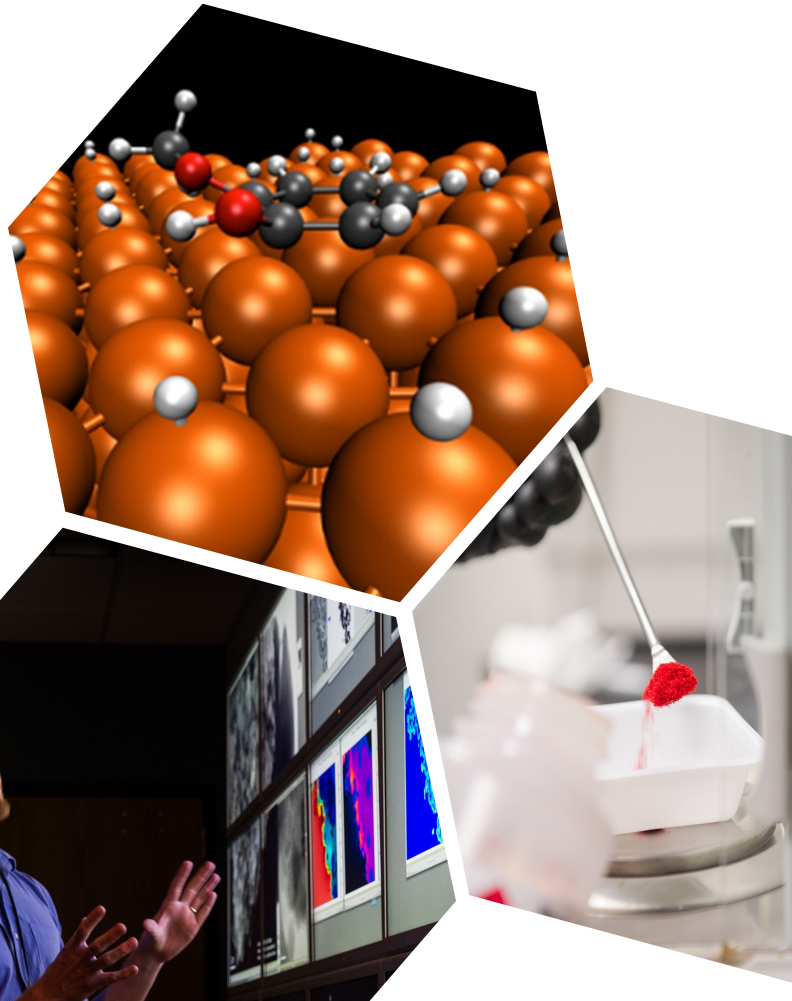
ChemCatBio
Chemical Catalysis for Bioenergy

BETO 2021 Peer Review: ChemCatBio Data Hub (2.6.2.500)

March 9, 2021

Catalytic Upgrading Review Panel

Kurt Van Allsburg, NREL



U.S. DEPARTMENT OF ENERGY | Office of ENERGY EFFICIENCY & RENEWABLE ENERGY
BIOENERGY TECHNOLOGIES OFFICE

The Data Hub Project within ChemCatBio

Integrated and collaborative portfolio of catalytic technologies and enabling capabilities

Catalytic Technologies

- **Catalytic Upgrading of Biochemical Intermediates** (NREL, PNNL, ORNL, LANL)
- **Upgrading of C1 Building Blocks** (NREL)
- **Upgrading of C2 Intermediates** (PNNL, ORNL)
- **Catalytic Fast Pyrolysis** (NREL, PNNL)
- **Electrocatalytic CO2 Utilization** (NREL)

Enabling Capabilities

- **Advanced Catalyst Synthesis and Characterization** (NREL, ANL, ORNL)
- **Consortium for Computational Physics and Chemistry** (ORNL, NREL, PNNL, ANL, NETL)
- **Catalyst Deactivation Mitigation for Biomass Conversion** (PNNL)

Industry Partnerships (Phase II Directed Funding)

- **Opus12** (NREL)
- **Visolis** (PNNL)
- **Sironix** (LANL)

Cross-Cutting Support

- **ChemCatBio Lead Team Support** (NREL)
- **ChemCatBio Data Hub** (NREL)

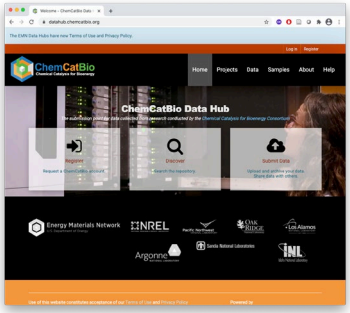
Project Overview

The Data Hub project:

Harnessing data to accelerate catalyst discovery

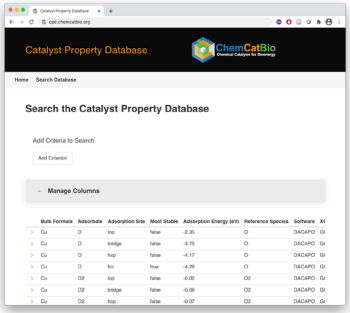
FY18–19: Develop the **Data Hub**, a general data repository

FY20–present: Develop the **Catalyst Property Database**, as part of a shift to focus on transformational tools enabling catalyst R&D



datahub.chemcatbio.org

Released in 2018



cpd.chemcatbio.org

Released in 2020

Project Overview: the Data Hub (FY18–19)

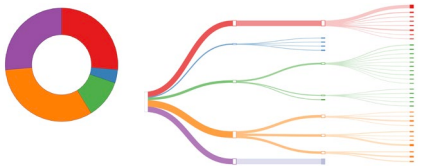
FY18–19: The **Data Hub**

- A requirement for ChemCatBio as an EMN consortium
- A dropbox for all kinds of scientific data, with advanced tools (e.g., visualization of common data types)
- Public and private datasets, curated by ChemCatBio

Data Hub, a framework for transformational tools enabling catalyst R&D

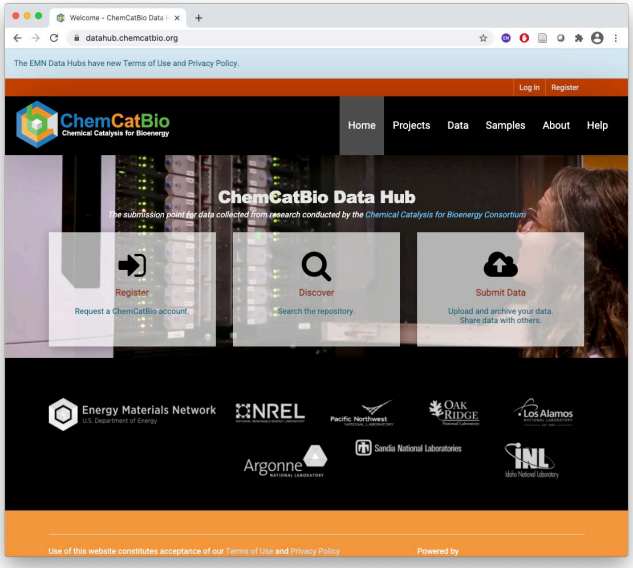
CatCost

(a free catalyst cost estimator, released in 2018)



