CDD VAULT®
Complexity Simplified

Customer Case Study

"CDD is a key enabler, because that's where all our private data is stored, and we need to extract information and structure it in the right way so that our LLM can analyze the data, and return results that the medicinal chemist can use."
- Sung Jin Cho, PhD, Founder and CEO of CIMPLRX

CIMPLRX
SIMPLIFY DRUG DISCOVERY

Bridging the Gap Between Medicinal Chemistry and AI: How CIMPLRX Uses CDD Vault to Power Explainable Drug Discovery

Founded in 2017 in Seoul, South Korea, drug discovery firm CIMPLRX is harnessing advanced computational tools to drive the identification and development of safer, more effective medicines. The company integrates a proprietary explainable AI-guided platform, CEEK-CURE (Create Explainable Knowledge and Collect and Uncover Relationships), with structured experimental datasets spanning chemistry and biology. This approach directs compound discovery and optimization, making workflows more insightful, actionable, and transparent.

Headed by founder and CEO Sung Jin Cho, PhD, CIMPLRX operates a flexible business model. Alongside an in-house drug discovery pipeline focused on neuropathic pain targeting kinase AAK1, the company collaborates with biotechnology and pharmaceutical partners across oncology, autoimmune disease, and metabolic disorders. Cho spent the early years after founding CIMPLRX developing its computational platform. After securing Series A funding in 2020, the company established its own chemistry and biology laboratories to advance its neuropathic pain program. CIMPLRX secured Series B funding in 2023 and has also received multiple government R&D grants.

Centralizing Data for AI Readiness

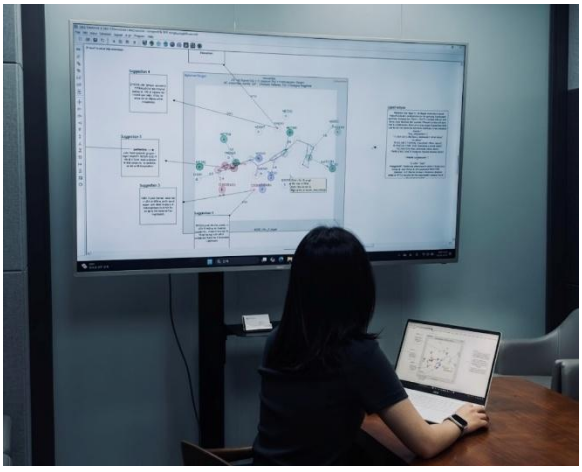
From the outset, data management has been a crucial component of the company's infrastructure and a cornerstone supporting downstream AI workflows. In 2022, CIMPLRX adopted CDD Vault as its central repository for compound registration, medicinal chemistry information, and experimental results. CDD Vault houses the company's experimental data layer, which is vital for SAR extraction, compound exploration, and explainable, structure-aware AI reasoning.

Managing complex experimental, biological, and chemical datasets remains a major challenge in drug discovery. Data diversity, lack of standardization, and difficulties in contextualizing different data types can all limit data utility, both for AI applications and traditional workflows. CDD Vault addresses these challenges through its API, enabling CIMPLRX to store all types of data in a single environment and seamlessly extract and structure it for downstream AI models and analytical tools.

Bridging Domain Expertise and Computation

CIMPLRX distinguishes itself from many AI drug discovery companies by starting from the perspective of medicinal chemistry rather than computer science. "My background is in medicinal chemistry," Cho said. "I'm a domain expert first, and it's from that perspective that I started developing platforms to answer the questions that medicinal chemists actually have." The company prioritizes these domain-specific questions and then applies advanced AI tools to solve them.

The key challenge, Cho emphasizes, is not just building AI models but providing the right context. "Large language models are very good at reasoning, but they need the right context to reason with," he explained. CDD Vault enables CIMPLRX to structure data in a way that supports effective reasoning by large language models. For CIMPLRX, the Vault serves as a trusted data warehouse. Scientists extract structured data from it and feed it into the internal computational environment. "The Vault's API-based architecture is essential," Cho noted. "Our platform relies heavily on data analysis, and so having strong API access to the data is very important. When we started looking for a chemical registration and experimental data management system, CDD Vault was the only platform we seriously considered."



