PEER REVIEWED

AI-Augmented Innovation in Drug Discovery

Innovators in pharma/biopharma seek to implement systems that support the discovery, development, and commercial launch of new products. Of particular interest are systems to support the implementation of continuous and mutually reinforcing digital-physical feedback loops. Here, digital tools and methods enhance physical processes, and feedback from these improved physical processes informs progressive digital advancements.

In traditional, non-digitalised drug discovery Design, Make, Test, and Analyse (DMTA) cycles, each transition between stages often demands substantial human effort to transpose and translate information, bridging disparate systems and domain-specific knowledge. Inefficient management of these transitions can result in productivity loss, as practitioners must frequently consult subject matter experts to translate critical, context-dependent information from design platforms to execution and analysis systems. This reliance on manual processes also increases the risk of transposition or transcription errors, where inaccurate transfer of numerical or textual data into digital interfaces may lead to failed experiments, flawed interpretations, or misguided decisions.

Introducing modern AI-powered tools into the DMTA workflow not only streamlines these transitions but also enhances the analysis phase: advanced algorithms can rapidly process experimental data; uncover patterns that might escape human notice; and generate actionable insights. To ensure that learnings are preserved and accessible for future cycles, results from such AI-enabled analyses should be systematically documented within integrated digital repositories, allowing teams to memorialise findings, trace decisions, and enable continual refinement of the DMTA cycle. By minimising manual intervention and harnessing AI's analytical capabilities, organisations foster a virtuous, resilient DMTA loop that seamlessly connects digital and physical domains.

Each transition within the non-digitalised DMTA cycle (from design to make, from make to test, etc.) often requires significant human transposition and translation of information from systems to bridge the gap between different stages and domains of expertise. Put simply, avoiding these risks across the various DMTA transitions and from digital-to-physical steps allows for a virtuous DMTA cycle.

Innovative DMTA Cycles in Drug Discovery: AI for Prediction & Orchestration

There are a variety of specific DMTA cyclesrequired for successful candidate nomination in drug discovery organisations.[1] The following summarises the dual-purpose of DMTA cycles during lead optimisation in drug discovery.

