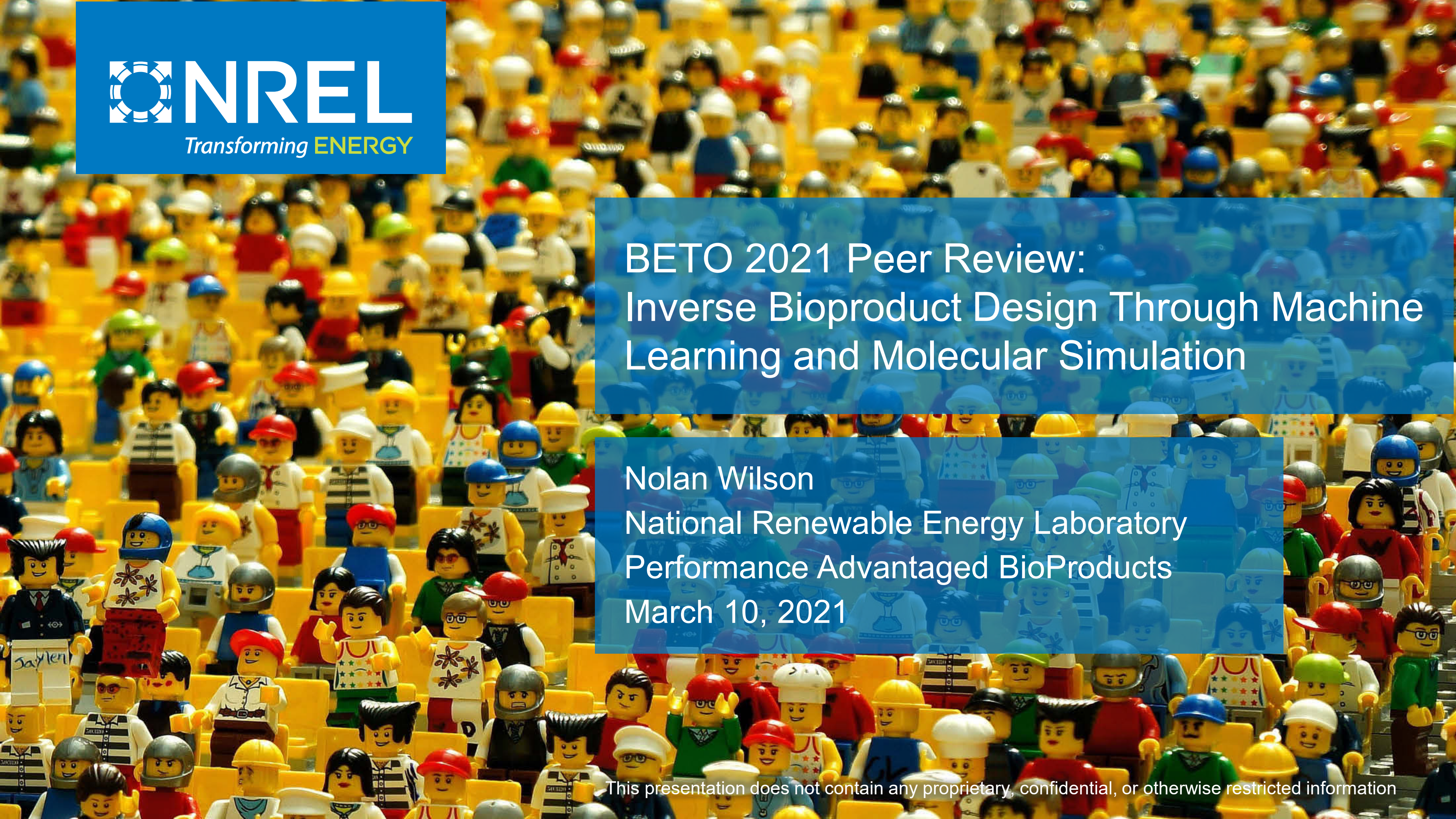


NREL
Transforming ENERGY

BETO 2021 Peer Review: Inverse Bioproduct Design Through Machine Learning and Molecular Simulation

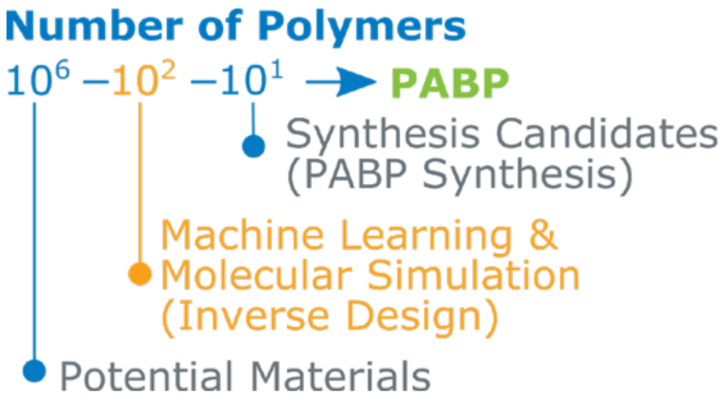
Nolan Wilson
National Renewable Energy Laboratory
Performance Advantaged BioProducts
March 10, 2021

This presentation does not contain any proprietary, confidential, or otherwise restricted information

Project Overview

Goal: Guide experimental synthesis and reduce time-to-market for PABPs by predicting properties from molecular structure.

Objective: Build machine learning (ML) and molecular simulation (MS) tools that enable high throughput property prediction of biobased thermoplastics, thermosets, and additives.



Number of Polymers

10^6 – 10^2 – 10^1 → **PABP**

Synthesis Candidates (PABP Synthesis)

Machine Learning & Molecular Simulation (Inverse Design)

Potential Materials

Today’s Technology

The Edisonian approach to materials discovery is insufficient to screen the >10^6 polymers accessible from biomass. Prediction approaches for polymers use hand-engineered features.

