What Are the Data Requirements for AI Drug Discovery?

Rasit Dinc

Rasit Dinc Digital Health & AI Research

Published: October 10, 2019 | Drug Discovery and Pharmaceutical AI

DOI: 10.5281/zenodo.17998794

Abstract

Artificial intelligence (AI) is rapidly transforming the landscape of drug discovery, promising to accelerate the development of new therapies and reduce the...

What Are the Data Requirements for AI Drug Discovery?

By Rasit Dinc

Artificial intelligence (AI) is rapidly transforming the landscape of drug discovery, promising to accelerate the development of new therapies and reduce the high costs and failure rates associated with traditional methods. At the heart of this revolution lies a critical component: data. The quality, quantity, and diversity of data used to train AI models are paramount to their success. This article explores the essential data requirements for AI in drug discovery, the challenges that need to be addressed, and the path toward building a robust data ecosystem for a future of data-driven medicine.

The Fuel for the AI Engine: Key Data Requirements

For AI algorithms to effectively learn and make accurate predictions, they need to be fed with high-quality data. The specific data requirements can vary depending on the AI application, but some general principles apply across the board.

High-Quality, Large, and Diverse Datasets

The adage "garbage in, garbage out" holds particularly true for AI in drug discovery. AI models require vast amounts of high-quality data to learn the complex relationships between chemical structures, biological targets, and clinical outcomes. This data can come from a wide range of sources, including:

Genomic, Proteomic, and Transcriptomic Data: *'Omics' data provide insights into the molecular mechanisms of diseases and can help identify new drug targets [1].* **Chemical Libraries:** Large databases of chemical compounds, such as PubChem, ChEMBL, and ZINC, are essential for training

AI models to predict the properties and activities of molecules [1]. ***Clinical Trial Data:*** *Data from past clinical trials can be used to train AI models to predict the efficacy and safety of new drug candidates.* **Electronic Health Records (EHRs):** EHRs contain a wealth of real-world data that can be used to identify patient populations for clinical trials and to monitor the long-term effects of drugs.

The Need for Data Standardization

One of the biggest challenges in using AI for drug discovery is the lack of data standardization. Data from different sources are often in different formats, making it difficult to integrate and analyze. This can lead to inconsistencies and errors in AI models. As noted by Ferreira et al. (2025), the lack of clear and standardized methodologies is a significant limitation, especially for 'omics' data [1]. To address this, there is a growing need for common data standards and ontologies that will allow for seamless data sharing and integration. The FDA has also recognized this need, emphasizing the importance of data standards in their draft guidance on AI in drug development [2].

Data Accessibility and Open Science

While many valuable datasets are publicly available, a significant amount of data remains locked away in proprietary databases within pharmaceutical companies and research institutions. This data siloing hinders collaboration and slows down the pace of research. The open science movement, which advocates for making scientific research and data freely available, is crucial for advancing AI in drug discovery. Initiatives that promote data sharing and collaboration are essential for building the large and diverse datasets needed to train robust AI models.

The Importance of Context and Metadata

Data without context is of limited value. For AI models to make meaningful predictions, the data needs to be accompanied by detailed metadata, including information about the experimental conditions, the analytical methods used, and the patient demographics. This contextual information is crucial for interpreting the results of AI models and for ensuring their translatability to real-world applications. As Zhang et al. (2025) highlight, the ability of AI to process vast amounts of data is a key advantage, but this data must be well-understood to be truly useful [3].

Challenges in Data for AI Drug Discovery

Despite the immense potential of AI, several data-related challenges need to be overcome to fully realize its promise in drug discovery.

Data Silos and Lack of Sharing

As mentioned earlier, data silos are a major obstacle. The pharmaceutical industry is highly competitive, and companies are often reluctant to share their data. This lack of data sharing limits the size and diversity of the datasets available for training AI models, which can in turn limit their

performance and generalizability.

Data Quality and Noise

Real-world data is often messy and incomplete. It can contain errors, missing values, and noise, all of which can negatively impact the performance of AI models. Data cleaning and preprocessing are therefore critical steps in any AI workflow. However, these steps can be time-consuming and require domain expertise.

Bias in Datasets

AI models can only be as unbiased as the data they are trained on. If the training data is not representative of the broader population, the resulting AI models may be biased and could even exacerbate existing health disparities. For example, if a model is trained primarily on data from a specific ethnic group, it may not perform as well for other ethnic groups. It is therefore crucial to ensure that the data used to train AI models is diverse and representative of the populations they are intended to serve.

The Path Forward: Building a Robust Data Ecosystem

To overcome these challenges and unlock the full potential of AI in drug discovery, a concerted effort is needed to build a robust and equitable data ecosystem.

Collaborative Data Initiatives

Collaborative data initiatives, where pharmaceutical companies, academic institutions, and non-profit organizations come together to share data and resources, are essential. These initiatives can help to break down data silos and create the large, diverse datasets needed to train powerful AI models.

FAIR Data Principles

The FAIR data principles (Findable, Accessible, Interoperable, and Reusable) provide a framework for making data more valuable and reusable. By adhering to these principles, we can create a data ecosystem where data is easy to find, access, and use for research.

The Role of Regulatory Bodies

Regulatory bodies, such as the U.S. Food and Drug Administration (FDA), have a crucial role to play in setting standards for data quality and in promoting data sharing. The FDA has already taken steps in this direction by releasing draft guidance on the use of AI in drug development [2].

Conclusion

Data is the lifeblood of AI in drug discovery. To build effective and reliable AI models, we need high-quality, large, and diverse datasets that are standardized, accessible, and well-annotated. While there are still many challenges to overcome, the ongoing efforts to build a more robust and collaborative data ecosystem are paving the way for a future where AI-driven

drug discovery can deliver on its promise of bringing new and better medicines to patients faster.

References

[1] Ferreira, F. J. N., Carneiro, A. S., & et al. (2025). AI-Driven Drug Discovery: A Comprehensive Review. *ACS Omega*, 10(23), 23889–23903. https://pmc.ncbi.nlm.nih.gov/articles/PMC12177741/

[2] U.S. Food and Drug Administration. (2025, January 15). *AI Drug Development: FDA Releases Draft Guidance*. Foley & Lardner LLP. https://www.foley.com/insights/publications/2025/01/ai-drug-development-fda-releases-draft-guidance/

[3] Zhang, K., et al. (2025). Artificial intelligence in drug development. *Nature Medicine*. https://www.nature.com/articles/s41591-024-03434-4
