DATA INTEGRATION, MODELING, AND AUTOMATION

Goal | **Breakthough Capability** | Milestone

Establish a computational infrastructure where easy access to data supports the DBTL process for biology.

Established standard and accessible repositories for biomanufacturing data and analysis methods.

Have developed a system of robust communication between academia and industry surrounding engineering biology data access and needs.	Biomanufacturing-specific data standards and repositories.
Develop findable, accessible, interoperable, and reusable (FAIR) data standards and open repositories for engineering biology.	

Common computational infrastructure for finding biological data and common APIs for search and analysis.

Demonstrate common data search and interchange among current biological and chemical repositories and existing microbial biofabrications.	Produce a common library of open design tools for more open medical and agricultural environments.
Produce a common library of open design tools, built upon standard APIs, and supported by portable/virtualized execution environments.	

End-to-end, industry-normed design software platforms for engineered biological systems.

Develop industry-accepted, sharable assessments of current data tools and uses in reducing cost and increasing reliability of executing the DBTL cycle.	Create an industry-accepted, open-source or publically-accessible version of industrially-relevant DBTL software and data.

Establish functional prediction through biological engineering design at the biomolecular, cellular, and consortium scale.

Fully-automated molecular design from integrated, large-scale design data frameworks.

Structure- and comparative analysis-based libraries for automated directed evolution, with feedback of large-scale results to algorithms.	Automated designs for integrated manufacturing to enable more successful, iterated workflows.	Use of large-scale design data in integrated frameworks.	Design and integration of thousands of critical catalytic activities into proteins for a set of model hosts and creation of standard tools for allosteric control of these activities.
	Large-scale design data generation to inform next-generation algorithms for molecular design.		

Use of enzyme promiscuity prediction algorithms to design biosynthetic pathways for any molecule (natural or non-natural).

Retro-biosynthesis software that can identify any biological or biochemical route to any organic molecule.	Data integration for certain classes of enzymes and pathways and predictable host-specific expression in model organisms.	Integrated data that allows on-demand characterization, standardization, insertion, and deployment of natural and non-natural pathways.

Scalable, data-driven host design for complex environments that enable high-level production of natural biomolecules.

Ability to make and screen multiple host mutations for epistasis mapping and synthetic interactions, making large-scale host optimization possible.	Thematic design rules for host system engineering inferred from data.	Data-driven domestication of any new host for new activities in any environment and scale.
	Tools to acquire and transfer data to a novel host to inform both genetic-domestication and prediction and determination of function.	
Better data on physiology and fitness in deployment environments suitable for informing design.	Novel design tools to support host design for more complex, natural (non-laboratory) environments.	

Enabled design of functional, self-supporting ecosystems.

Data-driven tools for selecting organisms for synthetic assemblies to achieve resistant, resilient activity.	Integration of molecular, pathway, and host design to create and build models of genetically-engineered communities that function predictably, in the context of deployment ecology.	Ability to design and build functional, enclosed, self-supporting ecosystems of multiple engineered microbial species for efficient industrial production.
Direct data collection for the most important communities in human, agriculture, and complex bioreactor work sufficient for informing design.		
Modeling tools to identify cross-organismal networks and ecological interactions.		Ability to design, model, and engineer microbial consortia to simultaneously and efficiently produce multiple products of interest with minimal by-products and waste.

Establish optimal manufacturing processes from the unit-operation to the integrated-screening scale.

Standardized informatics tools, data, and automation platforms for efficient and collaborative use and integration of data in order to develop novel products more quickly.

Establish communications and networks to develop democratized platforms for data exchange and automation across industry and academia.	Democratized platform for data exchange related to standard/model microorganisms.	Democratized platform for non-model organisms and microbial communities.	Full machine learning capabilities and ability of algorithms to run greater than 90% of the DBTL+automation cycle.
	Initial development of a non-model organism database to integrate predicted pathways and -omics data for production.	Democratized suite of platforms that can be utilized across different model systems.	

2 Years — **5 Years** — **10 Years** — **20 Years**