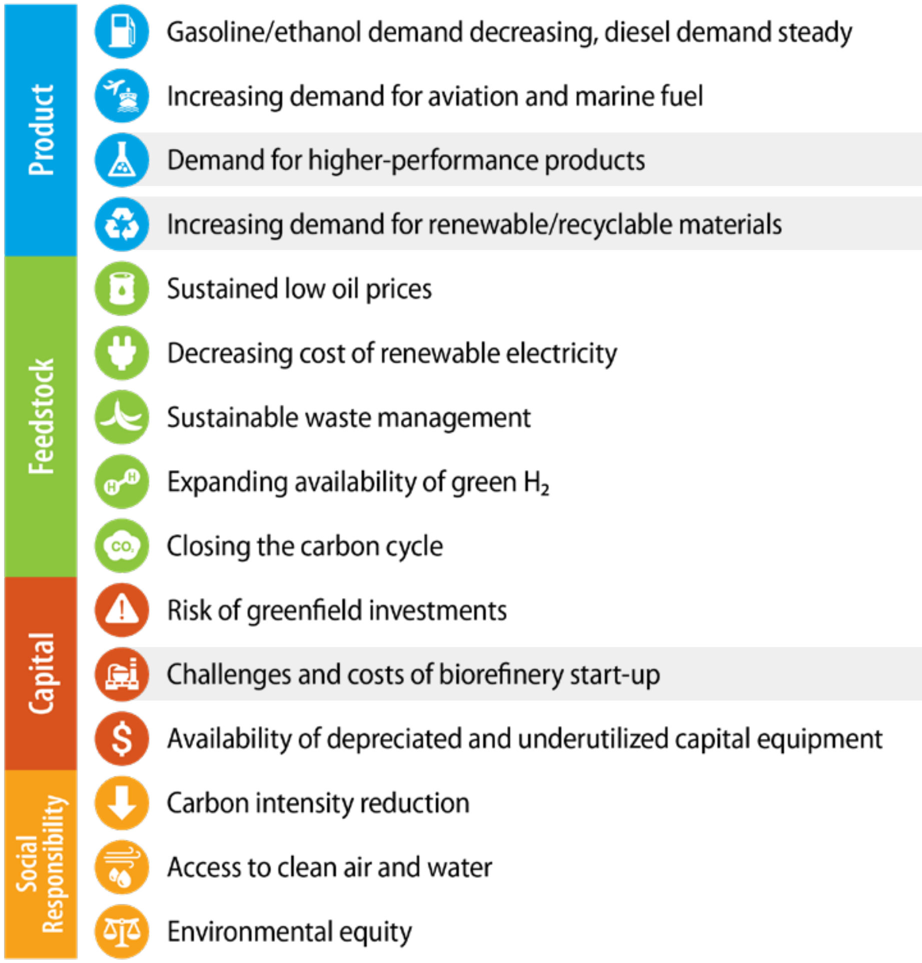
Importance

The unique chemical functionality resulting from biomass conversion can enable sustainable polymers with improved performance to supplant existing materials.

Risks

- Low accuracy and throughput
- Lack of interpretability for structure function relationships
- Low data availability & marginal structural embedding

Market Trends



Product	Gasoline/ethanol demand decreasing, diesel demand steady
	Increasing demand for aviation and marine fuel
	Demand for higher-performance products
	Increasing demand for renewable/recyclable materials
Feedstock	Sustained low oil prices
	Decreasing cost of renewable electricity
	Sustainable waste management
	Expanding availability of green H_2
	Closing the carbon cycle
Capital	Risk of greenfield investments
	Challenges and costs of biorefinery start-up
	Availability of depreciated and underutilized capital equipment
Social Responsibility	Carbon intensity reduction
	Access to clean air and water
	Environmental equity

NREL's Bioenergy Program Is Enabling a Sustainable Energy Future by Responding to Key Market Needs

Value Proposition

- Increase the rate of PABP discovery to reduce cost and time-to-product.
  - Predictions take seconds
  - Synthesis take days to months
- Down select from 10^6 to 10^2 candidates so experiments can focus on likely PABP

Key Differentiators

- End-to-end neural nets and high-fidelity structure generation can increase prediction accuracy and throughput
- Development of best practices for automated atomistic modeling of PABP polymer systems

1. Management

Management Approach & Team

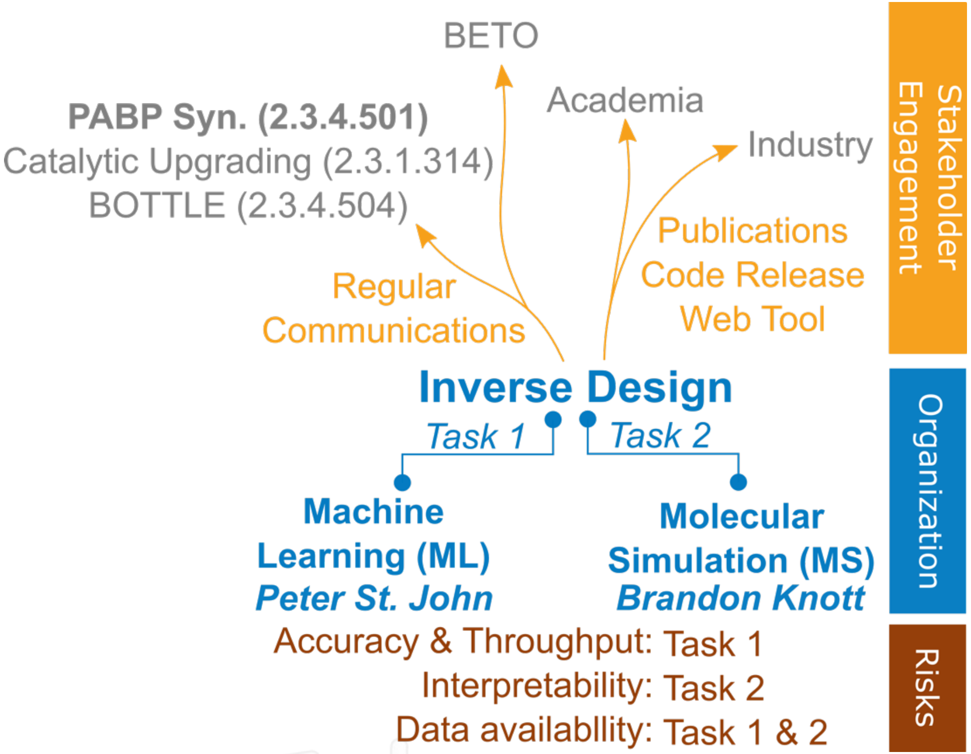
Use expertise in multiple simulation approaches to provide capabilities greater than sum of the parts.

