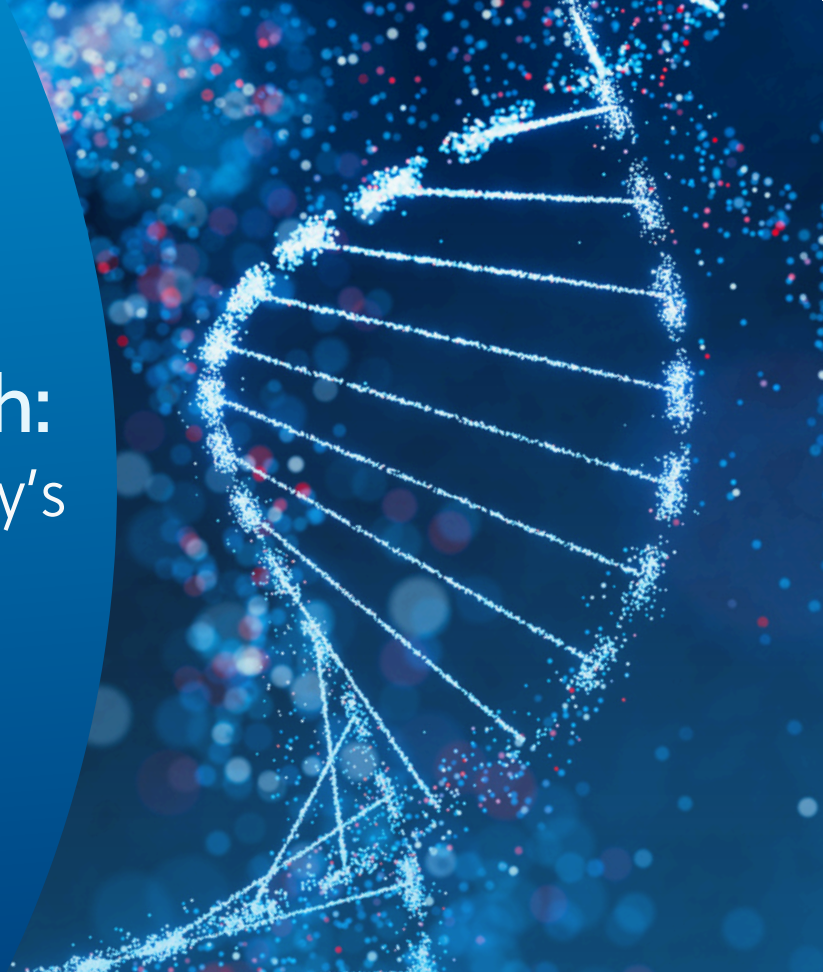# Charting a New Path:
Practical Advice for Today's Discovery Researchers

Discovery researchers have a wide range of responsibilities, and the total time they can devote to searching for scientific evidence and data is relatively small compared to their overall workload. Owing to how important this element of their role is, it's crucial that they have the right tools at their disposal to ensure they are as efficient and thorough as possible when they conduct research.

A unified view of evidence from a variety of sources can help researchers achieve focus and efficiency and avoid the dreaded fear of missing out (FOMO).

The Key Challenges of Discovery Research & Siloed Research

Efficiency & Focus

Bringing a wide range of content and data sources into a single tool reduces the frustrating process of recapitulating the same search in multiple different platforms and systems, stitching them together, deduplicating manually, and hoping that results will be comprehensive. When multiple platforms are required, users worry that they may not have "used the tools" properly to find the data they need.

This concern is compounded by the very nature of early research. Discovery researchers are often looking for an absence of data rather than its presence to try to confirm the novelty of a line of potentially unprecedented research. Concerns around false positives and negatives in search are top of mind — "Am I sure I used this platform properly?"

Discovery researchers have a lot to juggle and often fit their evidence and data searches around other commitments such as experimental lab work. It is also likely they are working on more than one drug program or idea at any time. This makes a unified view of data and information even more important, as eliminating unnecessary steps allows researchers to better maintain their focus.

Fear of Missing Out

The fractionated nature of data, information, and evidence to inform drug discovery across multiple platforms and databases can leave discovery researchers with a feeling that there is an important source that they have not searched or are not even aware exists.

Realistically, there are hundreds of resources out there and it is impossible for any one researcher to be familiar with all of them. They will of course gravitate to their favorite resources based on experience, but there is always a voice in the back of their mind suggesting that they have missed something.

