WORKFLOW INFORMATICS CORP.
GET IT DONE

From Data to Drug Discovery

An optimized approach to research informatics, unlocking the power of machine learning and generative AI in drug discovery applications.

Machine learning (ML) has become a cornerstone of the drug discovery process, offering new tools that enhance the effectiveness of pharmaceutical research. Today, many companies are taking these developments a step further using the transformative power of generative AI (Gen AI) to search for molecules under specific constraints, such as solubility, therapeutic success and patent status. In doing so, Gen AI is poised to enhance the efficiency, speed and creativity of the new drug discovery process in profound ways.

Yet despite their potential to revolutionize the way scientists explore molecular spaces, ML and Gen AI depend on the quality of the data that feeds them. The reality is, however, many companies neglect the critical data engineering steps in the rush to deploy ML and Gen AI technologies, undermining their efficacy as research tools.

Successful implementation of ML and Gen AI in drug discovery therefore requires an optimized approach to research informatics — merging data science practices into drug discovery pipelines to accelerate innovation.

The Problems With Unstructured Data

Approaching AI-powered drug discovery without deep expertise in medicinal chemistry and understanding of assay data can hinder model development and performance. Without this knowledge, you can introduce data management inefficiencies that bias model outputs toward chemically implausible or suboptimal structures.

One example of this is improper handling of tautomer data. Tautomers may look structurally different from each other, but they can be functionally equivalent. Two experienced chemists can treat two tautomers differently, and many chemical databases will represent these tautomers as completely different compounds. This can create data management challenges where relevant information like properties aren't associated with the correct structures.

Failure to identify a canonical tautomer is another issue. Without a standardized canonical tautomer form, the dominant resonance form won't be registered within the database and will incorrectly bias the AI model to predict compounds with less relevance.

Properly addressing tautomer data has cost implications as well. Many research organizations request assays of multiple tautomeric forms that are functionally identical. This duplication of efforts directly contributes to increased project costs and timelines.

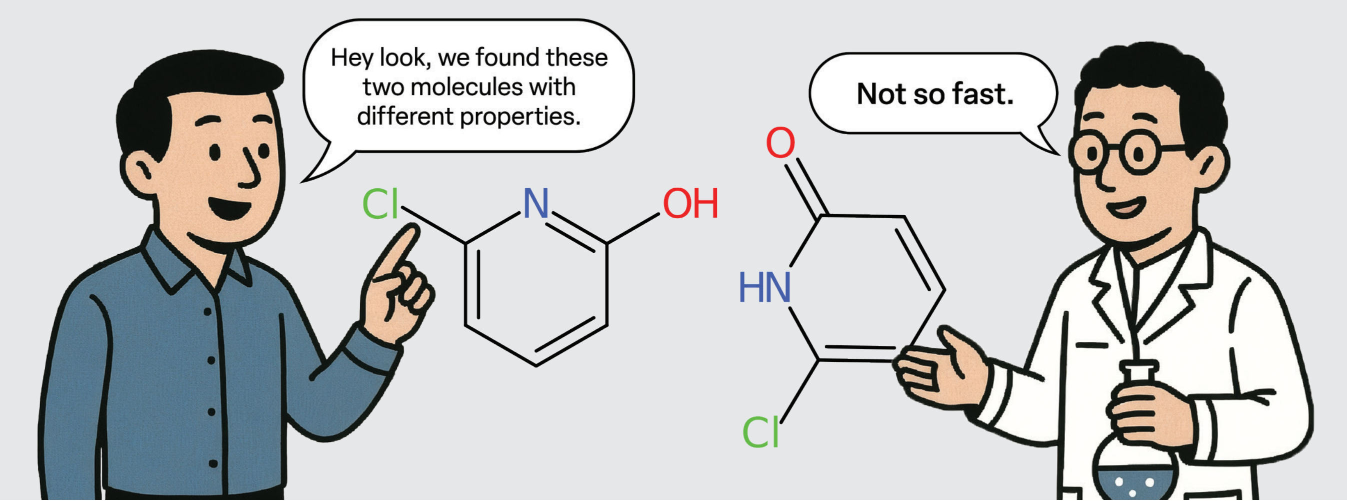
It isn't a given that data engineers would be aware of chemical distinctions to this degree of depth — although they might. However, oversights like mishandled tautomer data can be avoided when the data engineers structuring chemical databases for AI are also chemistry specialists.

