ELSEVIER

Reaxys®

The data bottleneck in AI-powered drug discovery

Contents

Note: This whitepaper was originally published in 2023 under the title "Fit-for-purpose data is key to meaningful AI for drug discovery."

AI: Transforming drug discovery while highlighting data gaps

In a world where ChatGPT has the public reckoning with artificial intelligence (AI), the pharmaceutical industry has been embracing AI's possibilities. Recognizing the inefficiency of a process where 30% of the US $1.1 - 2.8 million cost for a market-ready drug is lost research investment, established enterprises and a rapidly growing legion of AI biotech companies have explored opportunities latent in data for at least a decade.[1,2,3] That work promises a world where medicinal chemists can rely on "machines" to unravel disease, pinpoint targets, and guide discovery.[4] **(Table 1)**

Table 1. Areas in drug discovery where AI can make significant contributions [1, 3, 5, 6]

Area	AI-driven process acceleration
Market monitoring and product repurposing	Tapping into existing investigation and marketed drugs to identify unmet needs and product repurposing.
Target identification	Coalescing real-world and genomics data with published gene networks and biochemical pathways to generate hypotheses about novel targets.
Target identification	Compressing the determination of protein structure and their interactions with candidate drugs from months to just hours.
Target validation & hit identification	Increasing accuracy of high-throughput screening via AI-driven imaging analysis.
Hit identification	Predicting efficacy and toxicity-relevant properties of candidates in silico to shortcut lengthy compound library screenings.
Lead synthesis & optimization	Accelerating the design, synthesis and optimization of lead candidates.
Pipeline decision-making	Prioritizing indication areas for novel mechanisms of action, to optimize the life cycle of existing products, and to build efficiency into drug development programs prior to clinical stages.

In 2022, a selection of emerging AI-driven drug discovery companies had nearly 160 discovery programs and preclinical assets. Fifteen assets were in clinical trials, and novel drug candidates were emerging from AI-focused companies at a faster pace than from conventional pipelines.[7] These AI-supported discovery programs address a key pain point in industry: lengthy iterations of candidate design and testing.[5] Several companies have slashed the three to five years typically needed to identify preclinical candidates to only 12–18 months.[3]

But the AI journey is not that simple. Despite several eye-opening milestones in the last few years – like AlphaFold's prediction of 330,000 protein structures and the FDA recently designating an AI-discovered and designed drug as an "orphan drug"[6,8] – AI has still not penetrated the day-to-day R&D of most drug companies.[5] Limited access to good data is part of the bottleneck. This whitepaper examines data as the bedrock of an accelerated and innovative in silico supported drug pipeline.

A new take on DMTA: cycle less, but better

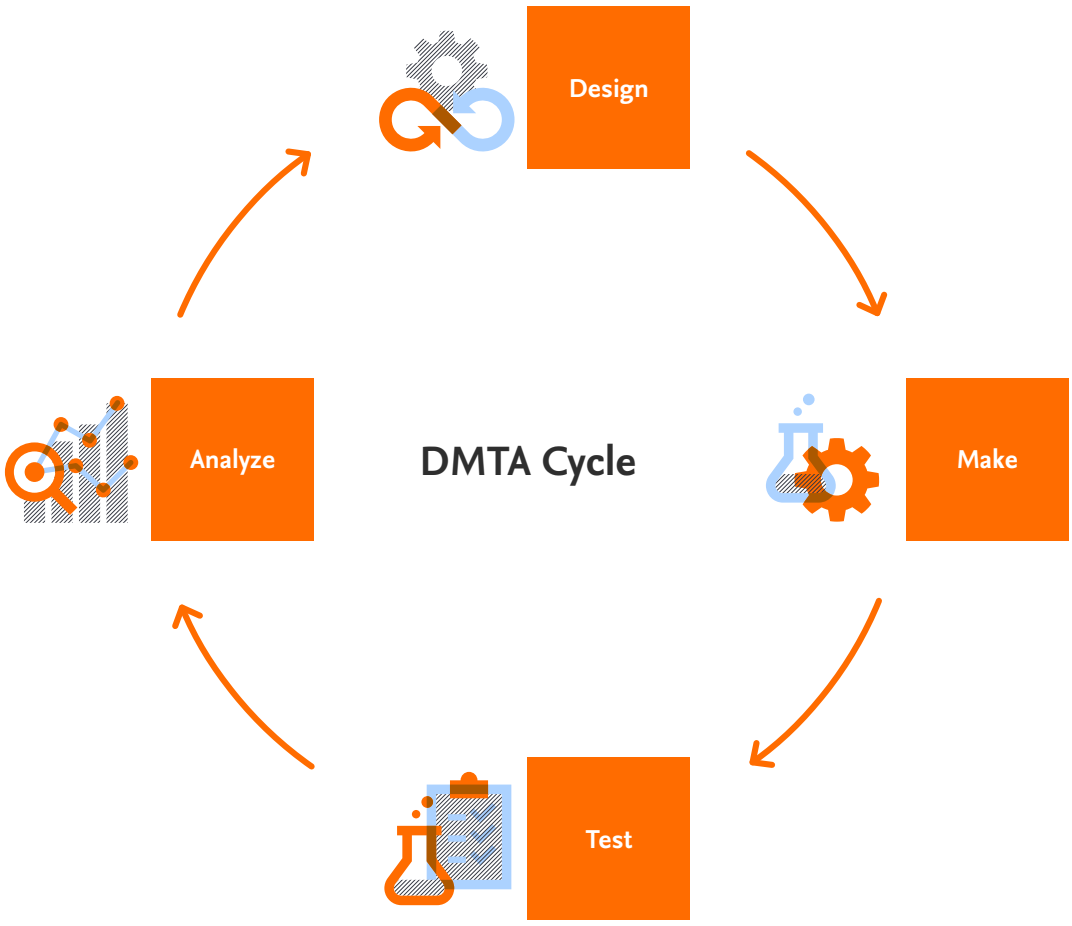
At the heart of the drug discovery process is the Design, Make, Test and Analyze (DMTA) cycle.