Nolan Wilson (PI)
Polymer engineering and design

Peter St. John
Machine learning for molecular property prediction

Brandon Knott
Molecular dynamics for structure function relationship elucidation

Michael Crowley (Former PI)
Macromolecular Simulation, QM/MM, CHARMM, Amber



1. Management

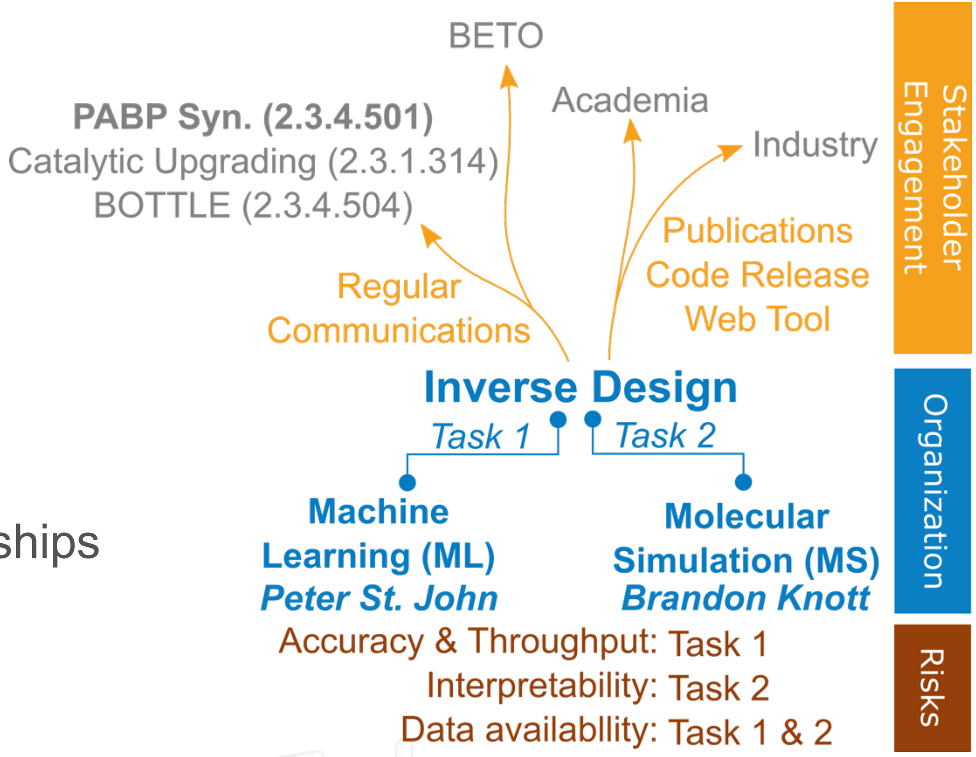
Addressing Risks

Project structured so risks are addressed by each task, which has the right expertise within the task

Task 1: ML can be accurate and make high-throughput predictions.

Task 2: MS can provide mechanistic insights into structure-function relationships and make predictions in absence of training data.

Task 1 & 2: MS data can augment ML training sets to increase size and domain of data



BETO

Academia

Industry

PABP Syn. (2.3.4.501)
Catalytic Upgrading (2.3.1.314)
BOTTLE (2.3.4.504)

Regular Communications

Publications
Code Release
Web Tool

Inverse Design

Task 1

Task 2

Machine Learning (ML)
Peter St. John

Molecular Simulation (MS)
Brandon Knott

Accuracy & Throughput: Task 1
Interpretability: Task 2
Data availablity: Task 1 & 2

Stakeholder Engagement

Organization

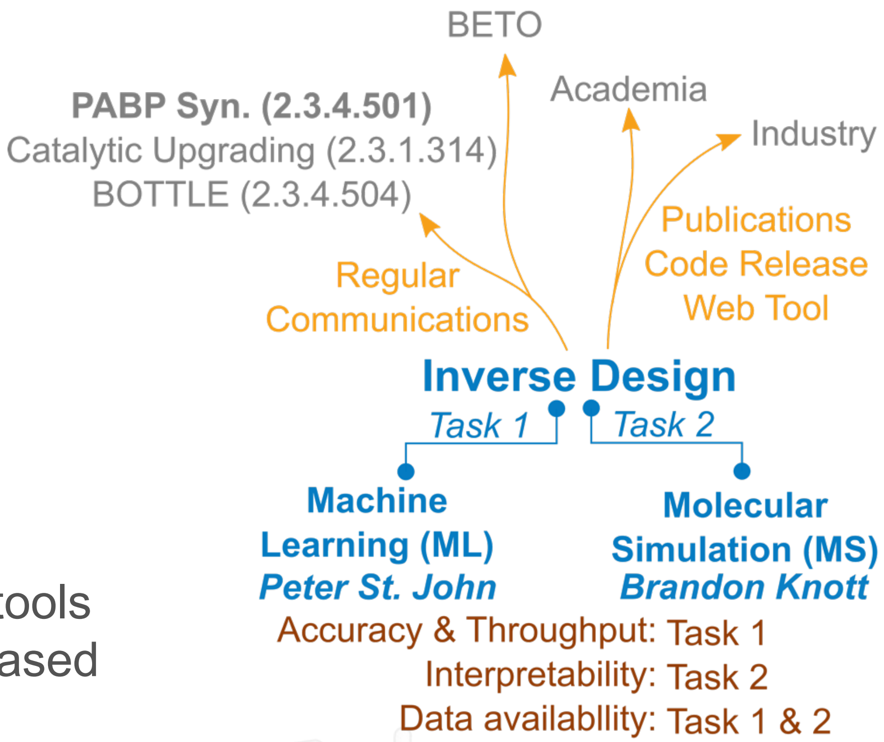
Risks

1. Management

DOE-BETO Related Projects: Predicted materials have been synthesized in PABP synthesis project. *Related Risk: Low Accuracy*

Research Community: Release of 3 open-sourced code stacks[1-3] and development of web-based tool for non-experts

Biomaterials Industry: Integration of tools into 2 projects in FY21 to develop biobased materials with commercial partners.



