Closing the Data Gap: A Pre-Competitive Alliance for AI-Ready Drug Discovery Data

Rick Dixon, Aishwarya Pathak, Kameron Wittmeyer, Nadya Urakova, Anna Chernysheva, Oleg Stroganov, Tania Khasanova, Alena Fedarovich

Rancho Biosciences, PO Box 7208, Rancho Santa Fe, CA 92067

Background

Despite significant advances in computational tools, drug discovery teams continue to face critical data gaps that existing platforms fail to address at the pace required by modern R&D. Predicting differential clinical outcomes, identifying the most promising therapeutic modality for a given target class, linking experimental data to drug exposure, and anticipating tissue-specific or off-target toxicities remain persistent challenges, in part because much of the needed data sits in regulatory documents and must be curated before it can be used. The Drug Data Alliance (DDA) addresses a critical need in drug discovery: consistent, harmonized, integrated and analysis-ready data.

Drug data alliance

Vision: Consistent, Harmonized, Integrated, Analysis-ready Data

Through a client-guided, pre-competitive membership model spanning small molecule and/or biologics tracks, DDA delivers expertly curated data on compounds, large molecules, targets, and indications integrated with pre-clinical and clinical PK, toxicity, and efficacy data at the study and experiment level from authoritative sources such as FDA regulatory approval documents.

Expanding a Trusted Data Foundation

DDA builds on Rancho's decades of expertise in extracting, harmonizing, and integrating biomedical information and years of experience curating drug-development data rooted in its NIH NCATS partnership that resulted in creation of Inxight Drugs[1,2] - a curated portal with extensive compound descriptors including toxicity, pharmacokinetics, and drug-drug interactions. DDA extends the trusted sources foundation to create a broader, deeper evidence landscape spanning emerging therapeutic modalities, PK/PD, efficacy, safety, and real-world data.

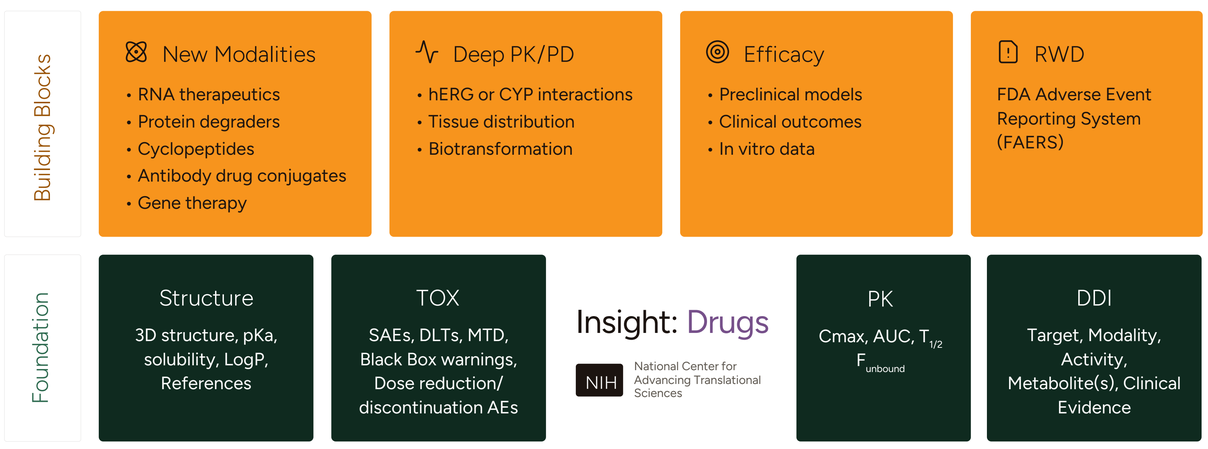


Figure 1. Expanding the Data Landscape. The established foundation (green) provides a common core of curated drug, target, safety, PK, and interaction data. DDA expansion areas (orange) add complementary evidence across new therapeutic modalities, creating a more connected view of therapeutics across preclinical and clinical development.

Member-Driven Priorities, Brought In-House

- Member-driven priorities: Members help define the highest-value data gaps, modalities, and scientific questions.
- Expert Curation: Rancho curates, harmonizes, and integrates prioritized evidence into a consistent, analysis-ready resource.
- Built for your environment: Bring the resulting data in-house, combine it with proprietary data, and use it across existing research and analytics workflows.
- A growing data asset, not another portal: Broad data rights enable members to incorporate curated content directly into their research ecosystem.