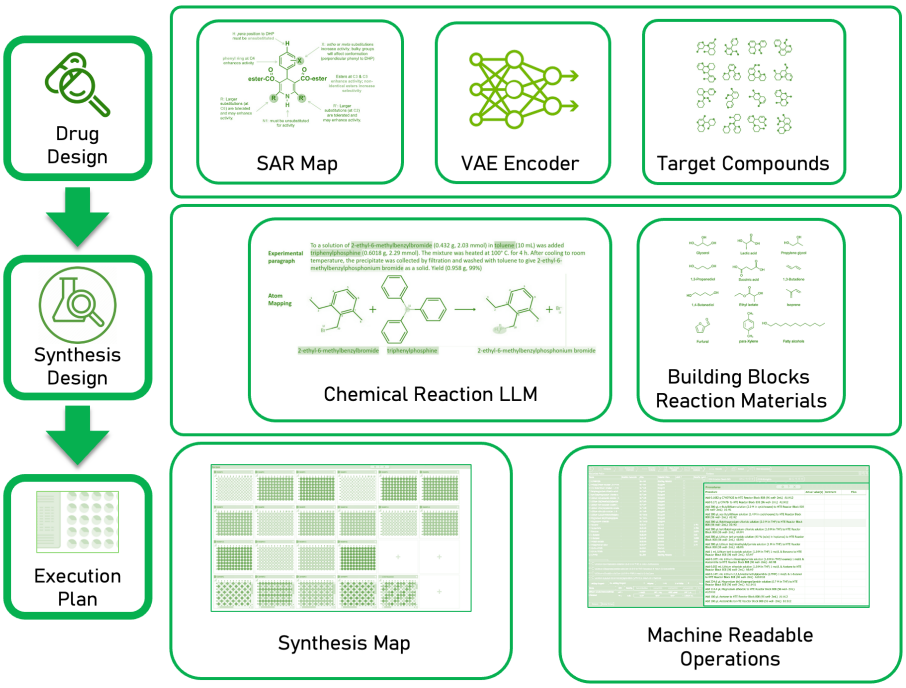

Figure 1. Machine-enabled virtuous design cycle in drug discovery. The drug design step involves the use of a structure-activity relationship (SAR) Map (2) and a variational autoencoder to generate a set of target compounds. The synthesis design step involves the use of chemical reaction large language models (LLMs),[3] retrosynthesis prediction tools, and material inventory web services to generate a set of reactions. Finally, the use of specialised planning applications supports the generation of a synthesis map with machine readable operations to support the transition to digitally supported and automated execution (make, in DMTA).

AI-Enabled Drug Design – The Design Step

The design phase of the DMTA Cycle in drug discovery addresses two key questions:

What to Make?

Virtuous approaches involve the use of generative AI tools to produce a structure-activity relationship (SAR) map. As described in Figure 1, AI components can serve to recommend specific chemical structures and sequences which should exhibit suitable physicochemical and pharmacological properties for corresponding patient populations. The resultant recommendations are tested via execution of confirmatory assays using physical composition of matter whose identity matches the AI-recommended structures.

After gathering enough SAR data on a lead series, medicinal chemists focus on optimising potency, selectivity, and druggability. Modern drug discovery organisations have endeavoured to implement a range of well-trained generative AI systems which produce reliably accurate sets of target compounds.[4,5,6]

How to Make it?

To confirm the success of sequential rounds of these generative AI outputs, medicinal chemists must design efficient synthetic routes for each of the target compounds. As described in the figure below, chemists will perform appropriate DMTA steps to ultimately produce test articles supporting confirmatory assay execution.

