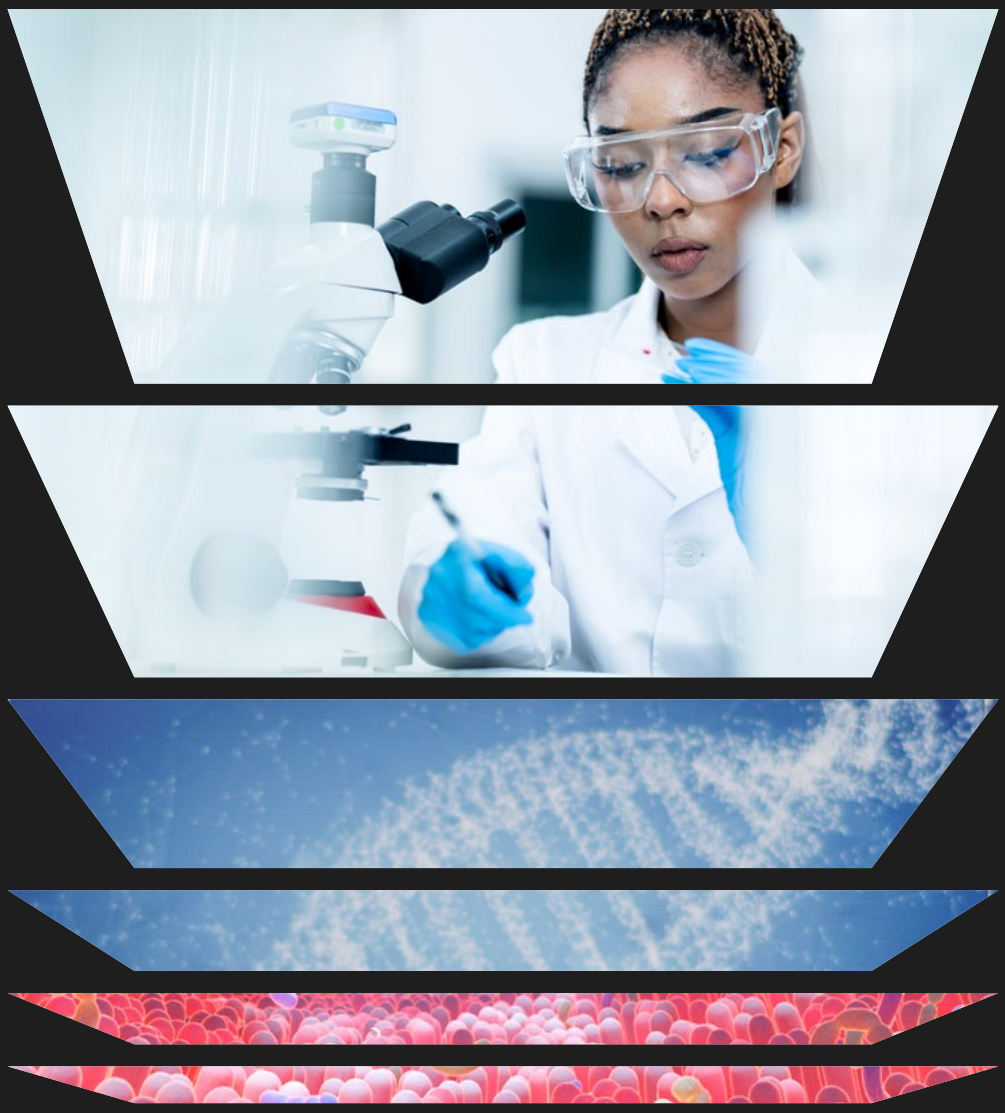
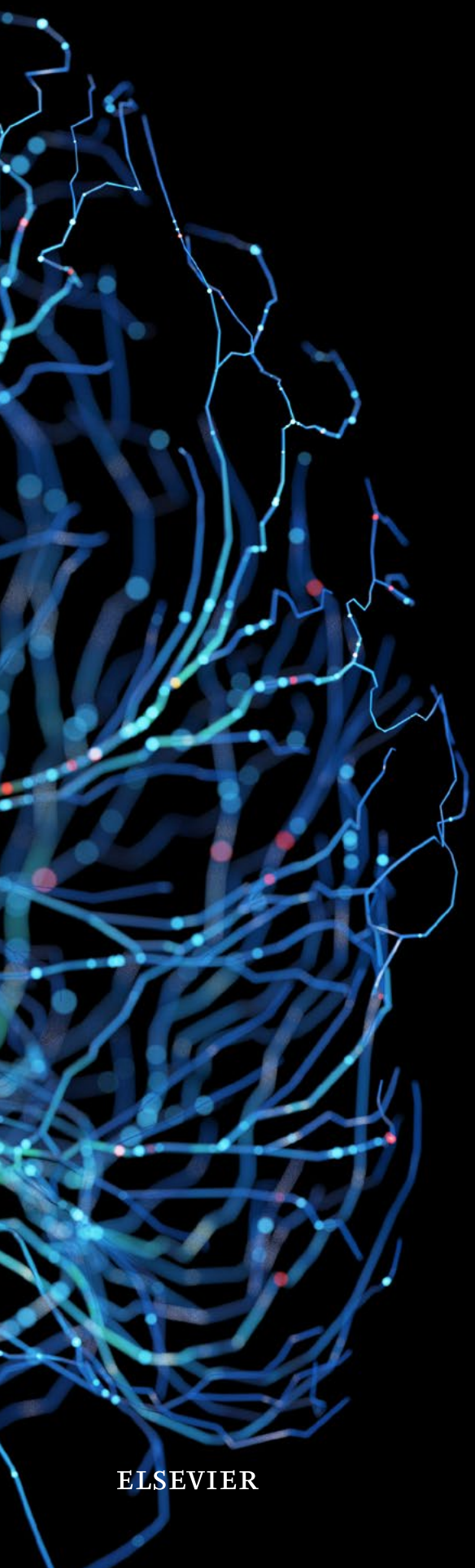