Seamless Integration via API

And while databases are all designed to store data well, Cho added, the data structures needed for applications are different. CDD Vault provides a centralized, structured data pool with rich metadata that can be seamlessly

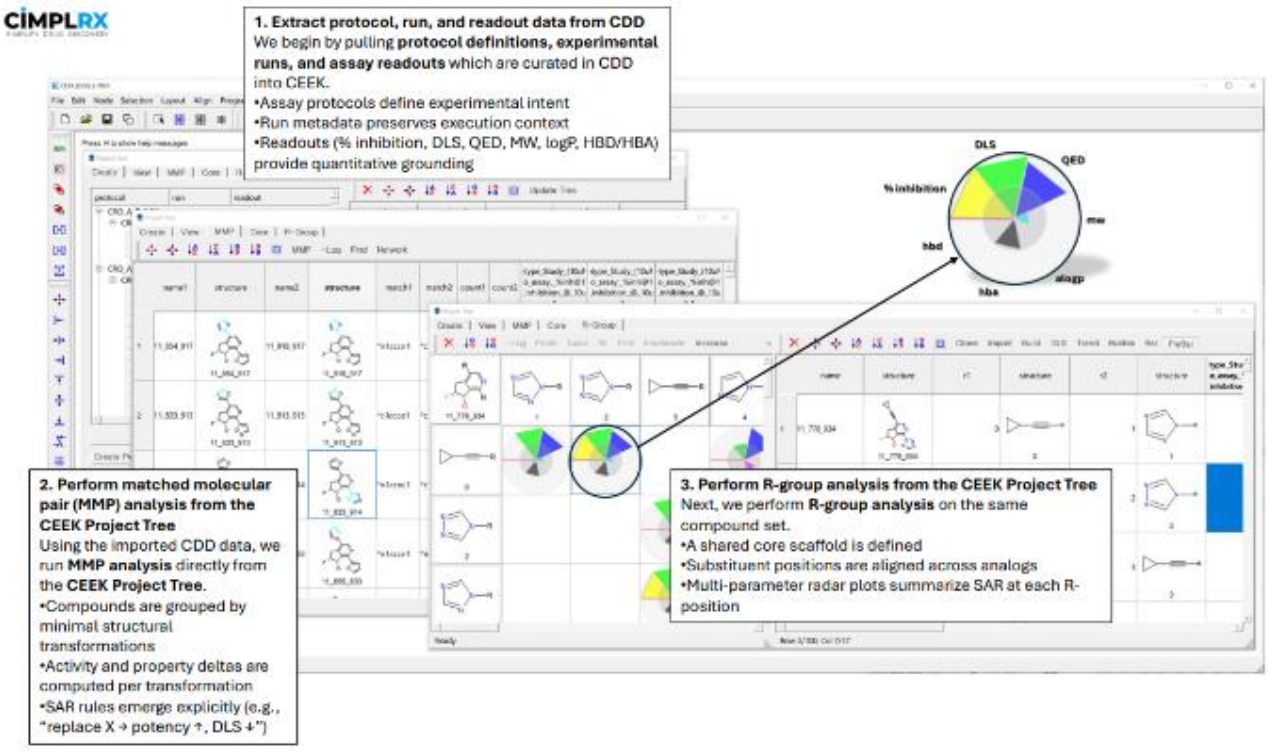
integrated into AI workflows without losing context or traceability.

Through the Vault API, CIMPLRX can pull compound and assay data directly into its workspace environment. “We can retrieve compound and assay data seamlessly from CDD Vault, get it into the right structure, and then incorporate it into analytical workflows,” Cho said. The team has built tools such as compound trees and project trees to organize this data. “With the API we can collect compound information, batch data, and assay results, as required. Then we format that information so our applications can understand, use, and analyze it all efficiently.”