Stakeholder Engagement

Organization

Risks

[1] https://pypi.org/project/nfp/
[2] https://pypi.org/project/m2p/
[3] https://pypi.org/project/common-wrangler/

2. Approach

Goal: Discover novel bioproducts by predicting properties from molecular structure, which will guide synthesis and reduce time to market.

Objective: Build machine learning (ML) and molecular simulation (MS) tools that enable high throughput property prediction of biobased thermoplastics, thermosets, and additives.

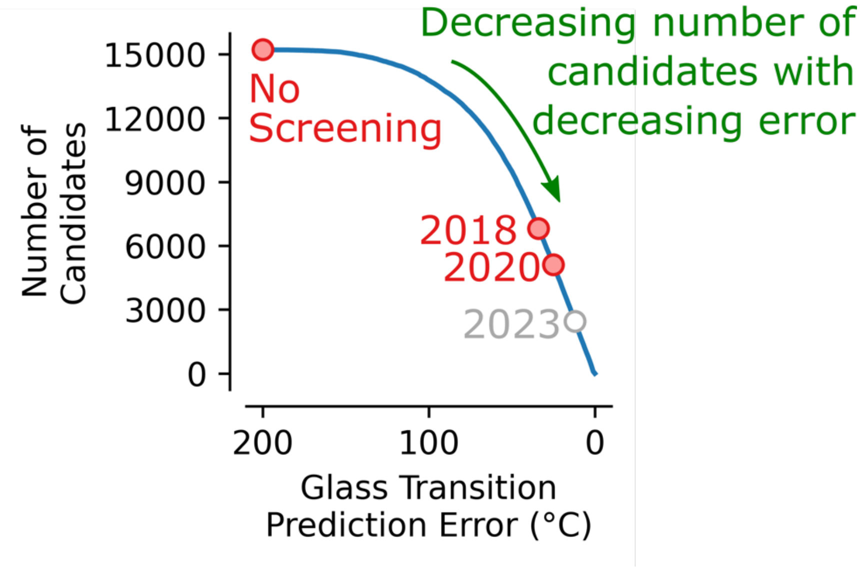
Key Milestones for Achieving Objectives	Metric	Quarter
Web-based tool for polymer prediction.	PolyML webtool deployment	FY20Q4
Validate ML & MS thermoplastic predictions with experiment	> 5 thermoplastics	FY21Q4
Demonstrate ML + MS can improve accuracy	> 10% improvement in mean absolute error (MAE)	FY22Q2 – Go/NoGo
Thermoset predictions Significant increase accuracy	Predict >100 PABP thermosets, >50 % improvement in MAE	FY23Q4

2. Approach

Research Approach

- Increase ML accuracy and screen using multiple properties to improve ability to down select
- MS can be used to describe structure-function relationships and inform experimental synthesis

- Augment training sets, improving network architecture
- Close coupling with experimental efforts (PABP Syn. Project)

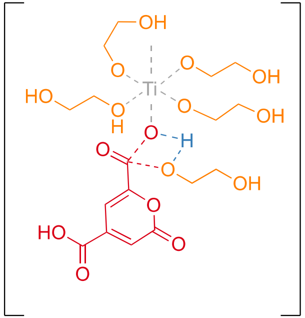


3. Impact

This technology moves bioproduct development from brute-force to informed discovery approach and will catalyze the adoption of biobased thermoplastics, thermosets, and polymer additives.

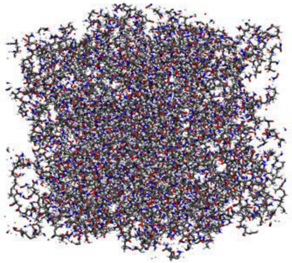
Bioeconomy & PABP Discovery

- Directed experimentalists towards PABP PET replacements
- Elucidated catalytic mechanism for PET replacement and directed experimentalists to new synthesis approach



Scientific Community

- Reaction mechanisms for polymerization of biopolymers
- Mechanistic understanding for structural design of biopolymers and bioproducts[1]
- New machine learning architectures for polymers
- Higher throughput and accuracy



[1] Rorrer *et al.*, in review

3. Impact

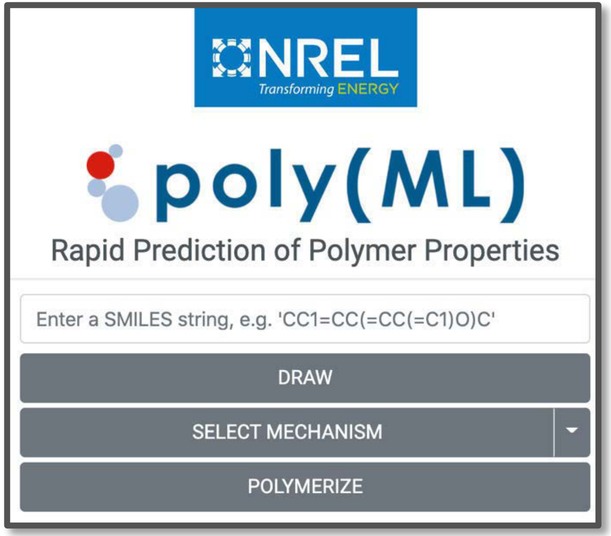
Industry

- Remove tradeoffs between performance and sustainability in new polymer design
- Providing access to state-of-the-art material design tools for experts and non-experts