This hypothesis-driven iterative loop begins with the design and selection of compound candidates based on structure-activity relationships and pharmacological profiles. After synthesis and purification, the selected molecules are tested to assess ADMET properties, selectivity, mode of action, and affinity. That information feeds a new round of design to ultimately generate candidates with a high probability of success in as little time as possible.[1,9]

The DMTA cycle is time-consuming: Completing an iteration can take four to eight weeks, and most discovery projects require multiple iterations.

However, AI-powered generative and predictive modeling can reduce the number of iterations needed. By leveraging existing and new data, both published and in-house, AI models can optimize compound design and assess synthetic routes in silico before any one candidate is progressed to time-intensive synthesis and testing.

The cycle's design phase especially stands to benefit from AI tools and modeling, not only in speed but also in boosted predictive power to push boundaries in design creativity. How well AI shortens the DMTA cycle depends on the quality of the data used to construct models.



"There has been great advances in the field of molecular ML, and models have permeated almost every step in the DMTA cycle."

– Volkamer, A. et al[9]

In silico insights derive from fit-for-purpose data

Trained on the right data, AI methods like machine learning can assimilate vast knowledge to accomplish creative and extrapolative tasks.

The question is, what constitutes "the right data"? Data that power meaningful AI must fit the purpose of a model in terms of type, how they were collected, and suitability for the intended use. **(Table 2)**

Those aspects of, for example, compound property and affinity data used to train a quantitative structure-activity relationship model ultimately define the quality and utility of that model. Tyrchan et al. additionally point to key attributes of appropriate data, including dataset size, the chemical and property space covered, diversity and noise.[11]

Table 2. Attributes of fit-for-purpose data [12]

Area	Description
Quantity	Training AI-based models to accurately cover the full scope of possible outcomes with high confidence is very data intensive. Quantity goes hand in hand with diversity.
Diversity	Because a model performs solely based on the data upon which it is trained, data diversity is a key aspect to eliminate bias, ensure inclusivity and grant a model more creative space.
Consistency	Consistency ensures that data are comparable and entails normalization across data types, sources and representation.
Accuracy	The data must objectively reflect the properties, events or relationships in question.
Relevance	The data should be up to date and pertinent.
Completeness	At one level, completeness is a combination of quantity and diversity for full coverage of the information space relevant to the problem. At another level, each point and relation in a dataset should include all necessary information for its use.
Machine-readiness	A dataset's access, format, structure and metadata all contribute to making it ingestible with minimal data preparation.