Optimizing Target ID with AI: A six-step framework

ELSEVIER

Advancing human progress together



Contents

Navigating the target ID maze with AI

Target identification (target ID) is one of the most complex and challenging stages of preclinical drug discovery. Researchers must refine a vast pool of potential targets into a focused, workable set that advances a project through the preclinical stages and brings novel treatments to market sooner.

AI is a powerful tool for streamlining target ID because of its ability to process large volumes of complex biological data at scale. By integrating disparate data sources and enabling deep analysis, AI can aid in uncovering new connections between genes, biological pathways and disease mechanisms to generate novel hypotheses far more efficiently than traditional manual methods.

Yet, generic off-the-shelf AI products lack the domain-specific knowledge needed to ensure outputs are both meaningful and scientifically valid. Public models also struggle with hallucinations and unclear data provenance.

For target ID teams, these trust and accuracy limitations pose a serious risk. A misidentified target can result in significant financial, reputational and patient implications. Early choices reverberate throughout the drug discovery pipeline, so evidence-based decision-making is essential.

Amid growing pressure to accelerate drug development, pharma companies face a critical question:

How can we implement AI for target ID in a way that is scalable, trustworthy, and delivers a strong return on investment?

The answer hinges on the triumvirate of technology, data and subject matter experts (SMEs). This article explores how these three pillars form a framework for successfully integrating AI into the target ID process.