Data Processing for Chemical Research

Medicinal chemists and other professionals implementing advanced ML and Gen AI models often need to apply strategies to their drug discovery workflows that are outside their expertise. While an additional university degree in data science isn't required, adopting methods from this field is extremely beneficial to optimizing pharmaceutical research informatics.

Domain-specific data management is fundamental to high-quality AI and ML implementations.

The Trouble With Tautomers


Hey look, we found these two molecules with different properties.
Cl
N
OH
Not so fast.
O
HN
Cl

Understanding tautomeric distinction is one example where chemistry domain expertise ensures accurate and efficient data management for Gen AI development. These structural isomers are functionally equivalent but their differing 2D representations can cause them to be registered as two separate compounds in a chemical database. Lacking a standardized canonical tautomer can incorrectly bias the AI model towards less-relevant compounds. Improperly addressed tautomers can also duplicate efforts, such as requesting assays for multiple functionally identical tautomeric forms.

Successfully integrating Gen AI with drug discovery will involve these key extract, transform, load (ETL) and processing tasks and more.

Processing Chemical Structure and Biological Data

Since ML and Gen AI models require standardized and structured data, it's important to address the wide range of data types and formats involved in pharmaceutical research to create an AI-powered cheminformatics and bioinformatics pipeline.

Case in point, while chemical formats like SMILES, InChI, SDF and MOL all describe molecular information, their different data types impede their integration into one standardized structure. SMILES and InChI are one-dimensional text data types, for example, while SDF/MOL files are multidimensional, preventing these formats from being easily ingestible by AI. How to handle stereochemistry, 3D conformation, salts and tautomers are additional considerations to keep in mind.

Biological data also requires processing. Although protein databases like PDB and UniProt are extremely useful, they store information in different data types. Further, the bioassay information in databases like ChEMBL and PubChem typically contains inconsistent units to measure binding effectiveness, requiring normalization into a standard unit. Among the various databases, proteins also tend to have different IDs and naming conventions, which can be resolved by mapping names to each other.

Integrating Text-Based Information

A wide variety of formats can represent text information, making it necessary to preprocess data before ingesting into a large language model (LLM). Harmonizing information contained within publications, patents, notes, safety data sheets and clinical trials enables an LLM to work with the data.

LLMs require extensive data preprocessing to accommodate various text formats like PDFs, tables, images and labels. Tools for web scraping or optical character recognition can extract raw and structured text data for the LLM. After extraction, the data must be normalized into a structured format like a JSON file. Converting this information into numerical vectors that the LLM can "understand" enables structured reasoning to form relationships like drug interactions.

Machine Learning and Generative AI in Drug Discovery

While ML and Gen AI in pharmaceutical research are closely related, their specific applications differ slightly. ML models are often used to identify or predict compounds based on properties or performance — for example, predicting a compound's toxicity, binding affinity or patient response.

On the other hand, Gen AI can create entirely new molecular structures based on learned characteristics of known compounds. Using the same example, Gen AI could be used to generate novel structures with desired toxicities, binding affinities and patient responses.

ML models used in Gen AI — including variational auto encoders (VAE), generative adversarial networks (GAN) and recurrent neural networks (RNN) — have the capability to generate new molecules based on existing molecules. Input data can be images like chemical drawings depicting structure, or text-based data like SMILES strings.

Similarly, Gen AI using these models could produce new chemical drawings and SMILES strings based on the data of known compounds. Keep in mind, Gen AI architectures are built with ML components — so the same data input requirements apply.