The advent of generative artificial intelligence (AI) created the erroneous belief that the problem of gaps in results has been solved. Systems based on large language models (LLMs) provide interfaces that can add narrative to a set of data, and summarize the content of one or many articles. However, they are unable to satisfactorily address some essential needs of discovery research such as contextual relevancy and domain specific semantics.

Addressing Research Challenges

It is important to remember that drug discovery is a full disclosure activity. Researchers need to be able to explain, ultimately to a regulator, what their idea is, what experiments they have done, what conclusions they drew, how they responded to the data, and how that shaped the drug program from idea to product. Having all the data and information in a unified view is critical to meeting those expectations and dealing with scientific and technical challenges along the way.

From here, we will explore three powerful approaches that empower drug discovery researchers: leveraging ontologies, utilizing search alerts, and effectively employing LLM-based AI systems.

Critical Questions in Target Identification

How can challenges around efficiency, focus, and fear of missing out for early researchers conducting target identification and validation be reduced by leveraging semantically and ontologically enriched solutions?

Critical Questions in Target Identification

One of the biggest challenges in target identification and validation is finding out as much as is known about the target, which usually leads to a series of vital questions:

- What is the normal function of the target?
- Where is it expressed?
- What evidence is there linking it to disease?
- Are there any drugs or other molecules that have been developed for this target?
- What are the cell-based or animal models that are used?
- Are there any programs for this target and/or disease in the clinic?

While these questions all sound straightforward, finding the answers might not be. Data can appear in research articles, company databases, websites, and other sources, including proprietary ones. With so many places to look, it's difficult to say with confidence that you have found all the data and information you sought.

Competing Naming Conventions Compound the Issue

To make matters worse, different resources may not use the same naming conventions for certain drugs, targets, or diseases. There are multiple naming conventions and standards, and these are not universally applied.

For example, the liver condition MASH — metabolic dysfunction-associated steatohepatitis — was until very recently known as NASH, non-alcohol related steatohepatitis. MASH can also be referred to as MASLD, metabolic dysfunction-associated steatotic liver disease. In disease indications where research is rapidly advancing, the naming and relationships of the processes involved also change quickly. Even though researchers in the field know about these permutations, remembering to query different data and content sources with all the possible synonyms is burdensome and error-prone.

Bringing Order to the Data Chaos

Semantic enrichment with curated ontologies does the hard work for researchers by connecting the different names, identifiers, and standards in use in bioscience to ensure a comprehensive search across different sources regardless of the particular terminology used in each source.

Said simply, an ontology makes it possible to search for MASH, NASH, MASLD, or any other name for this liver condition and get a more comprehensive, consistent set of results.

As the research community adopts AI and machine learning, the proper structuring of data and the relationships between data becomes ever more important, with ontologies and semantic enrichment playing a fundamental role.

Utilizing Search Alerts

Given the volume of new publications and data emerging daily, what proactive approaches can researchers take to prevent FOMO and stay up to date with new breakthroughs and emerging trends in their area of therapeutic focus?

Utilizing Search Alerts

New data and articles are being published at an ever-increasing rate. With the advent of pre-print servers, specialized data resources, and <u>1.5 million new citations</u> added to PubMed alone annually, keeping on top of the emerging content is an increasing challenge for discovery researchers.

New publications may offer evidence that invalidates a central hypothesis of a drug program, but it can be time consuming to run the same searches repeatedly, and alerting mechanisms are a great way to help automate review of the latest research.

Being able to set alerts on anything of interest in one unified, semantically aware platform is a great way for discovery researchers to stay on top of new content. This sort of alerting functionality can also be used by research teams to regularly scan for new publications and competitor activity, or product teams to be alerted to new publications mentioning their product.

In some companies, project-driven literature collections are curated by discovery researchers to back-up the in-house knowledge generated through experimentation, or to manage the competitive landscape. Alerting and sharing content properly is a great way to keep team members up to date on the current relevant content to their projects — as well as required reading for people new to the team. Using this type of automated alerting to build correctly managed literature sources is an important activity for many researchers.

The Benefits of Automating Alerts

01. Reduce the risk of missing key information related to your therapeutic area of focus such as emerging drug targets, biomarkers, or novel therapeutic strategies.

02. Save time by automatically receiving relevant info and reducing the manual effort of searching for new data, leaving you more space to focus on analysis and experimentation.