Building the bedrock for AI

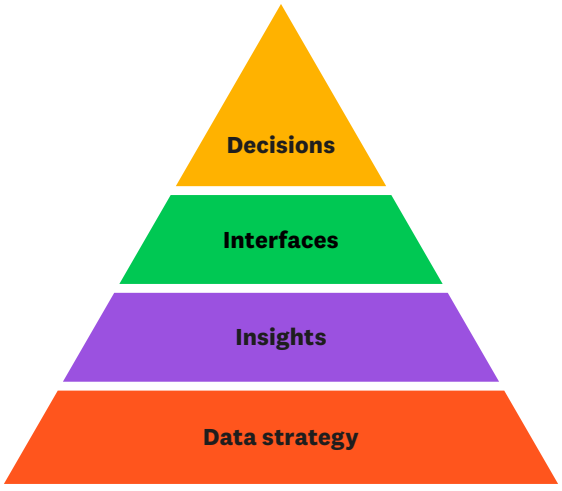
Before even considering how to apply AI to target ID, R&D organizations must first establish a robust data strategy. This strategy forms the foundation for any AI initiative, regardless of intended use case.

Data strategy shouldn't be approached as a one-off transformation. Rather it's a long-term, evolving process that requires continuous refinement as new knowledge, formats and datasets emerge.

A strong strategy enables consistent data management and enrichment across an entire organization. The best data strategies are supported by a semantic layer: a structured framework which adds meaning and context to raw data by mapping data to standard concepts, terms and relationships.

A semantic layer is underpinned by ontologies, which are explored in more depth later in this article. Data principles such as FAIR (Findable, Accessible, Interoperable, Reusable) also play a part in guiding data strategy, ensuring data is machine-ready and reusable.

This foundation can be visualized as a pyramid: **data strategy forms the broad base**, providing the stability and structure upon which AI models are built. **At the top are the insights, interfaces and decisions** that researchers interact with. But without a solid foundation, the whole structure risks crumbling.



With a data strategy in place, teams can then begin the process of applying AI to target ID.

Step 1: *Define* the research problem

Before deciding on which type of AI to use, how to build it, or how to apply it to target ID, organizations must first clearly define the problem they are trying to solve.

Taking a problem-first approach ensures that AI aligns with a targeted, measurable objective, rather than trying to retrofit AI to a target ID research question. Some examples of specific target ID challenges researchers might be trying to solve include:

- Searching for targets with structural or biochemical properties amenable to drug development. (e.g., that have binding pockets for small molecules or a certain number of binding sites)
- Focusing on specific genes
- Narrowing the scope of targets based on pathways where other drugs have been successful in trials

Step 2: Curate trustworthy, *relevant datasets*

Next, researchers can focus on curating AI-ready datasets that align with their research question. There are many datasets, formats and sources relevant to target ID that could be used to train AI models. However, not all will be valuable to the chosen research question.

Rather than training an AI model on all available data, organizations should curate datasets with the most relevant information. This includes filling gaps in their internal assets with high-quality, trusted external datasets to ensure completeness.

SMEs play a critical role in data curation, since their domain expertise enables organizations to refine the scope and format of the data to match the research question.

Checklist for an expertly curated dataset:

- Clearly defined scope that reflects the specific target ID research question
- Shows how data was collected, processed and modified to demonstrate provenance
- Come from trusted, high-quality sources
- Be machine-readable and FAIR
- Include full-text content, not just abstracts
- Contain relevant structural, textual and/or numerical data

Structural

- Shapes of proteins and binding sites
- AlphaFold, CHEMBL and other open source datasets

Textual

- Published scientific literature including journals, conference papers, patents
- Pre-prints
- Regulatory papers and submissions

Numerical

- Internal assay and experimental data
- Clinical trial results
- Omics sequencing data

Step 3: Structure and *normalize data* with ontologies

With the relevant datasets now identified, the next step is to check any external or new datasets are aligned and interoperable with the overarching data strategy and are mapped to ontologies. As mentioned at the beginning of this article, ontologies a critical element of the semantic layer.

What is an ontology?

Ontologies provide a structured framework for defining a scientific domain. Ontologies classify key concepts (e.g., diseases, targets and drugs) and the relationships between them in a manner agreed by subject experts.

Why are ontologies important to target ID?

AI models process data from multiple sources, each stored in different technical or syntactic formats, such as SQL databases, knowledge graphs or document indexes accessed via APIs.