AI-Enabled Synthesis Execution in Drug Discovery – The Make Step

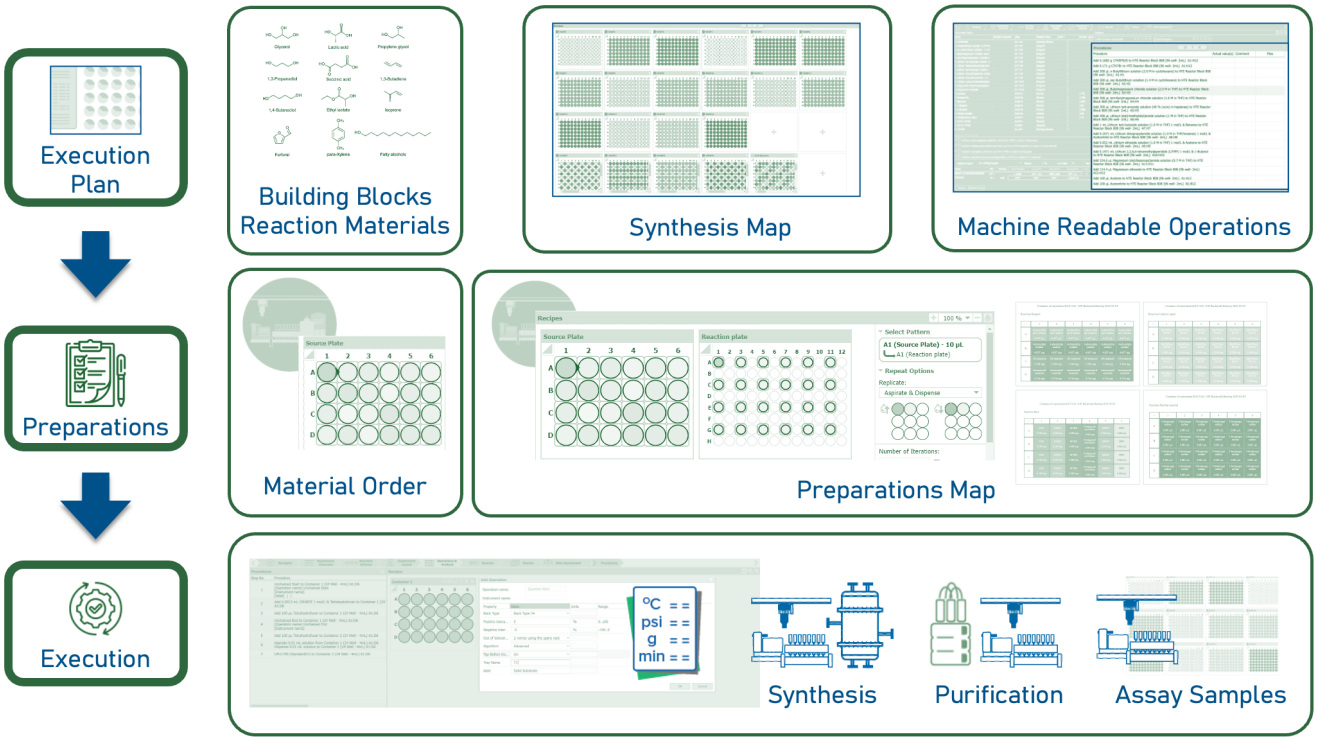

Figure 2. Machine-enabled virtuous "make" cycle in drug discovery. Leveraging the synthesis map and machine-readable instructions included in the execution plan, building blocks and other reaction materials are ordered and delivered in a reaction-friendly source plate format. The preparations map and corresponding operations are sent to synthesis, purification, and assay samples for execution. The digital representation of all assay samples (including applicable sample identifiers and related metadata) enables automated transition to the testing step.

Based on the similarity within each set of recommended structures, chemists seek to minimise the number of discrete planning and execution operations to produce test articles for the entire set. Consequently, chemists will identify common starting materials which can be used for execution of divergent-then-parallel synthesis operations. This results in a smaller total number of reactions required to produce confirmatory test articles for all AI-recommended target compounds.

Achieving an ideal number of minimum operations for parallel test article production involves the use of well-trained retrosynthesis tools.[7]

Furthermore, medicinal chemists use the recommendations from these retrosynthesis tools to map to experiment execution systems, including to applications which generate machine-friendly master procedure lists.[8] These instructions include machine-encoded material dispenses, reaction operations, and sample preparation operations. These machine-executed operations enable far more efficient preparation and execution of Make tasks required for the preparation of test articles (when compared to manual operations).

Upon completed design, medicinal chemists can execute each unit operation, from material dispensing, through reaction initiation, to final workup and sample preparation (Figure 2). In modern synthesis labs, the format of machine instructions must conform to the format requirements of each instrument supporting each unit operation.

The output results in a set of test articles, labelled with appropriate machine-and-human-legible material identifiers. These identifiers serve to associate test article information stored in appropriate software applications including output material identity, material metadata, container or vessel IDs.

Similarly to design, the testing of output materials (i.e., purified target compounds) serves a dual purpose:

AI-Enabled Testing in Drug Discovery—The Test Step

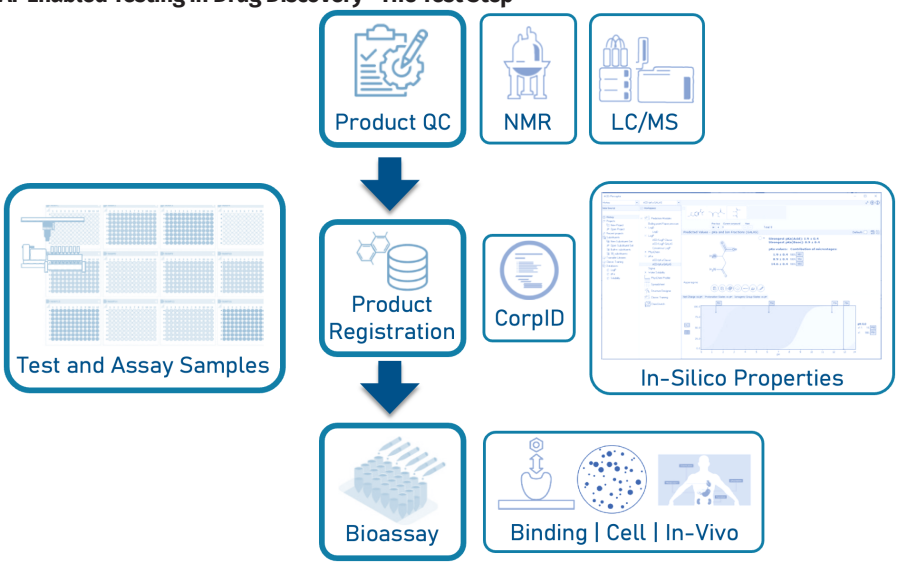

Figure 3. Machine-enabled virtuous "Test" cycle in drug discovery. For both product QC and bioassay, physical samples are labelled with human and machine-readable sample identifiers. Product registration generates a corporate identifier for each product structure. All assay results are related to these corporate identifiers (via the relationship between sample identifiers and corresponding assay results).

1. Subjecting materials to a wide range of project-applicable bioassays to confirm the performance of the target compounds as possible (future) drugs.

2. Identity and quantitative compositional testing to assure accurate SAR. [9]

As such, the modern DMTA cycle accounts for all applicable tests required to inform the overall lead optimisation effort. Samples of output materials from reaction execution steps are prepared, then labelled with system-derived sample identifiers, including:

- **Reaction quality control samples –** sample type (pre-process, in-process, and post-process), sample preparation information.
- **Product registration samples –** including chemical identity, purity, salt stoichiometry, and any applicable physical form information that is required for assessing bioassay results for SAR map.
- **Test article samples –** including sample concentration, assay role (e.g., control, standard, blank, replicate number, etc.),
- **Predicted properties –** intended to complement the measured bioassay and quality control (QC) results, enhancing the utility of the overall SAR Model:

Physicochemical	ADME	Toxicity
• Aqueous Solubility • Boiling Point/Vapor Pressure • Log*D* • Log*P* • pK_a • Sigma	• Blood Brain Barrier Permeation • Cytochrome P450 Inhibitors • Cytochrome P450 Substrates • Distribution • Maximum Recommended Daily Dose • Oral Bioavailability • Passive Absorption • P-gp Specificity • PK Explorer • Regioselectivity of Metabolism	• Acute Toxicity • Aquatic Toxicity • Endocrine System Disruption • Mutagenicity • Health Effects • hERG Inhibition • Irritation

Figure 4. Customary physicochemical, absorption, distribution, metabolism, and excretion (ADME), and toxicological descriptors generated upon product registration.

The first analysis activities support the processing and interpretation of product QC and bioassay data. The variety of tools that support this primary analysis step has been described elsewhere.[10,11,12]

Modern drug discovery organisations intend to leverage the output of these analyses, by:

1. Aggregating processed data into a data warehouse,
2. Implementing a rigorously enforced controlled vocabulary for all bioassay and product QC Results, as well as associated metadata structured as JavaScript object notation (JSON) objects,
3. Relating the test results by applicable sample, corporate identifier, and in-silico properties (generated for product registration).

Scientists can then update applicable SAR maps (Figure 1) based on the bioassay test results.

Value Realisation of AI & Digitalised DMTA

Implementing the modern DMTA cycle can improve productivity in drug discovery by supporting the generation of new target compounds and enabling an efficient, iterative process. The use of structured data, controlled vocabularies, and advanced analysis digital tools streamlines the drug discovery workflow and may accelerate the identification of potential therapeutics.