03. Continuously monitor the latest publications and relevant developments to ensure you and your teams stay informed of new advancements in your field.

04. Gain a competitive edge by staying abreast of the latest findings, trends, and breakthroughs.

05. Tailor your information with specific keywords, topics, ontologies, and filters to get actionable insights faster and minimize the risk of missing key info.

The Importance of Copyright Compliance

When keeping up with new content, are there any regulatory or copyright considerations that should be accounted for?

The Importance of Copyright Compliance

At the same time, remaining aware of applicable laws and regulations are essential in supporting the collaboration and organization of content, particularly for drug discovery research as reporting and compliance requirements are critical for regulators. Upholding a commitment to copyright compliance and adhering to terms and requirements set by content providers across teams, departments, and global geographies is vital to commercial research organizations. Without a solution that is copyright aware and promotes the lawful use of content, this can pose a significant challenge and hinder efficiency.

Effectively Employing AI

There is a lot of excitement about the potential applications of AI technology that can improve efficiency in drug discovery research — are there any limitations early researchers should consider when using generative AI tools, and what should researchers keep in mind when evaluating how to responsibly employ AI functionality?

Effectively Employing AI

Many researchers are now using generative AI chat tools to help them summarize the science in large sets of documents and get up to speed on areas of science new to them in a much quicker way than before; instead of reading a pile of papers, they can read the summary and focus on what is important to them.

However, in discovery research, there are some special circumstances to keep in mind that can complicate the use of LLM-based AI systems in exploring data and scientific evidence. When searching for established facts, many of the emerging AI chat tools will readily produce what is supposed to be the state of current understanding or summaries of current research in complex areas. Nonetheless, you should double check these results for accuracy; these tools may not be accurate and effective when you're looking for new or novel research.

LLMs are becoming embedded in many areas of pharma research. Despite this, they are only one tool in the arsenal of discovery researchers and other professionals in the pharmaceutical industry. A recent study [1] has shown that a significant percentage of AI-assisted work is not related to LLMs. Graph-based AI and other machine learning approaches are more appropriate in a wide spectrum of activities.

1 Yu Han and Jingwen Tao, "Revolutionizing Pharma: Unveiling the AI and LLM Trends in the Pharmaceutical Industry," arXiv (Jan. 24, 2024), https://arxiv.org/pdf/2401.10273.

There are other factors to keep in mind, as well. The quality of an AI tool depends on the quality of the data that was used to train it. However, when an LLM-based AI system is trained, its training data is not openly disclosed. This makes it difficult to ascertain the origin of the information that the AI tool provides in a certain output. This can be troublesome for drug discovery where explainability and the provenance of data are essential.

Therefore, discovery researchers should treat results that are generated by an LLM-based AI system with suitable caution and conduct controlled experiments in the same way as they do in the lab to understand how to assess the value and accuracy of the AI output.

It's also critical to consider copyright when using AI tools with third-party content.

A Solution to Address the Challenges of Drug Discovery Research

There are many challenges associated with gathering the data and evidence early drug discovery researchers require to remain on the cutting edge. A solution that tackles efficiency, focus, and FOMO through key functionality that leverages ontologies, search alerts, and effective employment of LLM-based AI systems can help researchers gain confidence that they've formulated a comprehensive query and found all the evidence available.

CCC's RightFind Navigate reduces the likelihood that you'll miss out on critical information while spending less time learning separate interfaces. By bringing together licensed third-party data sources with internal proprietary information and publicly available resources, searches can be conducted across multiple content types — including scientific articles, clinical trials, grants, preprints, patents, and more.