Interconnected Knowledge Graphs and Workflow Tracking

Importantly, the CIMPLRX platform is designed to capture not only experimental data but also the scientific decision-making process. It tracks how scientists interact with data and tools, converting these interactions into structured inputs for AI systems, enabling the systems to model the interconnected environment in which discovery happens. So, rather than treating steps as isolated, CIMPLRX records them as part of an interconnected knowledge graph.

“When you work in drug discovery, you go through many steps—getting data, analyzing compounds, running searches, generating new ideas,” Cho said. “Unlike other platforms that don't capture that workflow, we capture and track these workflows and processes.” Each



interaction becomes part of the evolving graph. “Rather than acting as a static image, each node is an application with which we can interact. Each of those nodes becomes part of the graph as you work,” Cho explained.

This architecture allows CIMPLRX to document workflows, decisions, and how compounds evolve over time, creating a comprehensive scientific landscape where both researchers and computational models can derive actionable insights. Once data is structured within the workspace, scientists can perform analyses commonly used in medicinal chemistry, including matched molecular pair analysis, fragment analysis, and R-group analysis, alongside structure-based modeling, docking, and dynamic simulations to assess protein interactions.

A Modern System of Record

CDD Vault plays a central role in this ecosystem. It acts as the authoritative system of record for compound and experimental data, enabling structured extraction of both empirical and process data. This supports advanced computational analyses within CIMPLRX’s platform and LLM environment, where not only results but also reasoning and connections between analyses can be traced.

Cho emphasizes that CDD Vault is far more than an ELN or LIMS. It provides structured, curated, experimentally grounded data that powers AI-driven drug discovery. Its data model preserves assay intent, execution context, and SAR, making it inherently suitable for AI integration. Using the Vault, CIMPLRX maintains secure data storage with traceable access while enabling flexible downstream analysis. “CDD is a key enabler, because that's where all our private data is stored, and we need to extract information and structure it in the right way so that our LLM can analyze the data, and return results that the medicinal chemist can use,” Cho said.

Advancing to Workspace-Augmented Generation (WAG)

This structured approach allows AI outputs to be returned in formats such as JSON, which can then be translated into chemically meaningful representations like SMILES strings. “Instead of just text output that you might get using ChatGPT, for example, the results become something a medicinal chemist can actually use.”

The workflow begins with data extraction from CDD Vault, including assay protocols, metadata, and quantitative results. Crucially, experimental intent, execution context, and SAR are preserved. Using this curated data, analyses such as matched molecular pair and R-group analysis are performed to extract SAR trends. Results are returned as both structured workspace objects and textual explanations. Because the LLM is workspace-aware, it understands molecular structures and relationships within the broader context, rather than reasoning in isolation.

CIMPLRX has extended the concept of retrieval-augmented generation (RAG) into what it calls workspace-augmented generation (WAG). While RAG retrieves external information, WAG injects the entire workspace—containing up-to-date compound structures, analyses, workflows, and project status—directly into the model’s context. “What we propose is