Interests & Partnerships

- **In FY21, we will be starting a project with Sealy, Patagonia, and Agilix for the "Commercialization of Fully Renewable Non-Isocyanate Polyurethanes"**
- Univ. Wisconsin, Univ. Maine, CSU, Lehigh Univ., LANL, IBM, Pyran, Checkerspot, BOTTLE Consortium

polyML's web tool enables state-of-the art ML polymer property prediction by non-experts (external release pending peer review manual)



NREL Transforming ENERGY

poly(ML)

Rapid Prediction of Polymer Properties

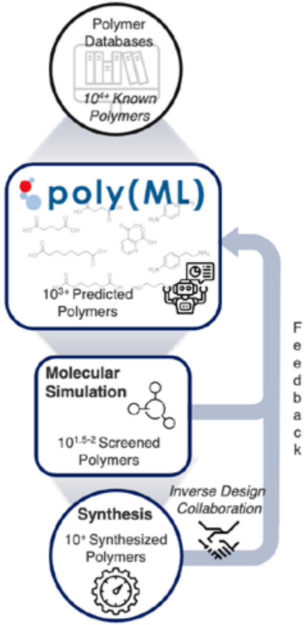
Enter a SMILES string, e.g. 'CC1=CC(=CC(=C1)O)C'

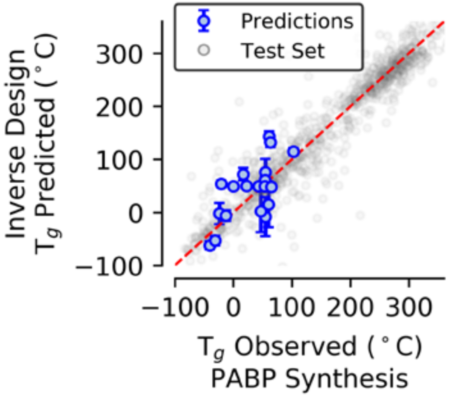
4. Progress and Outcomes

Task 1: Machine Learning

Made predictions for 1.4 x 10^6 biopolymers and benchmarked predictions with experimental data (*joint with PABP syn.*)

- Accuracy improvement: Tg_{MAE}= 11°C
- End-to-end embedding and automated structure generation


Polymer Databases
10^{4+} Known Polymers
poly(ML)
10^{3+} Predicted Polymers
Molecular Simulation
$10^{1.5-2}$ Screened Polymers
Synthesis
10^{+} Synthesized Polymers
Inverse Design Collaboration
Feedback



Throughput	10^2 predictions s^{-1}
Properties	Glass Transition & Melt Temp., Density, Modulus, Permeability of O_2, N_2, CO_2, & H_2O
Polymers	Olefins, Acrylates, Esters, Amides, Carbonates, Imides

4. Progress and Outcomes

Task 1: Machine Learning

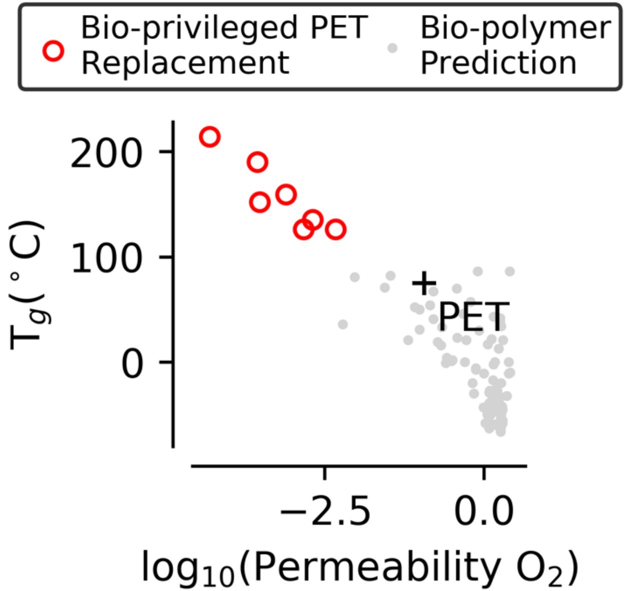
Case Study: Discovering a PABP PET replacement

Significance: Polyethylene terephthalate is used in films and bottles. A replacement PET with increased Tg and lower O_2 permeability will be performance advantaged.

Study: Use ML to predict polymers accessible from KEGG database to identify PABP PETs[1]

Results:

- Screened 15,222 polyesters
- 7 identified targets
- Polymer targets are currently being synthesized.
- Molecular simulations are investigating structure-function relationships