The Catalyst Property Database

(current focus of this project)



datahub.chemcatbio.org

Released in 2018

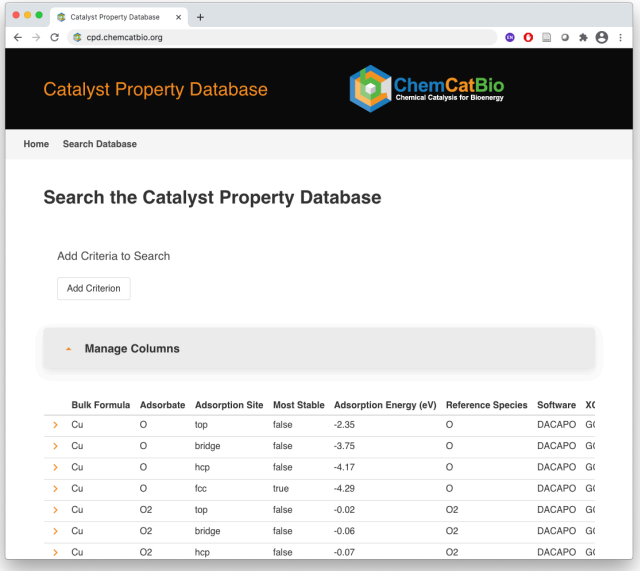
EMN = Energy Materials Network, a group of DOE-funded consortia

Project Overview: the CPD (FY20–present)

FY20–present: The **Catalyst Property Database**, a transformational tool enabling catalyst R&D

- Intended to accelerate materials discovery
- A resource for the entire catalyst community, not only ChemCatBio
- Open to public to view (FY20) and upload (FY21), subject to quality control

Informed by the vision for a **Catalyst Design Engine**

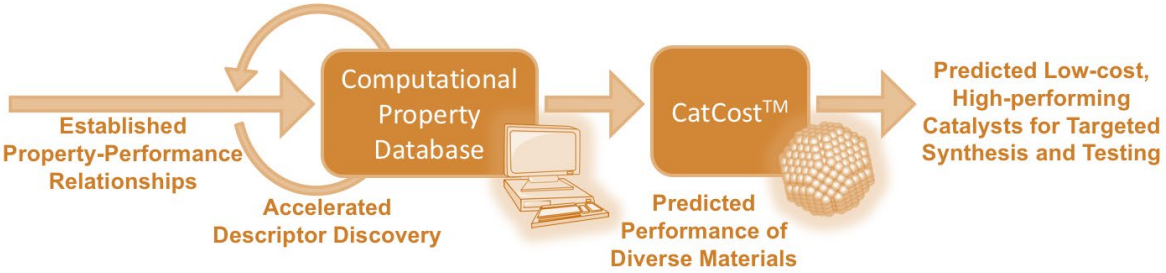


cpd.chemcatbio.org

Released in 2020

The Catalyst Design Engine

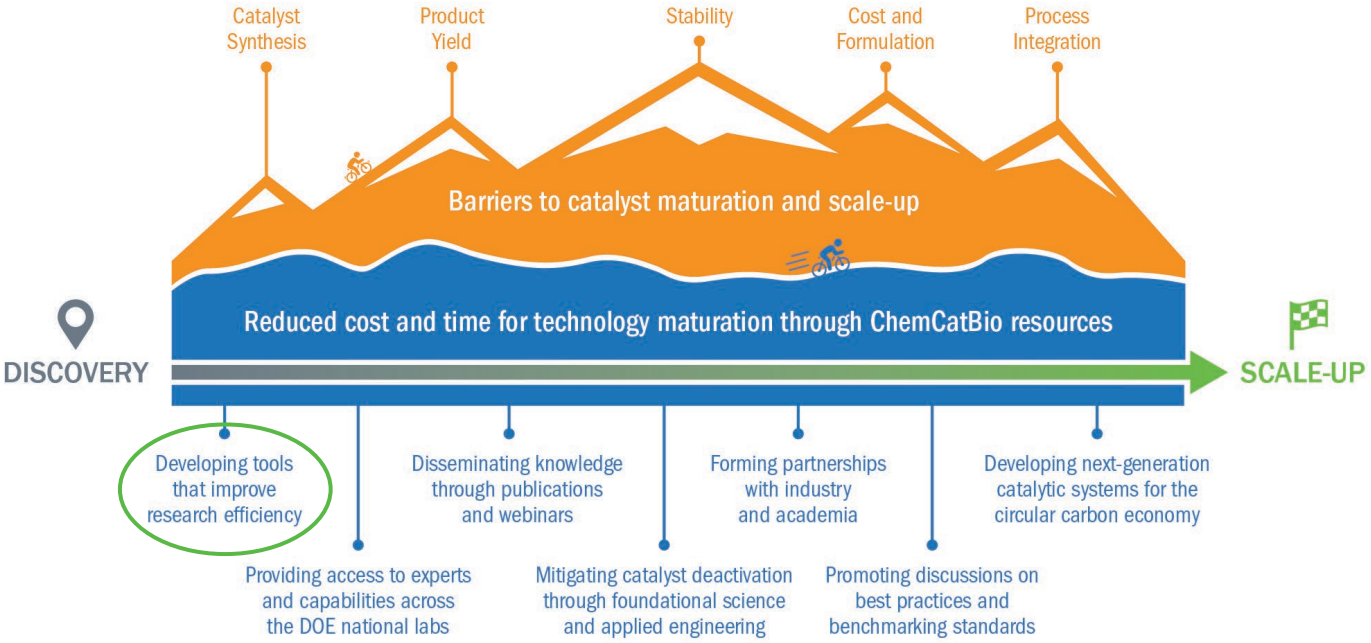
The Catalyst Design Engine is a vision for a free, publicly available tool that combines catalyst performance and cost insight



To work toward this vision, we first need to develop the building blocks: CatCost (released 2018) and the Catalyst Property Database (in development, initial release 2020)

Accelerating Catalyst Discovery

The path to catalyst deployment is slow and difficult.



ChemCatBio is accelerating the catalyst and process development cycle.