AI-Enabled Analysis in Drug Discovery – The Analyse Step

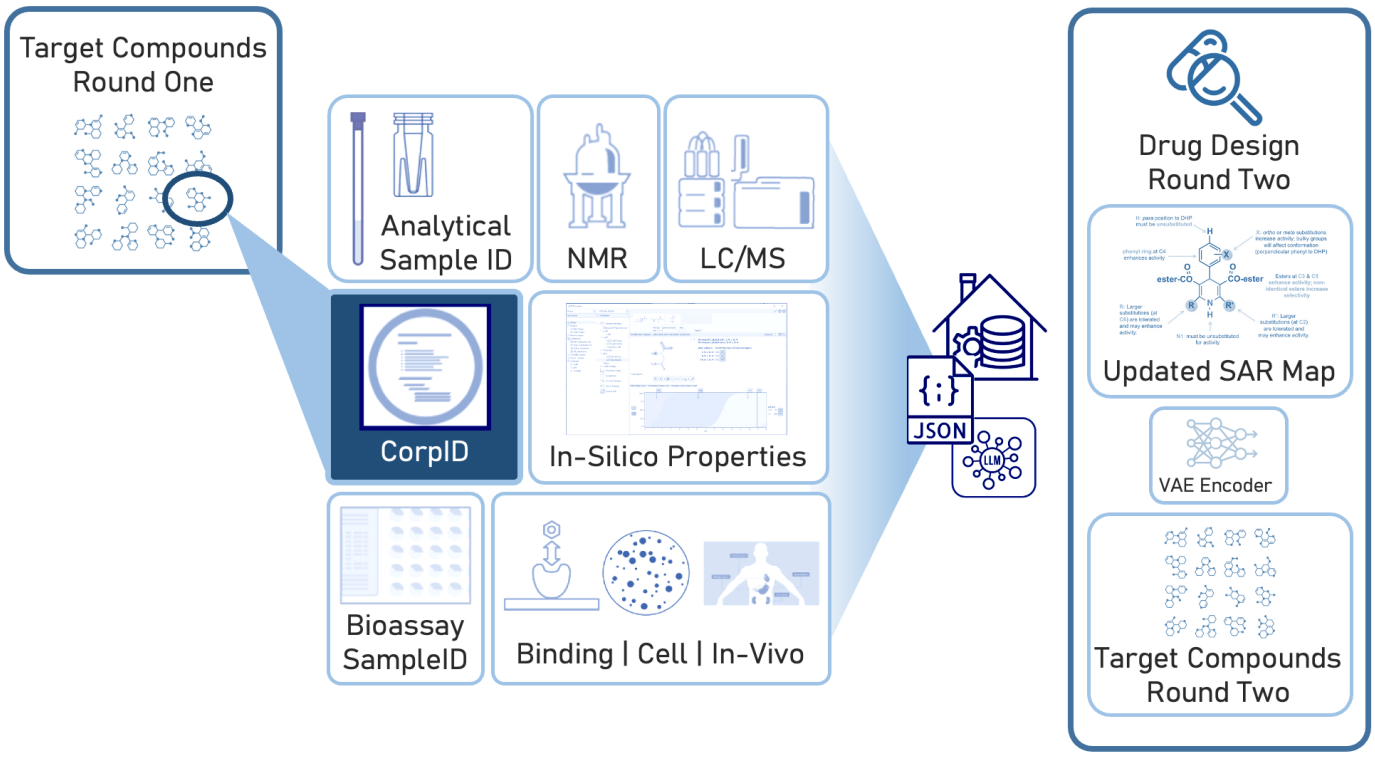

Figure 5. Upon completion of assay data analysis, results as JavaScript object notation (JSON) objects are used to re-train Generative AI Models, allowing for a virtuous next round of DMTA to start.

Furthermore, the use of AI-driven tools now plays a transformative role in the swift construction and refinement of predictive models based on structure-activity relationships (SAR). By leveraging expansive datasets that capture both molecular structures and their corresponding biological activities, these tools employ machine learning algorithms capable of discerning complex, non-linear relationships within the data. Generative AI further augments this process by proposing novel chemical scaffolds tailored to desired activity profiles, while continuous feedback from experimental results enables iterative model retraining, ensuring accuracy and relevance. This rapid, automated integration of SAR insights not only accelerates hypothesis generation and compound prioritisation but also enhances the reliability of chemical structure predictions – empowering scientists to select and optimise candidates with unprecedented precision and speed.