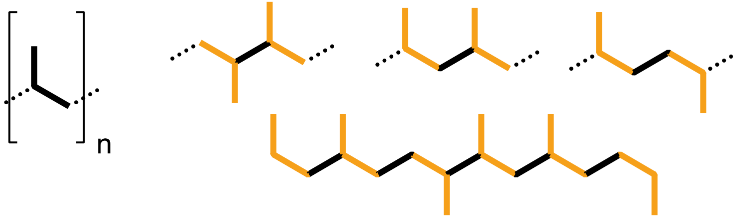
[1] Wilson *et al.* in preparation

4. Progress and Outcomes

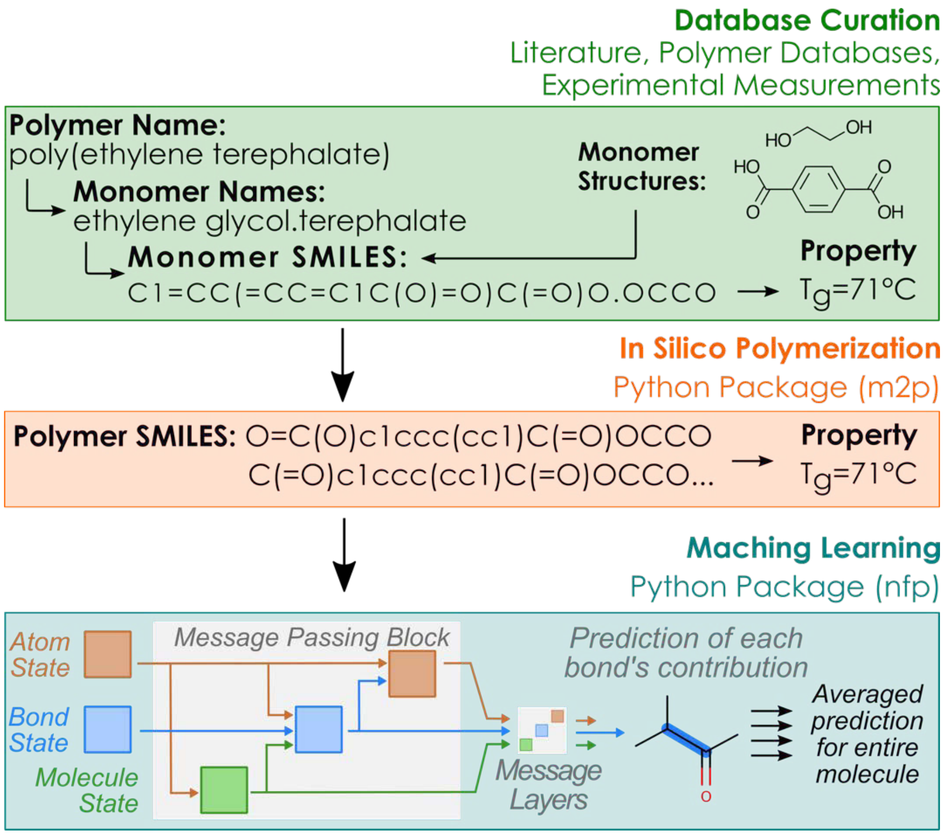
Task 1: Machine Learning

Polymer Database Development
Literature & databases, in-house experimental data, document discovery, transfer learning

Polymer Structure Generation
High throughput & high fidelity



Graph Neural Networks
End-to-end learning using message passing neural networks[1]



Database Curation
Literature, Polymer Databases, Experimental Measurements

Polymer Name:
poly(ethylene terephalate)
Monomer Names:
ethylene glycol.terephalate
Monomer SMILES:
C1=CC(=CC=C1C(O)=O)C(=O)O.OCCO → **Property** T_g=71°C

Monomer Structures:

In Silico Polymerization
Python Package (m2p)

Polymer SMILES: O=C(O)c1ccc(cc1)C(=O)OCCO C(=O)c1ccc(cc1)C(=O)OCCO... → **Property** T_g=71°C

Maching Learning
Python Package (nfp)

Atom State
Bond State
Molecule State
Message Passing Block
Prediction of each bond's contribution
Message Layers
O
Averaged prediction for entire molecule

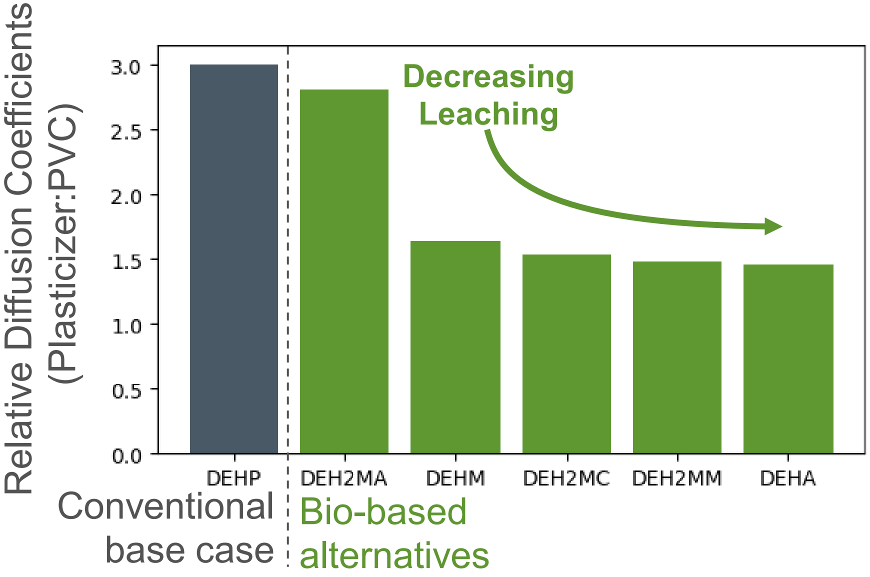
[1] St John, P. C. *et al. J. Chem. Phys.* **150**, (2019).

4. Progress and Outcomes *Task 2: Molecular Simulation*

Expanded Simulations to Include Polymer Additives

- Evaluated 5 low toxicity bio-based plasticizers in PVC over conventional plasticizer.
  - **Reduced Leaching**
  - **More effective Loading (less material)**
  - Glass Transition
  - Viscosity

- Established atomistic approaches vs. coarse grained for biobased plastics & bioadditives
- Evaluation of 3 forcefields based on property prediction for 5 polymers



4. Progress and Outcomes

Task 2: Molecular Simulation

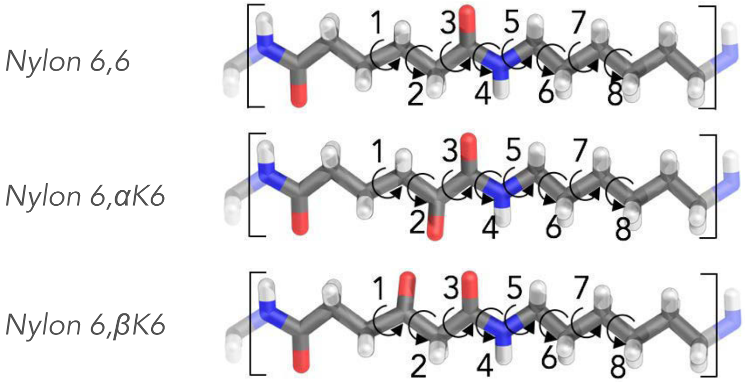
Case Study: Structure-Function Relationship for PABP Nylon