Project Overview: Heilmeier Questions

- What are you trying to do?

  Harnessing catalyst property data to accelerate catalyst discovery

- How is it done today and what are the limits?

  When data is used in catalyst discovery, it is **often collected/computed from scratch** because of the challenges in finding **reliable, directly comparable** datasets

- Why is it important?

  It's inefficient.

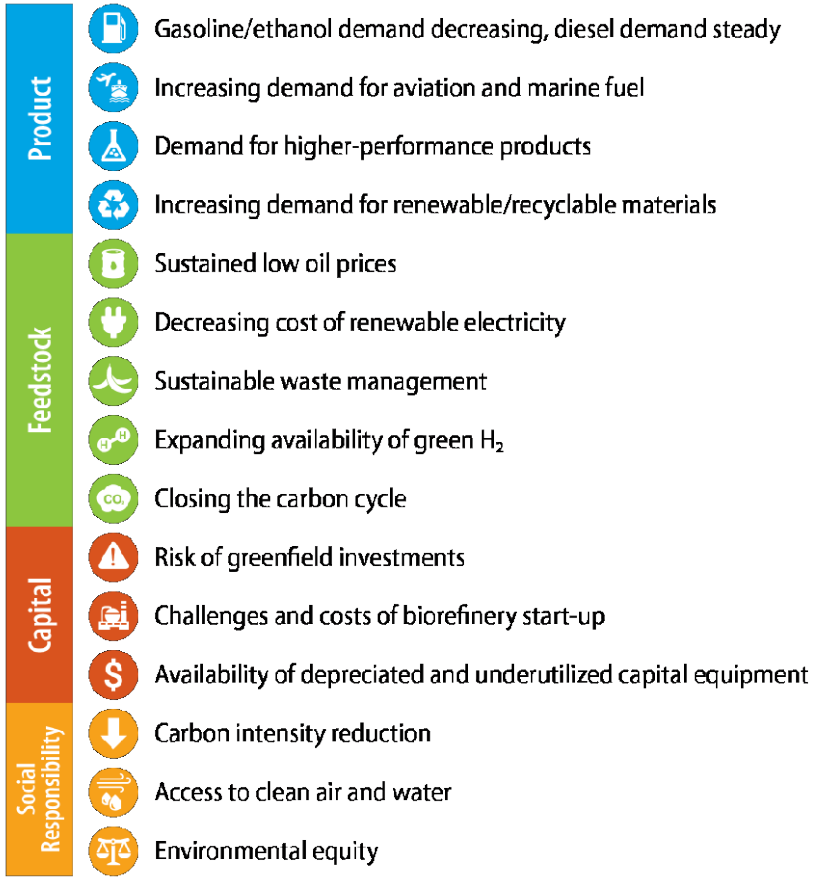
  Current practice results in **redundant calculations** being repeatedly performed

  Furthermore, only a **small subset** of published data is brought to bear

- What are the risks?

  Users find the database irrelevant or too difficult to use

Market Trends Addressed by the Data Hub



As a cross-cutting effort to enable accelerated catalyst discovery, the ChemCatBio Data Hub is most relevant to those market trends calling for new, advanced catalysts

1. Management: Team and Collaborations

Management Plan: A project team with diverse, targeted expertise

Kurt Van Allsburg, Ph.D.

PI

Experimentalist

Experienced developer of R&D tools such as CatCost

Nalinrat Guba, Ph.D.

Lead Developer

Software engineer (previously at Oracle)

Qiyuan Wu, Ph.D.

Experimentalist

Experienced web developer

Carrie Farberow, Ph.D.

Technical expert, former PI

Computational researcher

Sean Tacey, Ph.D.

Technical expert

Computational researcher

Advisors:

Josh Schaidle, CCB Director

Tom King, NREL, UI designer

Nick Wunder, NREL, web dev expert

Kathy Cisar, NREL Communications

Project Tasks:

- Task 1: Data Hub and Computational Catalyst Property Database Development
- Task 2: Data Hub Maintenance, Security, and Oversight

Project Collaborations:

- CCPC – Atomistic Modeling Task, (PI: Carrie Farberow)
- CatCost (PI: Fred Baddour)

CCPC Consortium for Computational Physics and Chemistry U.S. DEPARTMENT OF ENERGY BIOENERGY TECHNOLOGIES OFFICE

CatCost

1. Management: Tasks

Project Tasks:

- Task 1: Data Hub and Catalyst Property Database Development ($325,000)
  - New milestone-related features in Catalyst Property Database (CPD)
  - Bug fixes and usability improvements in CPD and Data Hub
  - Research demonstrations of CPD
  - Adding new datasets to CPD
  - External user interviews, user training & documentation, webinars/screencasts, quality control

Tracked & prioritized using Agile methods

- Task 2: Data Hub Maintenance, Security, and Oversight ($25,000)
  - Security upgrades to Data Hub and Catalyst Property Database
  - Managing site hosting with Amazon Web Services

1. Management: Risk Mitigation

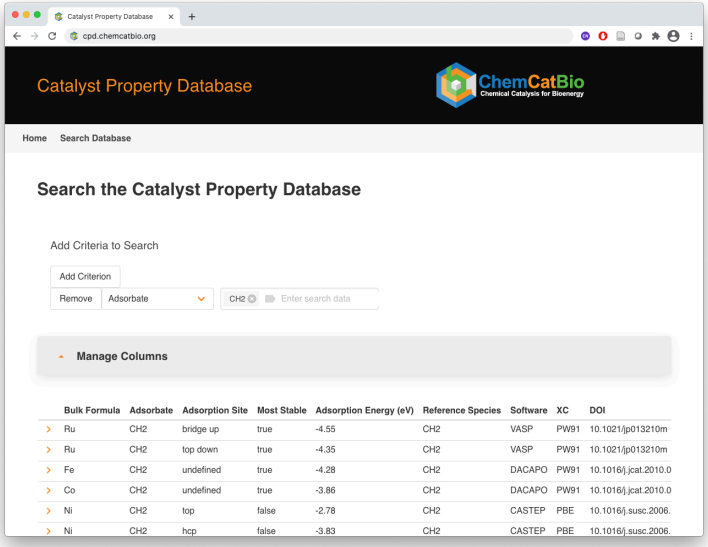
Risk 1: Development Scope & Schedule Creep	**Mitigation:** Use Agile (Scrum) software development systems & practices
Risk 2: Database Does Not Match User Needs (irrelevant / wrong features)	**Mitigation:** Go/No-Go milestone (FY21Q2) focuses on seeking expert / potential user feedback on development direction & pitfalls
Risk 3: Database Is Too Difficult To Use or Does Not Justify Required Effort	**Mitigation:** Ongoing input from catalyst researchers as part of Go/No-Go, NREL technical software developers & UI experts, and ChemCatBio stakeholders
Risk 4: Data Quality/Quantity Issues	**Mitigation:** Quality: Establish curation & training (FY21Q3) Quantity: Use partnerships with CCPC, CCB, external interviewees to gain buy-in & users, leverage “prestige” factor for data uploads