Implementing Cleanup and Quality Control

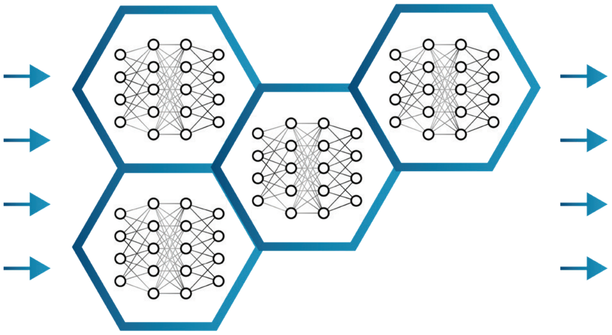
During this process, standard data processing quality checks — such as detecting and handling missing values in tabular data — still apply. Other procedures to implement include removing outliers and validating the correctness of chemical structure information. Because of the large volumes of data involved in drug discovery, it is also beneficial to optimize data structures to speed up Gen AI and ML processing.

Gen AI Drug Discovery in Action

Drug discovery projects often involve multiple organizations sharing information such as assay data and synthesis requests. Integrating the disparate information systems into one centralized platform will streamline cooperation and reduce project timelines. Implementing ML and Gen AI tools into this central system further amplifies these benefits.

Unfortunately, the incompatibility of each organization's database and content may hinder seamless integration into one consolidated structure. Designing a comprehensive data model that reconciles the differences between cooperating organizations provides a solution.

This unified data model (UDM) facilitates easy data-sharing among drug discovery partners and can even be designed to enable ML and Gen AI use.



Machine learning models like VAEs, GANs and RNNs rely on diverse sets of structured chemical data for drug discovery applications such as generating new molecular structures.

How We Did It

Our specialty in streamlining pharmaceutical workflows for drug discovery allows us to optimize information systems that integrate with chemistry, pharmaceutical, ML and AI applications. One example is the ***BioChemUDM***, which can represent compounds and assays for capturing, reporting and sharing biological and chemical data among pharmaceutical companies — regardless of the platforms used to manage chemical registration and assay data.

BioChemUDM registers compounds with a stereo-enhanced sketch according to specified drawing rules. For further categorization, this data model also applies string-based labels associated with the registered compounds. These labels, which inform users of useful information on molecules, batches, samples and assays, are easily parsed by commercial applications, reducing application integration complexity and time.

Because tautomerism plays an important role in drug performance, effectively managing tautomer data is crucial to designing a data model that is optimized for drug discovery. Within BioChemUDM, we addressed tautomers by normalizing tautomeric states according to SMIRKS patterns.

Assay data also deserves special attention. We organized assay information according to broad categories that would group similar assays together. This allowed us to differentiate assays based on types of activation, binding, inhibition, oxidation and more. Within the assay identification strings are fields to identify an assay's test article, format, target, documentation and other fields.

Note that BioChemUDM doesn't rely on a new standard of representing compounds. Instead, this data model uses the common V3000 format to register chemical structures according to their enhanced stereochemical connection table. Registration is done via sketches according to standardized drawing rules. By standardizing one identification format, BioChemUDM eliminates user-specific terms to describe stereochemistry. Instead, it relies entirely on chemical graph theory, enabling organizations to seamlessly register their individually siloed data into the UDM.

To date, multiple companies have embraced BioChemUDM, finding it easy to adopt string labeling and sketch rules to label their data. Using this system, critical information between organizations becomes sharable within the same day.

Emphasize Both Chemistry and Data Expertise

The challenges ML and AI implementation in drug discovery aren't automatically resolved by applying data engineering practices. Given the widely varied types, formats and structures of chemical and biological data used in pharmaceutical research — it can be a challenge to effectively apply data science methodology without possessing drug discovery domain knowledge.

Without domain-specific knowledge of drug discovery — chemistry, assay, and industry information — significant oversights can be made. Challenges are exacerbated when multiple organizations with siloed data systems must collaborate.

Workflow Informatics (WFI) is ideally suited for addressing the needs of drug discovery because of our expertise as both data engineers and medicinal chemists. We leverage our decades of cheminformatics experience to design UDM's that provide drug discovery collaborators with powerful AI and ML tools. Our expertise and solutions are applicable beyond pharmaceutical research as well, including other disciplines like materials, agriculture and food science, aerospace and automotive industries.

*To learn more, check out our paper on **BioChemUDM** or **contact us.***

WORKFLOW INFORMATICS CORP.
GET IT DONE