Bioactivity data for compound design

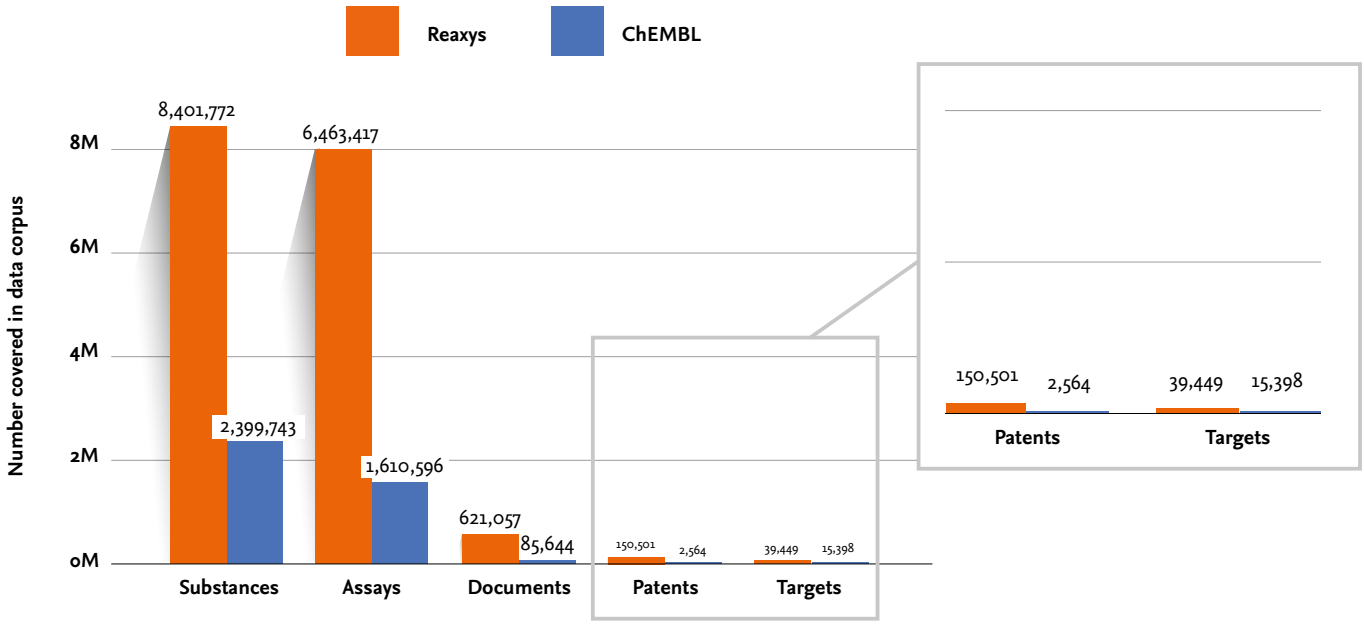
Fit-for-purpose bioactivity data – a description of how complete or partial molecules interact with potential targets – can accelerate compound design. Published and proprietary bioactivity data on millions of compounds are collected and organized into various databases, and the quality and quantity of those data depend on the excerption, ingestion and quality control policies of each repository. Thus, tapping into a database for AI projects is often associated with tradeoffs. For example, the relatively small corpus of commonly used public data repositories can make it necessary to merge multiple sources for data-intensive AI methods.

To better visualize this tradeoff, we compared the bioactivity data contained in ChEMBL with data in Reaxys. **(Figure 1)**

Briefly, ChEMBL is a publicly available, manually curated database with 2.4 million compounds, 15,000 targets and relevant chemical, bioactivity and genomic data from 88,000 documents. Also manually curated, Reaxys is an expertly organized medicinal chemistry database that contains normalized substance-target affinity data for over 8.4 million unique substances and 39,000 targets, sourced from 770,000 documents and patents.
It also includes comprehensive pharmacokinetic, efficacy, toxicity, safety and metabolic profiles, as well as data from in vivo animal studies. As a result, Reaxys not only incorporates more published documents in its database, it also excerpts and organizes details about vastly more substances and assays. **(Figure 1)**

Bioactivity data for compound design

Figure 1.
A quantitative comparison of the bioactivity data corpus of ChEMBL vs Reaxys. For each analysis category, the Reaxys data corpus offers two to seven times more coverage, except in patents, where coverage is several-fold higher.



Featuring deep data excerption and covering a range of assay categories, the massive body of bioactivity data in Reaxys is well-suited to train AI-based models that answer questions for compound identification and optimization. The following examples showcase how Reaxys target and bioactivity data have been used for virtual compound screenings and a priori risk assessment of adverse drug reactions.

Both uses decrease DMTA iterations by maximizing the likelihood that selected compounds will succeed before synthesizing and testing each.

