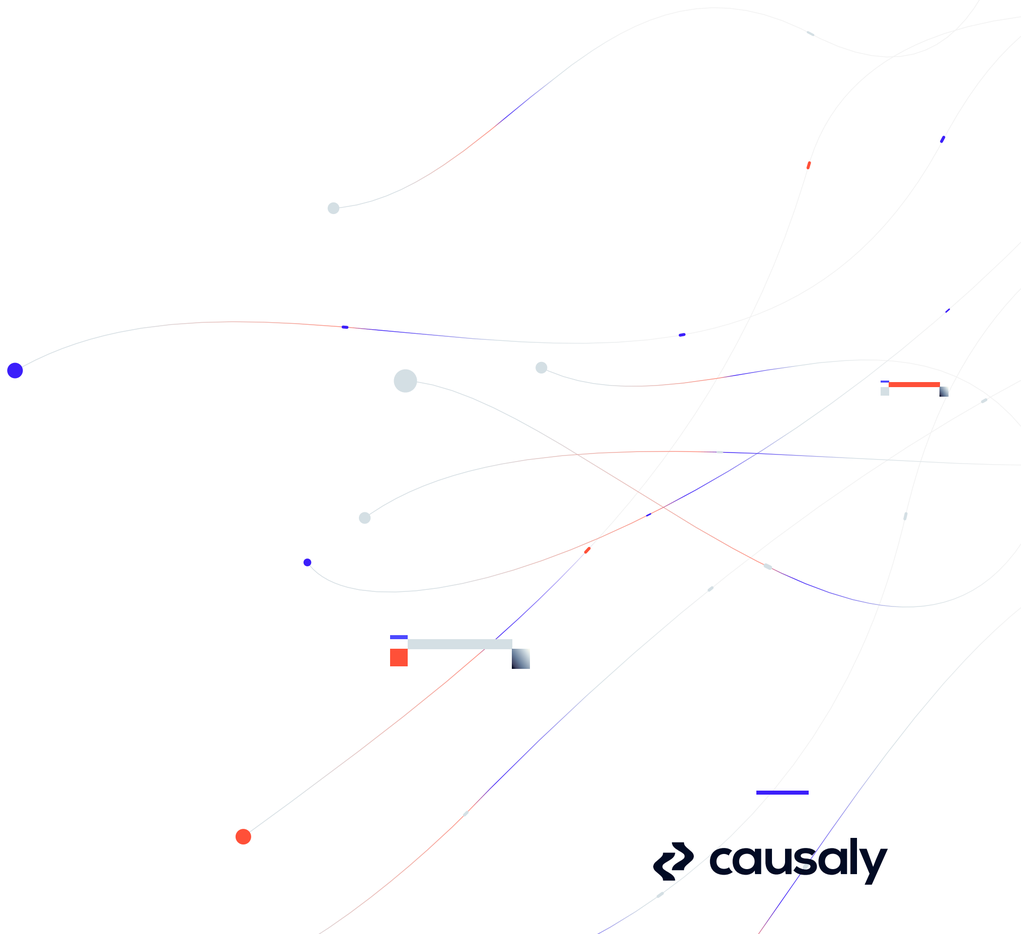
Pipeline Graph

Offers a structured database of more than 100,000 pipeline drugs from over 15,000 organizations, allowing users to search by target, indication, modality, or development phase.

Discover

An intelligent, domain-optimized document search system for navigating scientific literature. Discover includes:

- **Causaly Copilot:** Guides deeper exploration by recommending follow-up questions and related lines of inquiry based on search context.
- **Application Insights:** Enriches results with integrated information from external scientific resources such as UniProt, PubChem, multiple databases, etc.



causaly

waves, antibody
infiltrating
liver disease,
luorescence assay
/MS, intravenous
1, real-time cell
thrombocytopenia
me shifts phase
trial MAPK8
targeting
E11, PBMC, MCF-
frequency, site-
idization (CGH)
ance, alpha
-throughput
HTS), high-
assay systems
mmunoprecipitation
(ChIP-seq)
orted outcomes
metabolic flux
s, adalimumab
eration mapping
co-morbidity
serum, epigenetic
genicity,
immunofluorescence
off-target
xt-generation
(NGS), hybridoma
ntibody
on-alcoholic
er disease
igodendroglia
ity, 96-well
roptosis,
flow cytometry
iniaturization
cell-free circulat
NA (ctDN A)
ug conjugates
brolizumab,
anercept
lux disturbances
trating
(TILs),
tumor cells (CTC
ically disorder
DP) fluorescence
ridization (FISH
bispecific
ellular
turization
inal fluid
otoxicity,
umor
ty, Knockout/
KO/KD), nuclear
sonance (NMR)
y, HCT-8, double-
ks, cancer stem

Benefits for R&D CISO Stakeholders

SIRS provides notable enterprise-level benefits:

- **Reduced Integration Overhead:** The modular architecture simplifies innovation and updates.
- **Enhanced Data Governance and Security:** Comprehensive compliance controls ensure rigorous regulatory alignment.
- **Scalability and Flexibility:** Adaptable integrations facilitate agile responses to evolving research and technological needs.

Causaly's Scientific Information Retrieval System represents a significant advancement for biomedical research. By integrating sophisticated AI technologies and specialized retrieval methodologies, SIRS ensures unparalleled accuracy and scalability, fundamentally enhancing the pace and depth of discovery in life sciences R&D.

Ready to see how Agentic Research can help you make more breakthroughs?

SEE HOW IT WORKS

causaly