2. Approach: the Catalyst Property Database

In FY20–22, the Data Hub project is focused on developing the **Catalyst Property Database** as a collaboration and discovery tool

The Catalyst Property Database (CPD)

- A centralized, searchable repository of catalyst properties
- Publicly accessible to view and upload
  - Uploads subject to quality control
- Initial release: DFT-computed adsorption energies for intermediates on catalyst surfaces

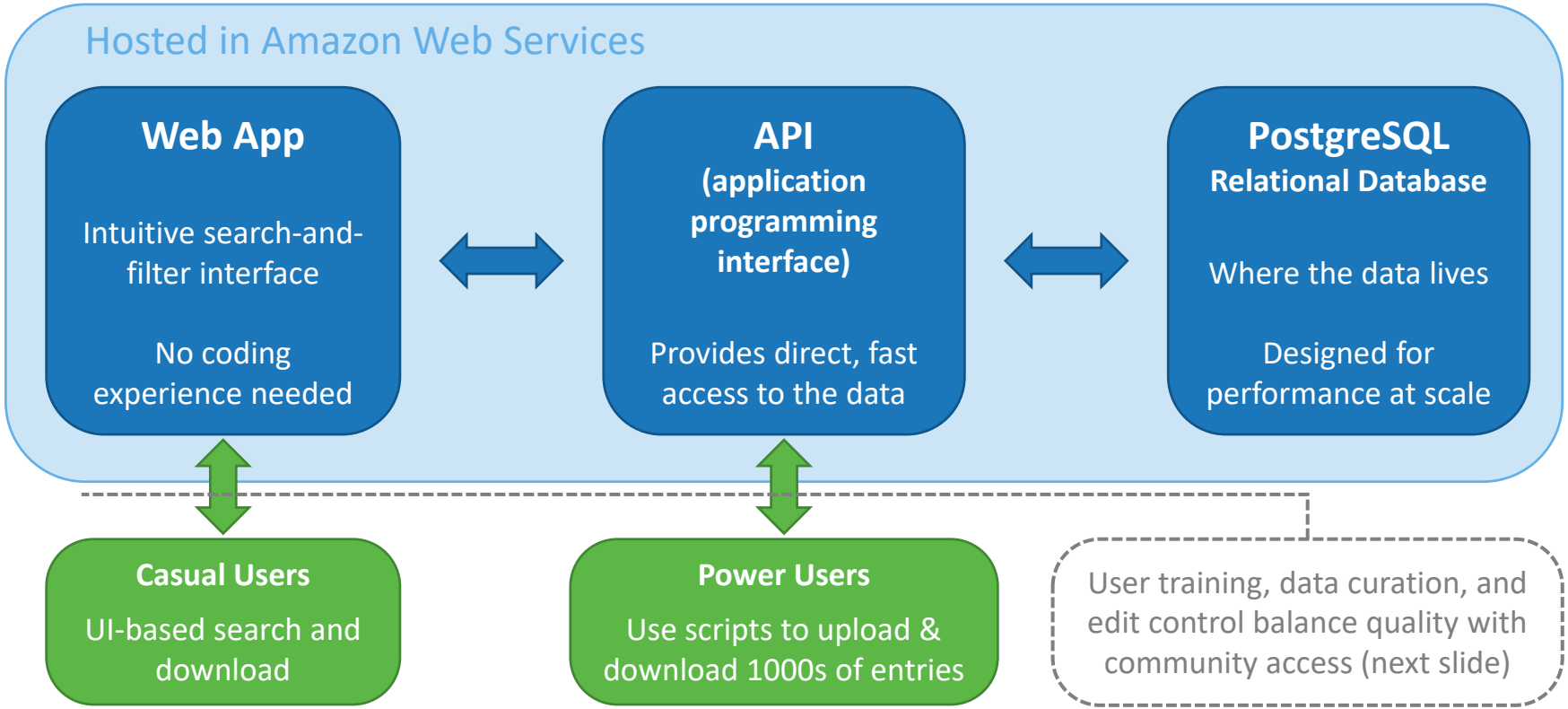


DFT = density functional theory

Note: The CPD was previously called the *Computational* Catalyst Property Database

2. Approach: CPD Architecture

The CPD uses a modern design for fast performance as the database grows



2. Approach: Data Growth & Curation

Problem:

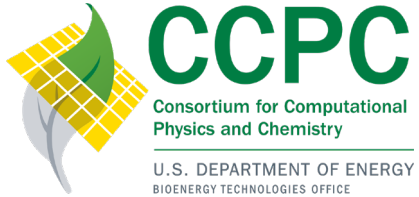
Achieving sustainable growth for the CPD while maintaining high quality

CPD Solution:

Open the database to public contributions, but use robust quality controls

Data Growth:

- Work with our partners to encourage large uploads



Quality Controls:

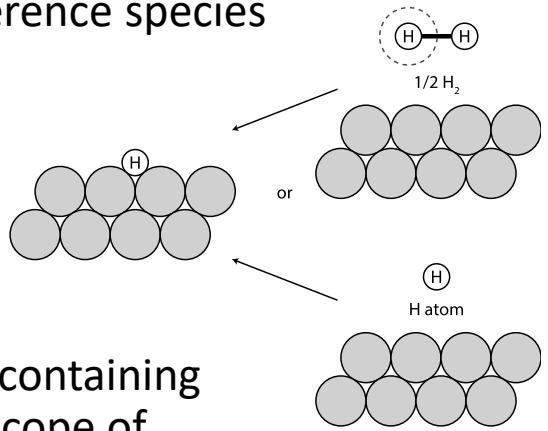
- Strict data requirements built into database
- Initial data curated by Data Hub team
- Long-term: use a Wikipedia model, but with qualified & trained curator-editors

2. Approach: Reference Species Translation Feature

Background: Computed adsorption energies, the critical DFT output included in the Catalyst Property Database, may be reported with different reference species

Adsorption energy (E_i*, in eV; 1 eV = 96.5 kJ/mol) for atomic H on a Pt(111) surface calculated with different gas-phase references.