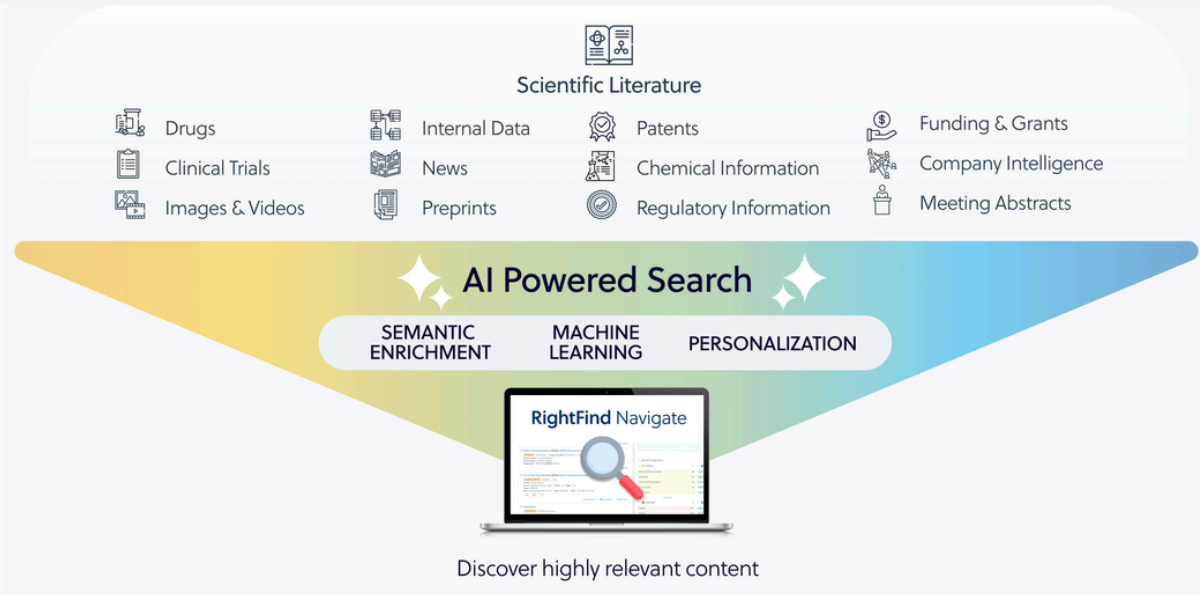## Sample data sources in RightFind Navigate include:

- PubMed
- ChemRxiv
- bioRxiv
- medRxiv
- ChEMBL Compounds
- Europe PMC
- Drugs@FDA
- NIH Clinical Trials

- NIH RePORTER
- PubChem Bioassays
- Compounds and substances
- Internal proprietary content
- Other 3rd party licensed data sources you rely on

Semantic Search

RightFind Navigate's semantic search makes it possible to systematically search multiple sources of evidence with a comprehensive query, saving researchers time and eliminating the uncertainty around using the right words in the right order to find needed intelligence. By using MeSh, SciBite VOCabs, or a custom ontology in RightFind Navigate, researchers can reduce the likelihood that critical content will fall between the gaps.

Search Alerts

Continuously monitor new and changing information relevant to your drug program or therapeutic area with search alerts across a multitude of data sources in RightFind Navigate. Researchers are equipped with the ability to automatically (and compliantly) collaborate with colleagues and stay abreast of up-to-date information that impacts their work.

Scientific Literature

Drugs · Internal Data · Patents · Funding & Grants
Clinical Trials · News · Chemical Information · Company Intelligence
Images & Videos · Preprints · Regulatory Information · Meeting Abstracts

AI Powered Search

SEMANTIC ENRICHMENT · MACHINE LEARNING · PERSONALIZATION

RightFind Navigate

Discover highly relevant content

AI-Enabled Features

CCC has been an industry leader in the discussions about generative AI and copyrighted materials, and our approach to introducing AI-enabled features keeps these responsibilities top of mind. With a focus on supporting organizations in responsibly and effectively employing generative AI, RightFind Navigate is currently operating a beta release of AI Chat functionality.

While RightFind Navigate has leveraged machine learning for personalization and contextualization of content for some time, generative AI features are an exciting forefront of innovation that we are approaching optimistically, but with the caution and careful considerations our clients in the highly regulated life science industry depend on.

Drug discovery research is a critical component of life science organizations. The rapidly evolving information landscape for those in early research roles demands technology solutions that prioritize the functionality essential to drug discovery. Ultimately, the best solutions must amplify researchers' expertise and help organizations deliver transformative therapies to the market more quickly and efficiently than ever before.

Keep Exploring

Visit our **RightFind Navigate page** to gain more insights into how our solution can help researchers solve many of the challenges they face around search, including staying current within their fast-moving fields while remaining copyright compliant.

Learn more about CCC Solutions

Help your team access, integrate, and collaborate more effectively with innovative licensing, content, software, and professional services from CCC.

Learn more

About CCC

A pioneer in voluntary collective licensing, CCC advances copyright, accelerates knowledge, and powers innovation. With expertise in copyright, data quality, data analytics, and FAIR data implementations, CCC and its subsidiary RightsDirect collaborate with stakeholders on innovative solutions to harness the power of data and AI.