Methodology: Scaling Expert Curation with AI-Enabled Pipeline

- Expanding DDA across new modalities and evidence types requires a curation approach that can scale with growing data volume and complexity.
- AI can accelerate labor-intensive activities such as information extraction, structuring, and quality checking, allowing larger volumes of evidence to be processed more efficiently.
- The goal is not to replace expert curation; it is to embed Rancho's established curation methods, standards, and quality controls into AI-enabled workflows.

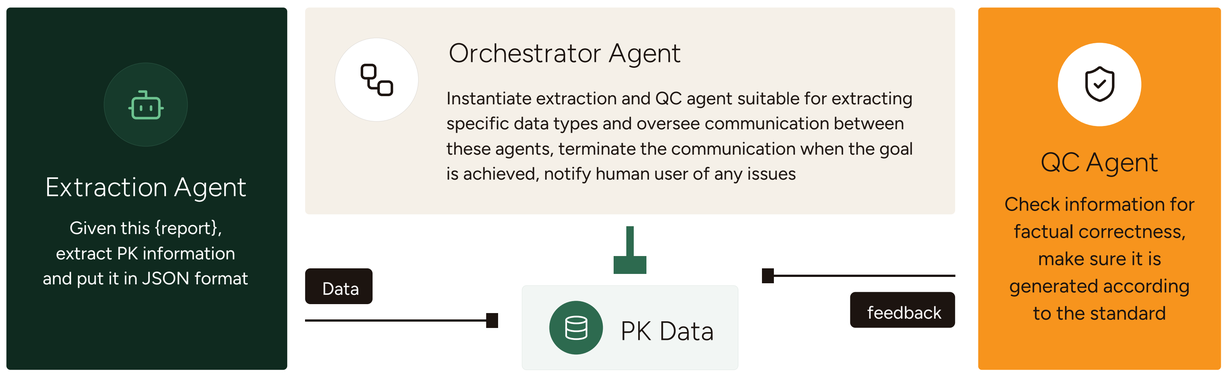


Figure 2. A high-level view of Rancho's AI-enabled curation workflow. Specialized AI components divide the work across extraction, coordination, and independent quality control rather than relying on a single AI-generated output. QC findings are fed back through the workflow for correction, while expert review and validated curation standards provide additional quality safeguards.

Ensuring Quality in AI-Enabled Curation

- Gold-standard foundation: Rancho's expertly curated datasets provide trusted reference data for developing, testing, and benchmarking AI-enabled curation workflows.
- Human-in-the-loop expertise: Subject-matter experts (SME) review outputs, resolve ambiguous and complex cases, and feed corrections back into the workflow to improve performance over time.
- Independent QC & validation: Dedicated quality checks provide an additional layer of validation, supporting accuracy, consistency, and traceability across the resulting data.
- Continuous improvement: SME feedback and QC findings are used to refine instructions, rules, and validation criteria as data sources, requirements, and workflows evolve.