This approach allows project teams to optimise lead series more rapidly and identify molecules that satisfy clinical candidacy requirements.

Conclusion

Ultimately, the integration of AI into drug discovery projects markedly elevates both project success rates and operational efficiency. Now, organisations can more accurately identify promising drug candidates earlier and faster – effectively prioritising compounds that exhibit optimal safety and efficacy profiles. This targeted approach minimises downstream failures and leads to a significant reduction in the attrition rates that historically plague clinical development pipelines – a critical key performance indicator for the industry. With AI continuously refining predictive models through rapid analysis of diverse datasets and real-time incorporation of experimental feedback, resources are allocated more judiciously, timelines are shortened, and costly late-stage setbacks are avoided. As a result, drug discovery teams are empowered to advance higher-quality candidates into clinical trials with greater confidence, thereby increasing the likelihood of regulatory approval and market success.

REFERENCES

1. DMTA Cycles include disease pathology research, biological target and pathway analysis, molecular modality selection, in-vitro assay development, in-vivo study design, and test article scale-up research. Certainly, all require effective DMTA cycles to matriculate through to clinical studies.
2. Frey, K. (2020). Structure activity relationship (SAR) maps. Curr. Pharm. Teach. Learn., 12(3), 339-346. https://doi.org/10.1016/j.cptl.2019.12.014
3. Vangala, S.R., *et al.* (2024). Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature. J. Cheminform., 16, 131. https://doi.org/10.1186/s13321-024-00928-8
4. S. Reddy N, *et al.* (2024, June 30). Leveraging Latent Evolutionary Optimization for Targeted Molecule Generation. 2024 IEEE Congress on Evolutionary Computation (CEC), Yokohama, Japan. https://ieeexplore.ieee.org/document/10611790
5. Ochiai, T., *et al.* (2023). Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity. Commun Chem, 6, 249. https://www.nature.com/articles/s42004-023-01054-6
6. M. C. Ramos, *et al.* (2025). A Review of Large Language Models and Autonomous Agents in Chemistry. Chem. Sci., 16, 2514-2572. https://doi.org/10.1039/D4SC03921A
7. Discovery at Your Fingertips | SYNTHIA™ Retrosynthesis Software. (2025, April 16). Synthia Online. https://www.synthiaonline.com/
8. Software to support HT experimentation workflows. Katalyst D2D, version 2024, Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada. www.acdlabs.com/katalystd2d
9. It should be noted that organisations will also subject output materials to such analyses in defence of composition of matter claims in IP Applications. Basics of claim drafting for utility patent applications. (Retrieved 2025, Apr. 22). USPTO. https://www.uspto.gov/sites/default/files/documents/ventionCon2021WhatsinaPatentClaimWorkshopFinalstakeholders.pdf
10. Mercer, K. (2023, Oct. 24). Enabling End-User HTE at GSK [Online presentation]. ACD/Labs Driving Efficiency with Spectrus Symposium. https://www.acdlabs.com/resource/enabling-end-user-hte-at-gsk/
11. Standardization of Analytical Data: Best Practices. (2025, Jan. 6). ACD/Labs. https://www.acdlabs.com/resource/standardization-of-analytical-data-best-practices/
12. FAIR Annotation of Bioassay Metadata. (Retrieved 2025, Apr. 22). Pistoia Alliance. https://fairtoolkit.pistoiaalliance.org/use-cases/fair-annotation-of-bioassay-metadata/

Andrew Anderson

Andrew Anderson, Vice President, Innovation and Informatics Strategy, ACD/Labs currently leads product and technology management efforts as VP of Innovation and Informatics Strategy at ACD/Labs. He previously worked in Technology Scouting at PepsiCo, and has served in a variety of Sales Management, Business Development, and Strategic Partnership roles in the Scientific Software industry. Andrew earned a BSc in Chemistry and MBA from San Diego State University and started his professional career as an Analytical Chemist and Regulatory CMC Specialist at Pfizer.