Gas-phase reference	E_i* (eV)	|Difference| (eV)
H	–2.80	2.26
½H_2	–0.54	

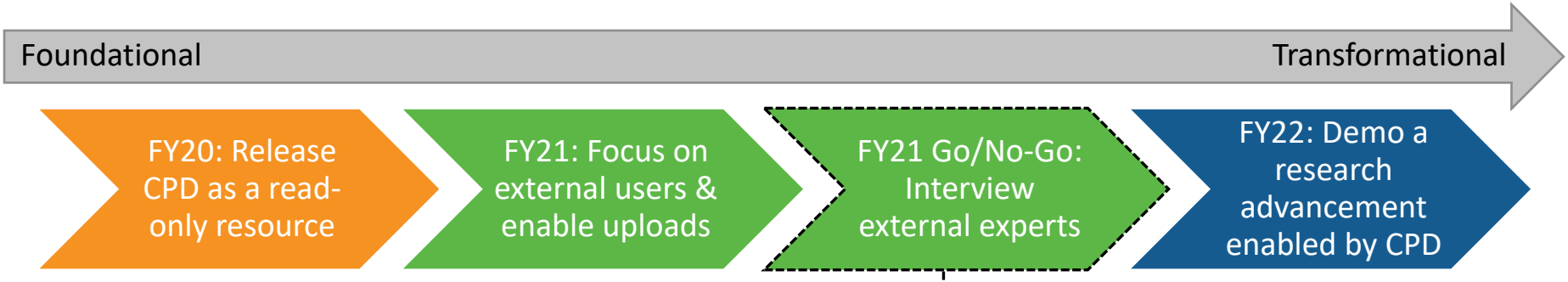


Problem: Data reported with different reference species, while containing similar information, **cannot be directly compared**, limiting the scope of data that can be applied in predictive catalyst applications

CPD Solution: Create a Reference Species Translation feature to enable interconversion between compatible reference species sets.

This is a key differentiator not found in any public database or resource.

2. Approach: Development Plan



FY21 Go/No-Go:
Interview 10+ experts to confirm that CPD development plans align with the needs of potential users and adjust as necessary.

Focus areas for interviews, to support CPD innovation:

- Preferred features
- Experience with competing solutions
- Strengths-Weaknesses-Opportunities-Threats
- Preferences for UI, scripting, etc.

3. Impact: Faster & Cheaper Catalyst Discovery

2019 Peer Reviewer Comment:

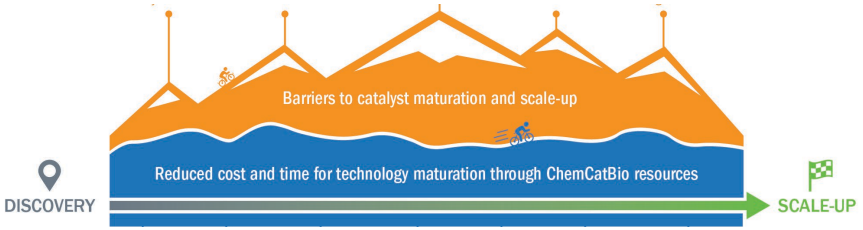
"We see considerable duplication of effort in catalysis research"

Every year, **more experimental and computational catalyst data is generated**, but the methods and **tools to apply this data have not kept up**.

The CPD is advancing the state of the art for application of computational data:

- Stop spending time and money **re-creating data** that already existed
- Enable **new approaches** to catalyst discovery that require large datasets

By harnessing the power of data, the Data Hub and the Catalyst Property Database are accelerating catalyst discovery



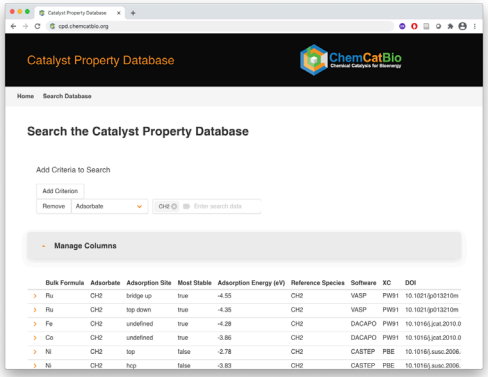
3. Impact: Diverse R&D Applications of the CPD

The Catalyst Property Database has diverse R&D applications, starting now and expanding as the database grows:

Computational Researcher
Data Validation/
Benchmarking

10–1,000 entries

cpd.chemcatbio.org



1,000–100,000 entries

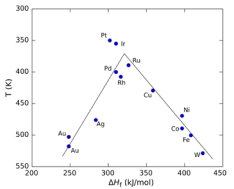
High-Throughput Calculations

1M+ entries

Catalysis Machine Learning

1,000–100,000 entries

CCPC Consortium for Computational Physics and Chemistry
U.S. DEPARTMENT OF ENERGY BIOENERGY TECHNOLOGIES OFFICE



Reactivity Descriptor Discovery

DFT = density functional theory

4. Progress: Create Database Data Structure

FY20: Developed the data structure for computed adsorption energies

Field	Data Type	Required Field?
Material Properties		
Formula	String	Yes
Primary Class	String	No
Secondary Class	String	No
Stretched?	Boolean	Yes, default False
Compressed?	Boolean	Yes, default False
Space Group	String	No
Lattice Constants (3)	Numeric	No
Surface/Particle Properties		
Nanoparticle Size	Numeric	No
1st Layer Composition	String	No
2nd Layer Composition	String	No
Facet	String	No
Termination	String	No
Cell Symmetry	String	No
Methods		
Software	String	Yes
Exchange correlation	String	Yes
Potentials	String	No
Basis Set	String	Yes
SpinPol?	Boolean	No
ZPE?	Boolean	No
Fixed Substrate?	Boolean	No

Field	Data Type	Required Field?
Adsorbate and Reference Species		
Adsorbate	Table	Yes
Adsorption Site	String	No
Coverage	Numeric	No
Reference Species (multiple)	Table	At least one entry required
Reference Species coefficient	Numeric	Yes
Metadata		
DOI	String	Yes
Notes	String	No
User	String	Yes
Adsorption Energy Data		
Adsorption Energy	Numeric	Yes
Most Stable?	Boolean	Yes, default True