Example 1

A model trained on Reaxys bioactivity data finds matrix metalloprotease inhibitors among a library of natural products in Reaxys[13]

Matrix metalloproteases (MMPs) are responsible for the degradation of extracellular matrix components. Excess expression and activity induced by ultraviolet light contribute to skin aging, which may be ameliorated by an MMP inhibitor. Gimeno, A. et al. developed a virtual screening (VS) workflow to identify candidate compounds that target the conserved catalytic region of binding sites in a set of five MMPs. The VS included four filtering steps:

(1) A random forest model trained on bioactivity data, such as IC50 and Ki for over 50,000 compounds, from Reaxys and ChEMBL

(2) Protein-ligand docking using structures from the Protein Data Bank

(3) A pharmacophoric filter

(4) An electrostatic similarity analysis

They applied the VS to the Specs compound library (more than 45,711 compounds) and extracted hits identified in two or more VS. Of those, they sourced 20 compounds to validate the VS workflow in vitro. Having validated the method, they ran all natural products in Reaxys with a molecular weight of 300–600 Da through the VS workflow. The screening resulted in 183 identified candidates, of which 49 were hits in three or more VS. That two compounds had already been reported to inhibit MMPs and another two were natural products already used in skin applications underscores the quality of the hits. The authors plan to examine the remaining compounds for possible skin treatments.

Example 2

Reaxys structure-activity data train a virtual screening model that improves hit rates for bromodomain inhibitors[14]

Bromodomains are variations on a protein domain that recognize acetylated lysine residues and transduce the corresponding signal into normal or abnormal phenotypes. As such, bromodomain inhibitors are actively pursued as clinical candidates to treat cancer and multiple sclerosis. Seeking to identify novel binders of the bromodomain BRD4, Casciuc, I. et al. used docking and structure-activity data from 1,221 compounds in Reaxys and 672 compounds in ChEMBL to train automated virtual screening (VS) models. They built several support vector machines (SVMs), generative topographic mapping, and structure pharmacophore models to virtually screen 2 million compounds in a proprietary library from Enamine.

An initial compound selection based on consensus between the different models underwent docking analysis to further reduce the pool to 3,000 molecules that were then tested as ligands of BRD4. Concurrently, 3,000 compounds were randomly screened from the same library for similar testing. The VS models delivered 29 experimentally confirmed BRD4 ligands, representing a 2.6-fold improved hit rate over the random screening.

Example 3

Pharmacological and chemical data from Reaxys reveal patterns to predict adverse drug reactions[15]

Looking to anticipate adverse drug reactions (ADRs), Ferro, C. J. et al. used physicochemical, blood-brain barrier, pharmacokinetic, and pharmacological property data to predict the likelihood of ADR for each of four commonly used oral anticoagulants: apixaba, dabigatran, edoxaba and rivaroxaban. They built a predictive model with Reaxys data covering off-target effects, normalized target-affinity data, volume of distribution, plasma protein binding, renal excretion, and blood-brain barrier penetration properties like pKa and clogD7.5. The model highlighted property thresholds predictive of ADR risk.

Based on these, the authors made predictions about possible ADRs associated with each of the four anticoagulants and used real-world data from the MHRA Yellow Card database and prescription rates in the UK to confirm or refute the predictions. In general, the predictions held true. Importantly, the authors predicted that dabigatran would have the least clean off-target profile based on chemical properties related to on-target efficiency, like the degree of nonspecific interacting lipophilic components in a drug. And indeed, dabigatran showed the most overall ADRs and the highest rate of fatalities.

Fit-for-purpose bioactivity data: the engine of AI-supported drug discovery

AI tools have already improved design, optimization and safety evaluation of candidate drugs. While the first AI-generated candidates remain to be fully tested in the clinic, AI is estimated to save 25-50% of the cost of developing a new drug.[1,5]

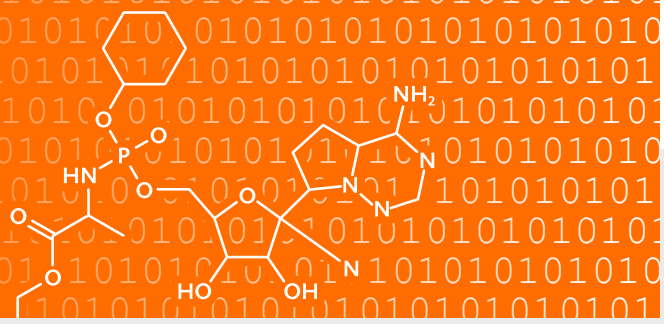
Until now, the use of AI has been narrowly focused on disease characterization, target discovery and small-molecule optimization for just a handful of therapeutic areas. Research efforts have been biased toward oncology, neurology and COVID-19,[5] but guidance and acceleration from good generative and predictive AI could push areas like infectious and environmental diseases into the limelight.