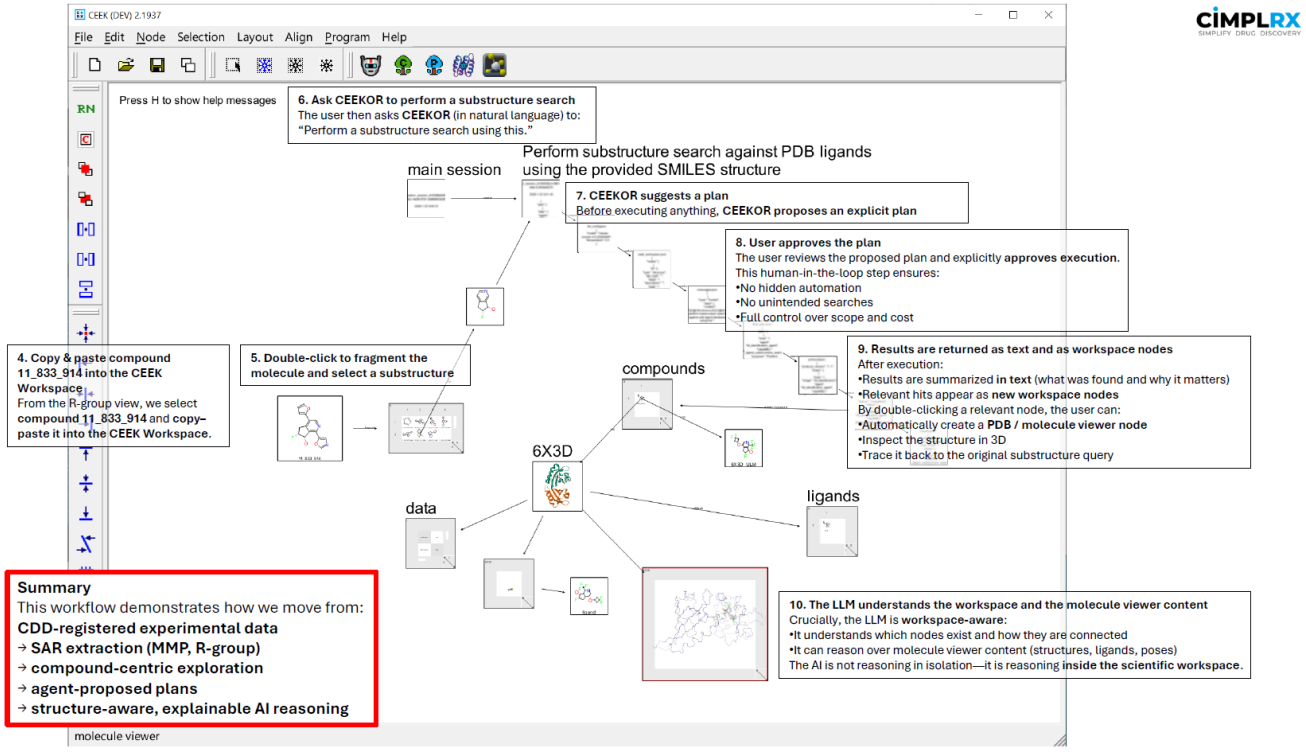
workspace-augmented generation—WAG," Cho said. "Instead of just retrieving documents or data, we inject the workspace itself into the model's context… The LLM understands our environment and uses intent."

Better Science Together

This approach enables more intelligent and targeted compound design. In one oncology collaboration, CIMPLRX demonstrated a dramatic improvement over traditional screening methods. "They screened over a thousand compounds from their internal libraries and had a 0.1% hit rate. In contrast, we hand-picked about 36 compounds and achieved around a 30% hit rate. We may need to test only 10 or 20 or so compounds and still get meaningful results," Cho said.

While this ability to better understand and apply collective data will help to streamline the drug discovery process, CIMPLRX's goal is not simply to make drug discovery faster or cheaper. Instead, the focus is on improving scientific decision-making and increasing the likelihood of downstream success. "It's not really about saving money or resources," Cho explained. "The real question is how far a compound can go. If you design the right compound—safe and effective—it is far more likely to progress further—and potentially all the way—through development."

Currently, CIMPLRX is a team of eight, including seven scientists. Its laboratory work is focused on internal projects, particularly the neuropathic pain program. "For our neuropathic pain project we are in the later stage of lead optimization, so it's possible that by the end of this year [2026], we may be ready to start preclinical studies," Cho said. The company remains flexible in its approach to partnerships. "Depending on the nature of the project and how we set up the collaboration,



we might consider expanding the scope of work with our partner," Cho noted.

CIMPLRX's philosophy centers on combining structured data, human expertise, and AI reasoning. By using CDD Vault as a reliable data foundation and building a workflow-centric discovery platform around it, the company supports both internal programs and global collaborations. "We're a small team," Cho said, "but we want to show what's possible when you connect the data, the workflow, and the science together."

About Collaborative Drug Discovery

Collaborative Drug Discovery provides a modern approach to drug discovery informatics that is trusted globally by thousands of leading researchers. Our CDD Vault is a hosted informatics platform that securely manages both private and external biological and chemical data. It provides core functionality including chemical registration, structure-activity relationship, inventory, visualization, and electronic lab notebook capabilities. For more information, visit us at www.collaborativedrug.com.