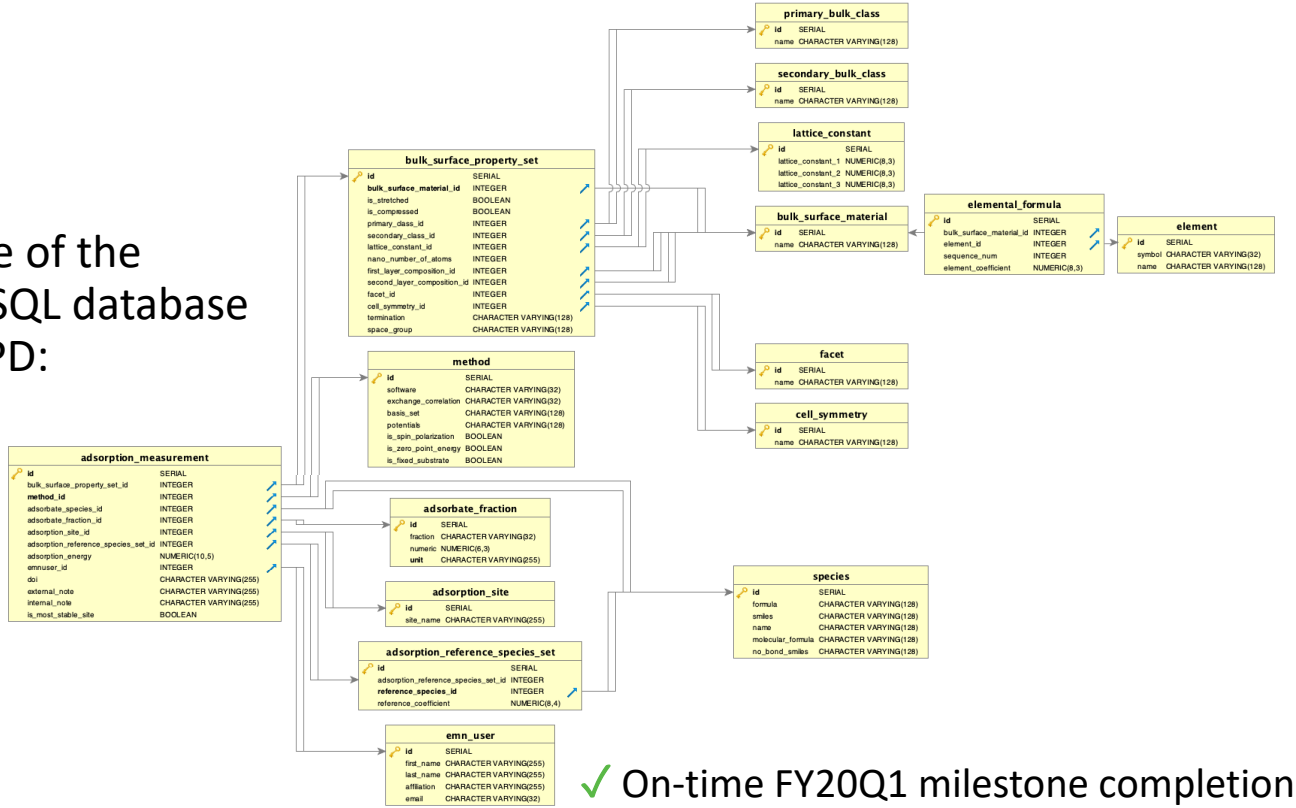
Outcome: This structure and the associated rules are essential to data quality control & curation

✓ On-time FY20Q1 milestone completion

4. Progress: Create Database Data Structure

FY20: Converted the data structure into a PostGreSQL database

Structure of the PostGreSQL database in the CPD:



Outcome: Database ready to efficiently scale up to millions of entries

✓ On-time FY20Q1 milestone completion

4. Progress: Standardize Data & Create Rules

FY20: Created dictionaries of recurring entries, such as adsorbate/reference species, and rules for naming, data input/curation, etc.

	Molecular Formula	Name	SMILES
	~~O2 (g)~~	~~oxygen (g)~~	~~[OO]~~
	~~OO~~	~~dioxygen~~	~~OO~~
	O2	oxygen	O=O
	~~HH~~	~~hydrogen (g)~~	~~[HH]~~
	H2	hydrogen	[H][H]
	O	oxygen, atomic	[O]
	N2	nitrogen	N#N
	CH3COOH	acetic acid	CC(=O)O
	HCO	formyl	[CH]=O
	CH2	methylene	[CH2]
	COOH	carboxyl	[C](=O)O

Outcomes from dictionaries & data rules:

- Quality control
- Faster performance
- Faster, easier search
- Gives users confidence in the data

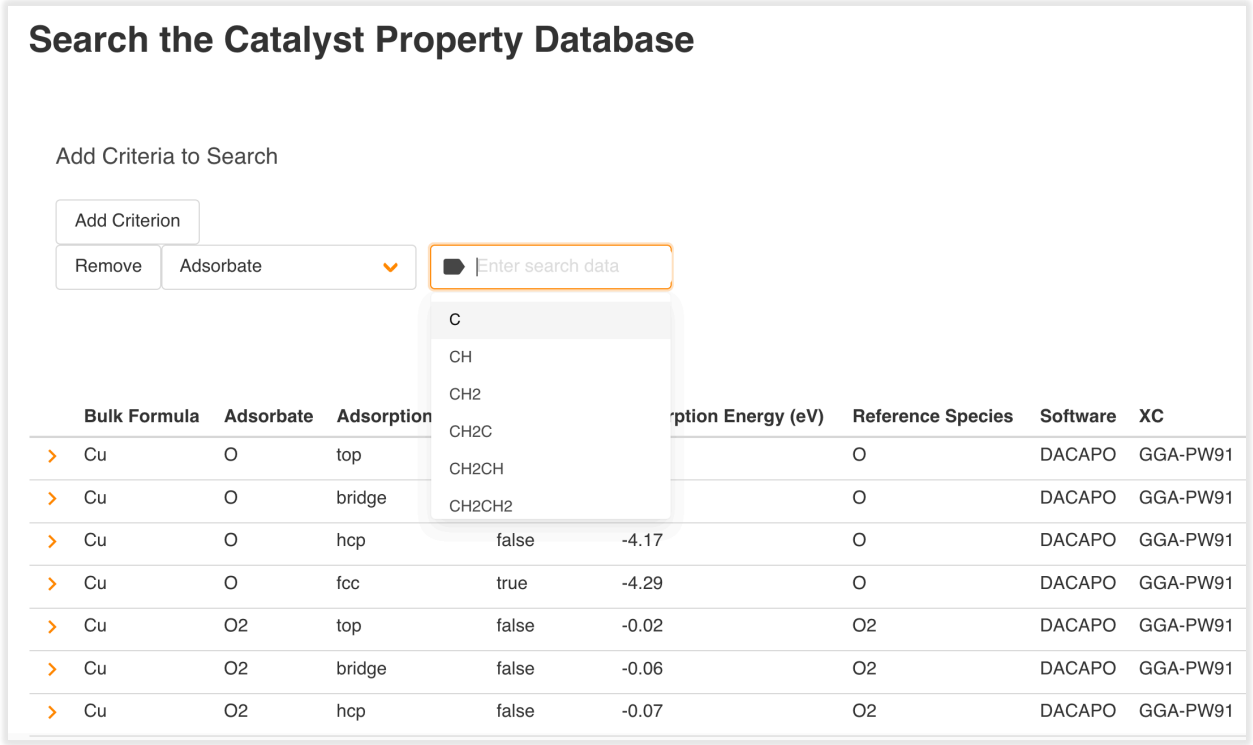
✓ On-time FY20Q1 milestone completion

4. Progress: Create User Interface

FY20: Developed the UI for the Catalyst Property Database, including a live-updating search/filter approach:

Outcome:

The clean, modern interface allows catalyst researchers of all backgrounds to access valuable data



✓ On-time FY20Q3 milestone completion

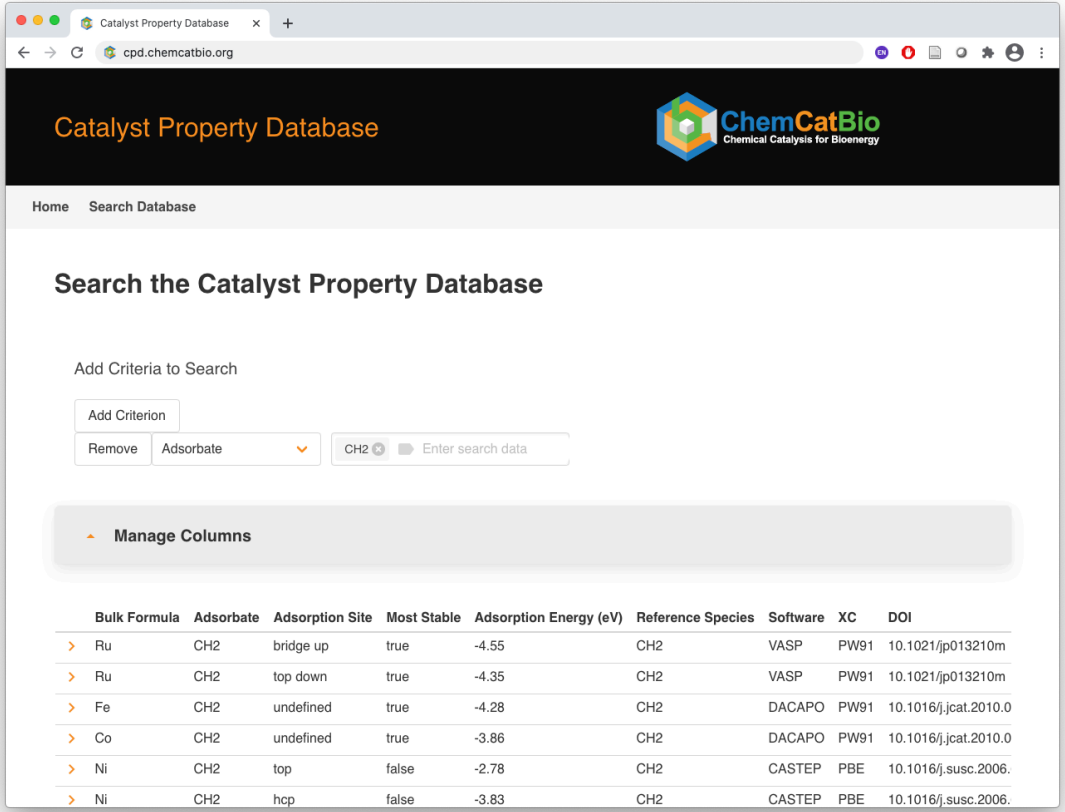
4. Progress: Initial Public Release

FY20: Released the Catalyst Property Database to the public in Sept 2020

cpd.chemcatbio.org

Outcome:

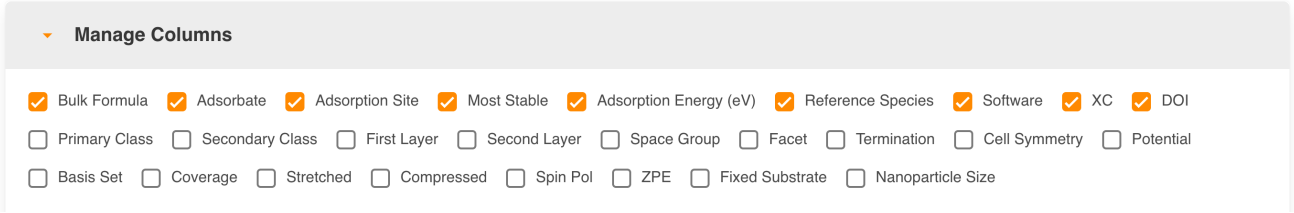
The CPD is now available for users to test and provide feedback.



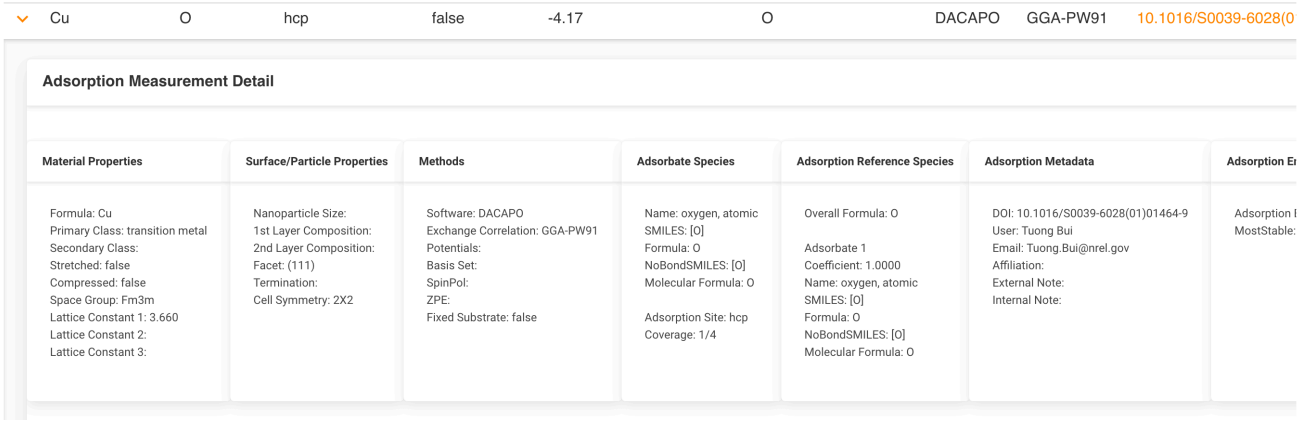
✓ On-time FY20Q4 milestone completion

4. Progress: Ongoing Development

FY21: Added a Manage Columns dialog to allow users to select columns for display...