These sources use different semantics to represent the same concepts. Inconsistencies hinder the accuracy and trustworthiness of AI outputs, since models may misinterpret or fail to connect relevant evidence.

Engaging SMEs to build ontologies solves this challenge by creating a common vocabulary and mapping relationships between synonyms in line with what is currently accepted in the field, ensuring integration across datasets. One example would be linking the text string Glucagon-like peptide-1 to the identifier (ID) GLP-1.

Some commonly used ontologies in target ID include MeSH, Gene Ontology, The Disease Ontology (DOID) and Protein Ontology (PRO).

By capturing domain knowledge and scientific context, ontologies provide an information retrieval framework for an AI model.

Just like in the data curation step, engaging SMEs to build and implement ontologies is critical.

This is because ontologies are not static – they evolve alongside our knowledge of a scientific domain. Therefore, SMEs are needed to ensure information is classified and organized according to the latest understanding of a field.

Step 4: Determine the *correct tools and models* for the research question

At this stage, the target ID problem is clearly defined, and the relevant datasets have been selected and structured to be AI-ready. Now, organizations can decide which AI tools are best suited to provide researchers with answers.

AI is a broad term. Different AI architectures have their own limitations and strengths that must be considered. For example:

- Many organizations are attracted by the usability, accessibility and interactivity of natural language Generative AI (GenAI) chatbots based on Large Language Models (LLMs)
- Small Language Models (SLMs) are also of interest in certain use cases due to their specific, task-oriented approach, which may suit target ID
- Machine learning (ML) techniques and statistical models are widely used for analyzing large-scale biological and chemical datasets to uncover patterns that can aid target ID
- Agentic AI uses AI "agents" capable of the autonomous execution of multi-step workflows. This architecture can bring together multiple components, whether ML, LLMs or SLMs, to support complex research processes like target ID

LLM

- Suited to textual data, such as journals, patents, and conference papers
- Good foundation for natural language chatbot
- Can lack deep domain understanding or scientific context
- Struggle with numerical assay data

ML

- Excel at analyzing numerical, imaging and structured biological and chemical datasets for pattern recognition
- Not as user-friendly as an LLM and requires data science expertise for training, fine-tuning, and interpretation

SLM

- Focus on a particular data type or problem, making them more accurate in certain use cases
- Can be more cost effective, requiring less computational power
- Narrower scope: less adaptable to broader, cross-disciplinary research questions

Agentic AI

- Autonomously execute multi-step workflows to deliver significant cost/time savings
- Like LLMs, agentic AI may also lack scientific context
- Reasoning behind an agent's chosen order to execute a workflow can be untraceable, creating a 'black box'

Step 5: Strengthen *reliability and transparency* by grounding AI in fact

For AI to be a trusted tool in target ID, outputs must be grounded in high-quality data and fact.

However, AI architectures like LLMs cannot automatically judge which sources are scientifically reliable. This means when used in isolation to answer research questions tasks, LLMs are prone to hallucinations, bias and irrelevant responses. Such limitations are problematic for target ID, where accuracy and scientific validity are paramount. Hallucinations could even divert drug discovery efforts away from viable therapeutic targets.

Grounding an LLMs in fact is achieved through fine-tuning with domain-specific datasets and by integrating information retrieval (IR) techniques, such as retrieval-augmented generation (RAG).

What is retrieval-augmented generation (RAG)?

RAG ensures an AI model only retrieves responses from a defined set of contextually relevant, up-to-date documents. This enables the model to cite the information it uses in its response, improving transparency, accuracy and users' trust in the system.

Why is RAG important in target ID?

Target ID involves analyzing large corpora of data. RAG optimizes this process by narrowing the search space, reducing the time and computational resources needed to extract actionable insights. This enables researchers to focus on high-confidence targets faster, accelerating early-stage decision making.

RAG enhances data provenance and supports regulatory compliance – both critical in target ID. Evidence behind each decision is documented at the "snippet" level, making transparency easier to demonstrate. This strengthens scientific rigor and builds trust in AI-generated hypotheses.