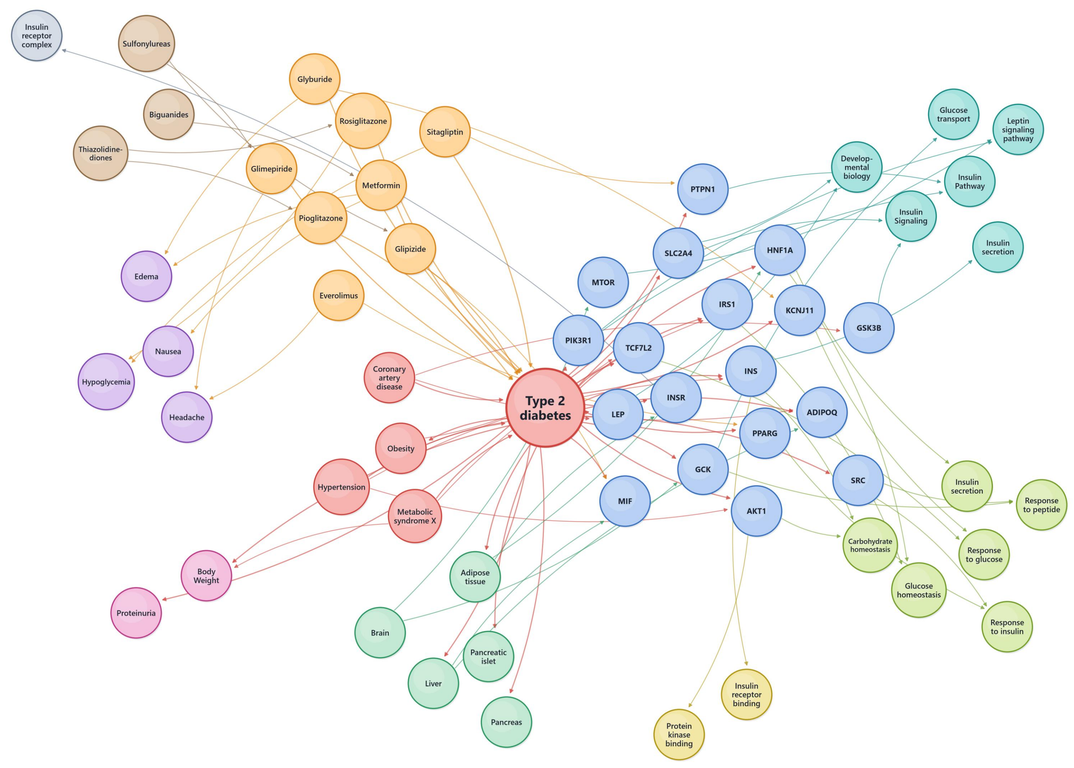
Value: AI-Ready, Experiment-Level Data

Potential Directions for Consumption Layer

DDA business model offers a scalable, cost-effective pathway for organizations to bring high-value, decision-ready drug experiment data in-house and to accelerate decisions across early- and late-stage discovery. We deliver experiment-level data you can use anywhere.

"Plug-and-play" content is delivered in flexible, ingestion-ready formats suitable for direct API integration into internal databases, AI/ML models, and knowledge graphs. Members retain perpetual rights to the data, which is continuously updated as new approvals are published.

- Data accessibility and formatting (API's)
  Connection to internal infrastructure and platforms
- AI/ML/Knowledge Graph-Ready
- Maintenance



Potential Applications Areas - More Value

Content applications include predictive safety, pharmacokinetic modeling (dosing, exposure, and drug-drug interaction prediction), biomarker validation (leveraging target and off-target data for companion biomarker selection), and personalized medicine (identifying optimal treatments based on specific target alterations).

Predictive Safety	Overlay known adverse events to look for patterns and predict AE in the clinic	Clinically relevant AEs linked to drug concentrations
AI-Ready	Provide content for AI/ML tools: training sets, features	Curated content delivered in a format consumable by AI and ML tools
Pharmacokinetics	Answer clinical questions like "what doses are safe in patients?" "How does concentration relate to dose?"	PK Properties overlay onto drug interactions to predict Drug-Drug Interactions
Biomarkers	Target and off-target information helps with selection of companion biomarkers	Validation of biomarker hypothesis
Personalized Medicine	Identify drugs most suitable for treating patients with a specific target alteration	Deep annotations for compounds allow for fine-tuning when selecting treatments

Investigational Drug Value Add Stories

Find targets that need a drug, find drugs that need a target

mGlu Antagonist Pomaglumated & Related Compounds:
- Studied in Early 2000's for treatment of Schizophrenia and psychotic disorders
- Discontinued due to lack of efficacy
- No FDA approved mGlu-targeting drugs to date

BACE1 Inhibitors:
- Investigated for treatment of Alzheimer's disease by Merck, AZ and Eli Lilly
- Halted in 2010s due to lack of efficacy
- No BACE1 inhibitors are FDA approved to date.

ALK inhibitors:
- >10 ALK inhibitors in CTs for the treatment of ALK+ lung cancer
- Many of these "me-too" compounds were discontinued for economic reasons (5 drugs approved by FDA for these indications)
- Compounds have different selectivity, some were successfully re-purposed: Entrectinib in addition to ALK inhibits TrkA, B, C kinases and is approved by FDA for TRK-fusion driven cancers

References

1. Siramshetty VB, Grishagin I, Nguyen D, Peryea T, Skovpen Y, Stroganov O, Katzel D, Sheils T, Jadhav A, Mathé EA, Southall NT (2022) NCATS Inxight Drugs: a comprehensive and curated portal for translational research [https://doi.org/10.1093/nar/gkab918]
2. https://drugs.ncats.io/