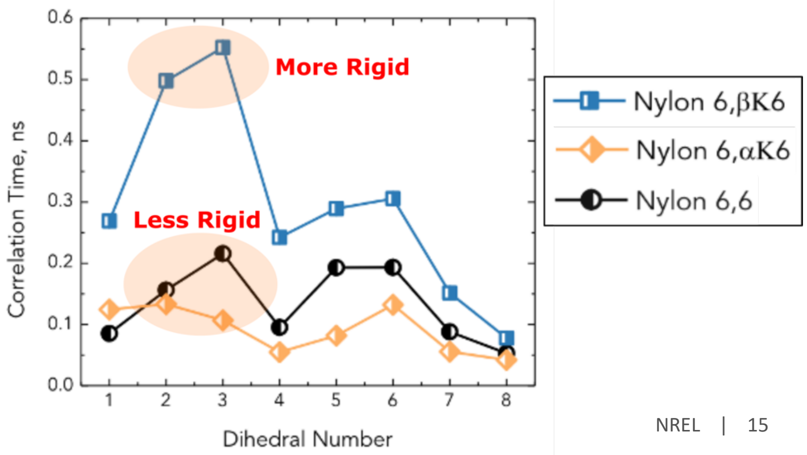
Significance: Experimental observation of β-ketoadipate increase of performance, but not for α-ketoadipate.

Study: Use MD to interrogate structure-function relationship for design principle around ketone containing monomers.

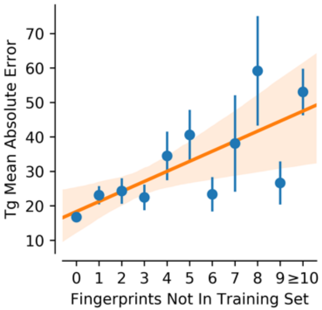
Results:

- Dihedral in nylon 6,βK6 is "locked" into a single confirmation, in contrast to nylon 6,6 and nylon 6,αK6, increasing glass transition temperature.
- Enhanced interchain hydrogen bonding is observed when ketone is introduced into nylon 6,6 at the β, but not the α, position

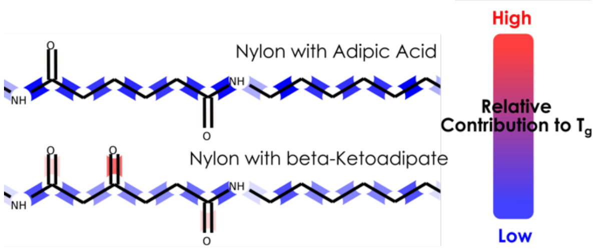




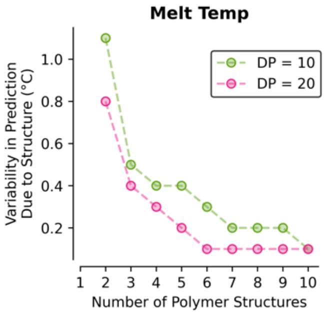
Additional ML & MS Development



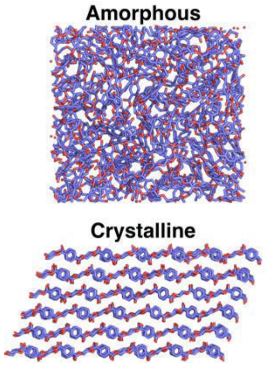
Domain of validity method development for polymers



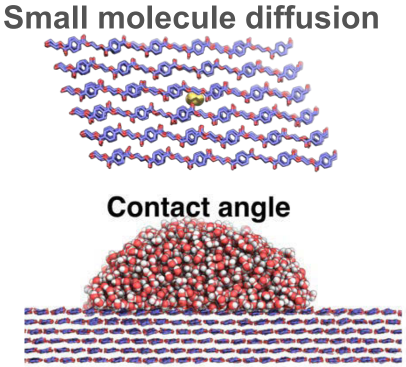
Structural heat mapping for structure function information



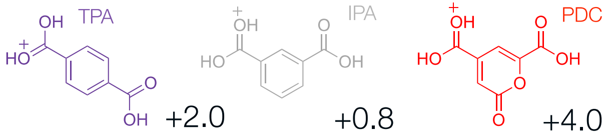
Network topology & structure optimization for prediction



Phase dependent MS system building for polymers



MS based property prediction



DFT based reactivity estimate for biomolecules

Quad Chart Overview (for AOP Projects)

Timeline

- Start: FY18 – FY20
- Renewed: FY21 – FY23

	FY21	Active Project (FY21-FY23)
DOE Funding	*$400K*	$1,200K

Project Partners: Lehigh University

BETO Projects*:* Synthesis and Analysis of PABP project, BOTTLE Consortium, Catalytic Upgrading of Pyrolysis Products, Biological Lignin Valorization, Bioconversion of Thermochemical Intermediates

Barriers addressed

(Ct-J) Identification and Evaluation of Potential Bioproducts

(Ct-K) Developing Methods for Bioproduct Production

(Ct-N) Multiscale computational framework accelerating technology

Project Goal

Creating new opportunities for advanced biomaterials by predicting properties and performance of novel biomass-based materials based upon molecular structure, which will guide synthesis and reducing time to market.