For R&D leaders, these benefits translate to faster target validation, reduced risk of false leads and a scalable AI-driven research framework.

Interested in Agentic AI? Why creating ‘guided agents’ is a must

AI agents present a huge opportunity for expediting time-consuming multi-step tasks.

For example, a target prioritization agent could take a therapeutic area as input (e.g., Type 2 Diabetes) search across multiple datasets, evaluate evidence, and generate a ranked list of potential therapeutic targets (e.g., GLP1R).

In this example, an AI is allowed to independently decide tools, define steps and execution order. However, this total agency approach creates a “black box” and removes human oversight. Researchers cannot follow the “logic” of decisions the agent took to produce its target list, risking low trust, poor reproducibility, regulatory hurdles and potential false leads.

For evidence-based use cases like target ID, organizations should work with SMEs to create predefined workflows that clearly delineate the tools and steps an agent should use, and in what order.

This guided approach ensures the tools and sequence of tool usage – the logic – are explicitly outlined and documented, so all data considered by the model is from an approved source and relevant to the research question.

Go to the database of disease-gene relationships and pull a set of genes

Go to the druggability info and retrieve information for each gene from step 1

Review the literature for:

a. **evidence supporting the most druggable targets** and associated adverse events

b. **other targets associated with the disease that were not captured in step 1** and can therefore be assumed as being novel

Step 6: Conduct *sophisticated analysis* with knowledge graphs

Next, organizations can move onto analysis. Target ID involves analyzing a network of complex relationships between genes, drugs, disease pathways, and more. Knowledge graphs (KGs) are a sophisticated way of analyzing relationships relevant to a hypothesis at scale.

What are knowledge graphs (KGs)?

KGs are a powerful method of data science representation. Enabled by enriched, structured data, they visually map complex connections and support specialized algorithms and query languages designed for networks.

This makes it easier to uncover hidden relationships and generate deeper insights than is possible with flat, tabular or textual data.

The role of LLMs in KGs

LLMs can support KGs in three primary ways:

- Building KGs by extracting relationships from unstructured biomedical literature
- Querying KGs using natural language processing (NLP) to query syntax
- Ground LLMs in fact, providing a structured knowledge base that improves AI reliability and reduces hallucinations, as explored in step 4

How to make KGs a success in target ID

Just as organizations must define the scope of their data, they must also determine the scope of a KG. While a large, enterprise-wide KG may be beneficial for some applications, it is not always necessary for target ID hypothesis generation, where a targeted, domain-specific approach can yield faster and more relevant insights.

SMEs can define the parameters of what makes a “good” target, and which biological relationships need to be analyzed to reach meaningful conclusions. For example, they might limit the KG to focus on targets with a specific number of binding sites, or pathways where previous drugs have shown success.

By constraining the KG to only relevant relationships, organizations ensure AI-driven insights remain focused, interpretable and actionable for target ID.

What *success* looks like in target ID

If organizations follow the framework outlined and embed the triumvirate of expertise, technology and data, they can expect to realize the following benefits in their target ID projects:

- Productivity gains
- Early identification of non-viable projects
- Cost savings in target safety assessments
- Agentifying manual, routine workflows

Your *partner* for progress

One of the biggest challenges R&D-intensive companies face is to quickly and reliably extract meaning from a vast and growing sea of scientific data. Elsevier seamlessly integrates trusted quality content, advanced technology and scientific expertise to help innovators accelerate discovery and innovation. Partner with us and turn data into discoveries.

Datasets

Power your custom applications and third-party tools with domain-specific curated and enriched Datasets.

SciBite

Get more from your data with text analytics and data enrichment tools from SciBite.

Professional Services

Solve data integration and applied analytics challenges with support from the domain and data science experts on our Professional Services team.

Reach out for a consultation by filling out the form here:
https://www.elsevier.com/industry/pharmaceuticals/contact-us

Let’s shape progress together.