While mostly AI-first biotechs use AI tools routinely, the pharmaceutical industry as a whole is embracing AI, investing in talent and prioritizing fit-for-purpose data.[3,5] Data – its quantity, quality, diversity and readiness for use – are the engine of meaningful AI-supported drug discovery. Given the speed at which chemistry and biomedicine evolve, applying good AI means tapping into databases that maintain data relevance and accuracy through timely ingestion, repeated updates, and careful normalization for comparability across source, data type and time. Those data exist and should be used to realize the full transformative power of AI.

Streamline drug discovery with data fit for the purposes of your AI.

Talk to our Reaxys experts to learn more.

Reaxys®



1 McKinsey & Company. 2022. AI in biopharma research: a time for focus and scale. https://www.mckinsey.com/industries/life-sciences/our-insights/ai-in-biopharma-research-a-time-to-focus-and-scale (accessed July 2023)

2 Wouters, O.J. et al. 2020. Estimated research and development investment needed to bring a new medicine to marker, 2009–2018. JAMA, 323: 844. doi: o.1001/jama.2020.1166

3 Ayers, M. et al. 2022. Adopting AI in drug discovery. https://www.bcg.com/publications/2022/adopting-ai-in-pharmaceutical-discovery (accessed July 2023)

4 Roberts, M. and Genway, S. 2019. How artificial intelligence is transforming drug design. https://www.ddw-online.com/how-artificial-intelligence-is-transforming-drug-design-1530-201910/ (accessed July 2023)

5 Unlocking the potential of AI in drug discovery. A report from BCG, commissioned by the Wellcome Trust. https://web-assets.bcg.com/86/e5/19d29e2246c7935e179db8257dd5/unlocking-the-potential-of-ai-in-drug-discovery-vf.pdf (accessed July 2023)

6 Chun, M. 2023. How artificial intelligence is revolutionizing drug discovery. https://blog.petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery (accessed July 2023)

7 Jayatunga, M.K. et al. 2022. AI in small molecule discovery: A coming wave? Nature Reviews Drug Discovery, 21: 175. doi: 10.1038/d41573-022-00025-1

8 Insilico Medicine. Press release, 8 February 2023. Insilico Medicine receives FDA Orphan Drug designation for generative AI discovered and designed drug for idiopathic pulmonary fibrosis. https://www.globenewswire.com/news-release/2023/02/08/2604040/31533/en/Insilico-Medicine-Receives-FDA-Orphan-Drug-Designation-for-Generative-AI-Discovered-and-Designed-Drug-for-Idiopathic-Pulmonary-Fibrosis.html (accessed August 2023)

9 Volkamer, A. et al. 2023. Machine learning for small molecule drug discovery in academia and industry. AILSCI 3: 100056. doi: 10.1016/j.ailsci.2022.100056

10 Schneider, P. et al. 2019. Rethinking drug design in the artificial intelligence era. Nature Rev. Drug Disc. 19: 353. doi: 10.1038/s41573-019-0050-3

11 Tyrchan C., et al. 2022. Chapter 4 – Approaches using AI in medicinal chemistry. Pp. 111-159 In Computational and Data-Driven Chemistry Using Artificial Intelligence; Fundamentals, Methods, and Applications. Ed. Takashiro, A. Elsevier. Doi: 10.1016/B978-0-12-822249-2.00002-5

12 Ataman, Altay. 2023. Data quality in AI: challenges, importance and best practices. https://research.aimultiple.com/data-quality-ai/#:~:text=What%20are%20the%20key%20components%20of%20quality%20data,to%20incomplete%20or%20biased%20results.%20...%20More%20items (accessed August 2023)

13 Gimeno, A. et al. 2021. Identification of broad-spectrum MMP inhibitors by virtual screening. Molecules 26: 4553. doi: 10.3390/molecules26154553

14 Casciuc, I. et al. 2019. Pros and cons of virtual screening based on public "Big Data": in silico mining for new bromodomain inhibitors. Eur. J. Med. Chem. 165: 258. doi: 10.1016/j.ejmech.2019.01.010

15 Ferro, C.J. et al. 2020. Relevance of physicochemical properties and functional pharmacology data to predict the clinical safety profile of direct oral anticoagulants. Pharmacol Res Perspect. e00603. doi: 10.1002/prp2.603

ELSEVIER

For more information or to book a demo, visit
Elsevier.com/products/reaxys/drug-discovery

Reaxys is a trademark of Elsevier Ltd. Copyright © 2024, Elsevier. May 2024

Reaxys®