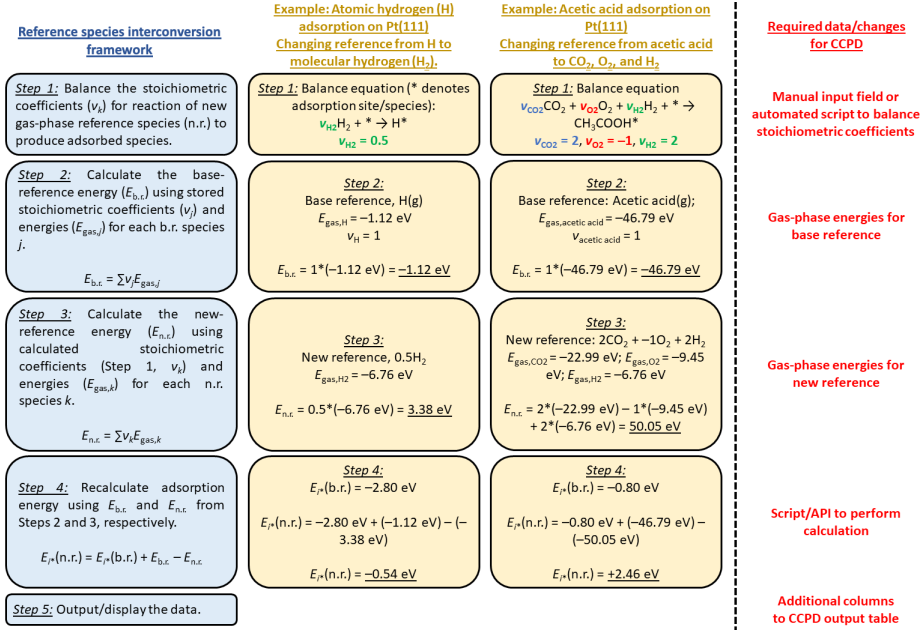


... and a "detail view" to show all available metadata for a row



4. Progress: Reference Species Translation

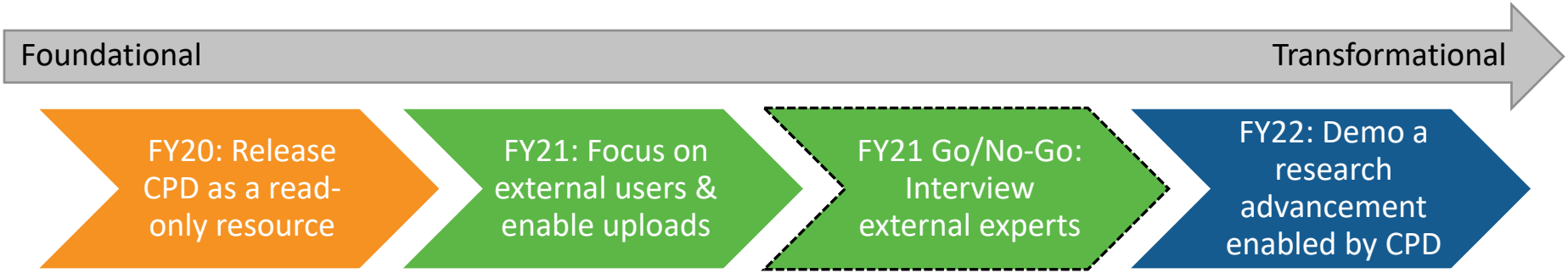
FY21: Created a detailed development plan for this feature, including required data



Outcome: This key differentiator is complex to implement, so we broke it down into achievable development steps.

✓ On-time FY21Q1 milestone completion

Future Work



Upcoming work in the Data Hub project:

FY21Q2: Go/No-Go – Interviews with external experts; guide development direction

FY21Q3: Develop training, documentation, and curation procedures

FY21Q4: Add batch upload capability, including examples (e.g., Python script)

FY22: Demonstrate a research advancement enabled by the unique data-sharing capabilities of the CPD. Time savings of 2–10X compared to best alternative methods.

Quad Chart

Timeline

- Project start date: 10/1/2019
- Project end date: 09/30/2022

	FY21	FY20–FY22
DOE Funding	$350k	$1.05M

Barriers addressed

- Ct-F – Increasing the yield from catalytic processes
- Ct-G – Decreasing the time and cost to develop novel industrially relevant catalysts

Funding Mechanism

AOP

Project Goal

Enable ChemCatBio and the bioenergy industry to accelerate the catalyst and process development cycle through development of publicly available advanced analytics tools. Demonstrate application of the Computational Catalyst Property Database (CPD) in a predictive design application that is dramatically accelerated (2–10X) by the use of the CPD's unique capabilities.

End of Project Milestone

Demonstrate a research advancement enabled by the unique data-sharing capabilities of the CPD. Pilot and document a research use case of the CPD that results in a time savings of 2–10X compared to the best alternative methods. Perform this demonstration with a suitable partner, such as a high-throughput computational research group.

Acknowledgements

Data Hub Team

Carrie Farberow

Nalinrat Guba

Sean Tacey

Matt Jankousky

Tuong Bui

Nick Wunder

Kris Munch

Courtney Pailing

Josh Schaidle

Advisors

Kathy Cisar

Fred Baddour

Tom King

BETO

Trevor Smith

Jesse Glover

Andrea Bailey

Nicole Fitzgerald

This research was supported by the DOE Bioenergy Technology Office under Contract no. DE-AC36-08-GO28308 with the National Renewable Energy Laboratory

This work was performed in collaboration with the Chemical Catalysis for Bioenergy Consortium (ChemCatBio, CCB), a member of the Energy Materials Network (EMN)

ChemCatBio
Chemical Catalysis for Bioenergy

Energy Materials Network
U.S. Department of Energy

U.S. DEPARTMENT OF ENERGY | Office of ENERGY EFFICIENCY & RENEWABLE ENERGY
BIOENERGY TECHNOLOGIES OFFICE

Thank you!

Data Hub: Informed by 2019 Peer Review

"We see considerable duplication of effort in catalysis research"

- The Catalyst Property Database aims to make data easier to find and reuse, cutting down on redundant computations/measurements

Should allow contributions from non-CCB researchers

- The CPD is designed for contributions by the broader catalysis community

"It would be nice to see...a more universal repository of the sort that NIST curates"

- The newly-renamed Catalyst Property Database is designed to work toward this vision

Should use Agile project management principles

- We are using Agile/Scrum in our development

User feedback and confirmation of usefulness is essential

- We have made external user feedback the core of our Go/No-Go milestone