Technical Approach

- Deploy **machine learning (ML)** tools to rapidly predict molecular properties from chemical structure; broaden application to thermosets and small molecules
- Employ **molecular dynamics (MD)** simulations and **quantum mechanics (QM)** calculations to predict and understand properties at molecular-level
- Leverage previously developed high-throughput MD pipeline to augment ML data sets

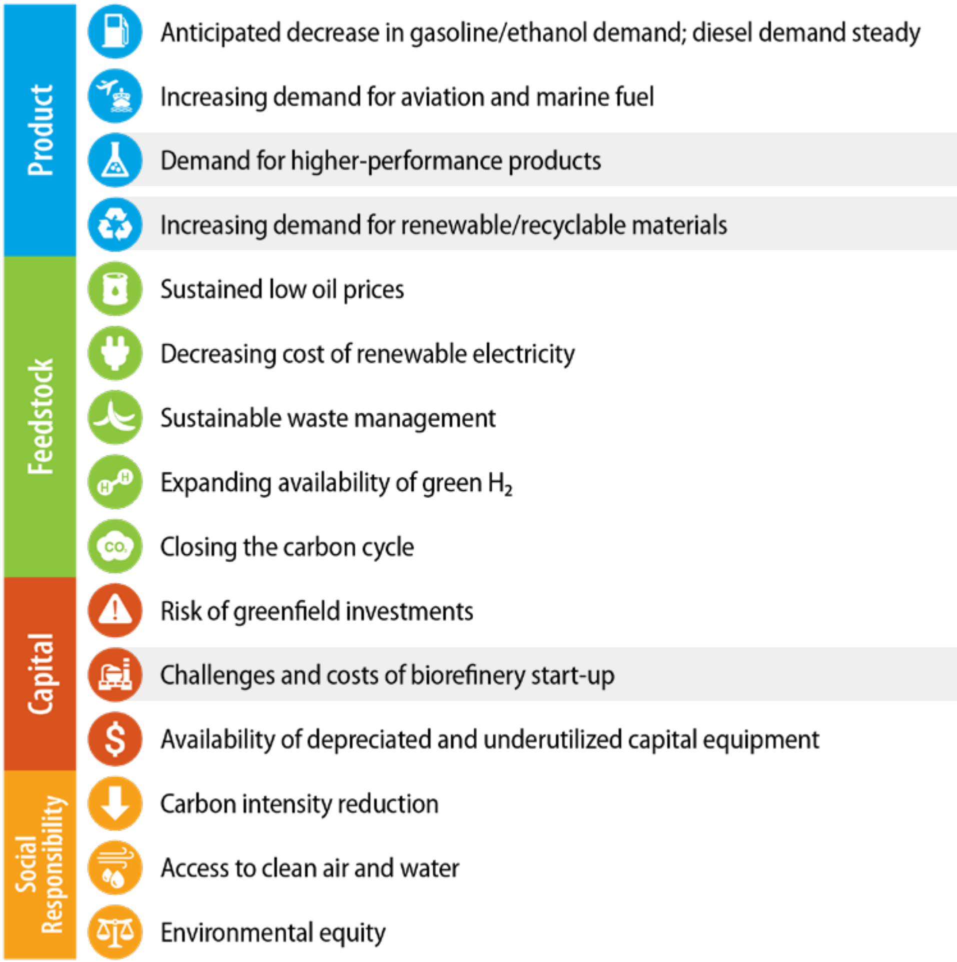
End of Project Milestone

Improve the accuracy of ML by 50% and identify 10 PABP thermoset materials

Funding Mechanism

Bioenergy Technologies Office FY21 AOP Lab Call (DE-LC-000L079) – 2020.

Summary



Product	Anticipated decrease in gasoline/ethanol demand; diesel demand steady
	Increasing demand for aviation and marine fuel
	Demand for higher-performance products
	Increasing demand for renewable/recyclable materials
Feedstock	Sustained low oil prices
	Decreasing cost of renewable electricity
	Sustainable waste management
	Expanding availability of green H_2
	Closing the carbon cycle
Capital	Risk of greenfield investments
	Challenges and costs of biorefinery start-up
	Availability of depreciated and underutilized capital equipment
Social Responsibility	Carbon intensity reduction
	Access to clean air and water
	Environmental equity

NREL's Bioenergy Program Is Enabling a Sustainable Energy Future by Responding to Key Market Needs

Management

Expertise across computational methods enables capabilities beyond any single approach

Approach

Aligned milestones to objective of ML & MS prediction tool for PABP discovery

Impact

Guide PABP synthesis & reduce time to market

Move from brute-force to informed discovery

Progress and Outcomes

- 1.4 x 10^6 biopolymer predictions
- 7 PABP PET
- 5 biobased plasticizers
- Design principle for PABP nylons

Acknowledgements

DOE Technology Manager Andrea Bailey (and Nichole Fitzgerald formerly)

PABP Synthesis Team (PI Gregg Beckham)

Inverse Design Team

Michael Crowley, Heather Mayes, Brandon Knott, Shivani Kozarekar, Mark Nimlos, Peter St. John

Thank You

www.nrel.gov

NREL/PR-2800-79420

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Bioenergy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

NREL
Transforming ENERGY

Additional Slides

Responses to Previous Reviewers' Comments

Summary of Key Questions/Criticisms

- Data availability and methods for sourcing data on non-commercial polymers and biopolymers
- Application of modelling approach to small molecules

Response

- The team is implementing natural language process techniques for document discovery to increase ability to pull data from literature. This will expand the breadth of polymers within that database as well as the rate at which the database size can be increased.
- The team is developing new methods to augment experimental data with computation data to ultimately increase data set size and prediction accuracy.
- The team is pursuing polymer additives (*e.g.*, plasticizers) as a relevant and related research area of small molecules.

Publications, Patents, Presentations, Awards, and Commercialization

Manuscripts in Press

- St John, P. C. *et al.* Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **150**, (2019).

Manuscripts in Preparation

- Wilson, St John, *et al*., Discovering Bio-privileged Materials with Machine Learning. In Preparation. (2021)
- Rorrer, Notonier, Knott, *et al.* Performance-advantaged nylon from bio-based β-ketoadipic acid. In Preparation. (2021)

Python Packages

- Neural Fingerprints
  - https://pypi.org/project/nfp/ (pip install nfp)
  - https://github.com/NREL/nfp
- Monomers to Polymers:
  - https://pypi.org/project/m2p/ (pip install m2p)
  - https://github.com/NREL/m2p
- Common-wrangler:
  - https://pypi.org/project/common-wrangler/ (pip install common-wrangler)