CITRINE INFORMATICS

White paper

DOMAIN KNOWLEDGE INTEGRATION

To accelerate development – to increase knowledge transfer

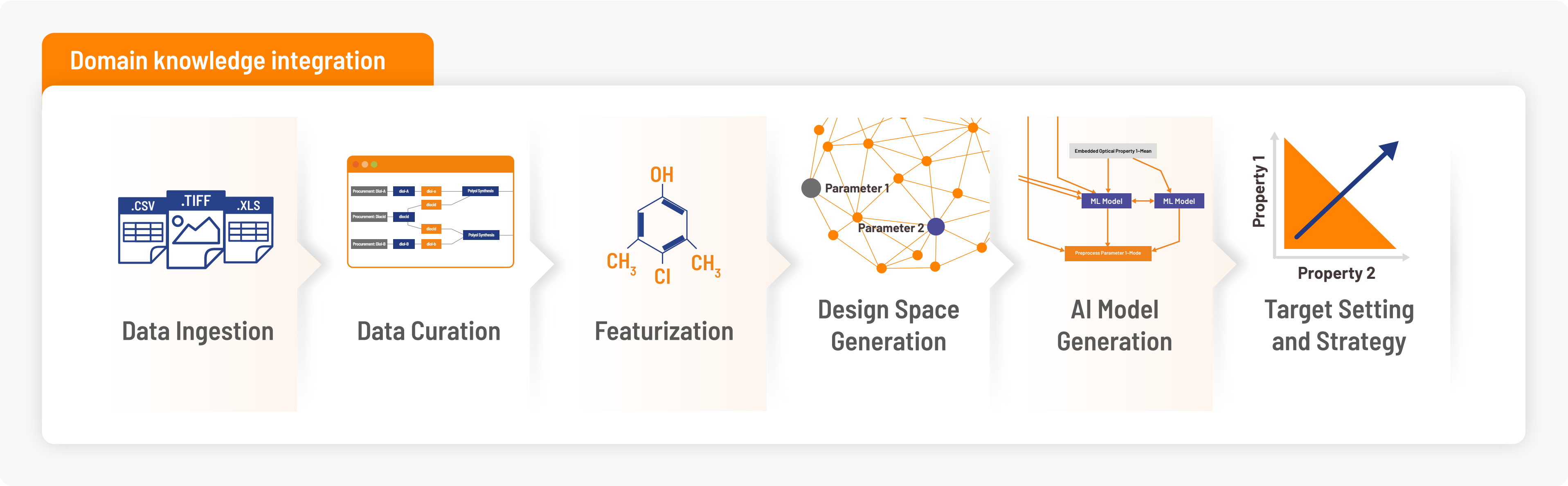
Domain Knowledge Integration

To accelerate development – to increase knowledge transfer

OVERVIEW

Scientists and engineers are your company's most valuable asset. Their scientific and business intuition is critical to developing new products, responding to customers and market trends, and training new hires. However, their knowledge is often locked in lab notebooks or in their brains, and their valuable insights are not shared across the organization. Domain knowledge integration (DKI) is the practice of capturing and institutionalizing these valuable insights, and is a key component to a digital strategy for R&D and applications engineering. It is critical that this knowledge is captured and leveraged across a digital workflow from data ingestion to target setting and research strategy.

Oftentimes, domain knowledge integration is synonymous with featurization, where aspects of materials and processes are represented in a way that is conducive to statistical analysis. Adding known physical properties or analytical relationships in this way to existing materials and chemicals data sets boosts the predictive accuracy of AI models. However, there are additional opportunities within an AI workflow to incorporate scientific and business insights. The Citrine Platform enables a holistic approach to DKI, where expert knowledge can be incorporated at each stage of the AI workflow.



For each of these stages, this paper will describe how the domain knowledge of your team can be leveraged to accelerate development, improve customer responsiveness, and disseminate valuable insights across your teams.

DATA INGESTION

Data ingestion is the process of importing historical data and setting up data pipelines to import future data.

Useful Knowledge in Your Team

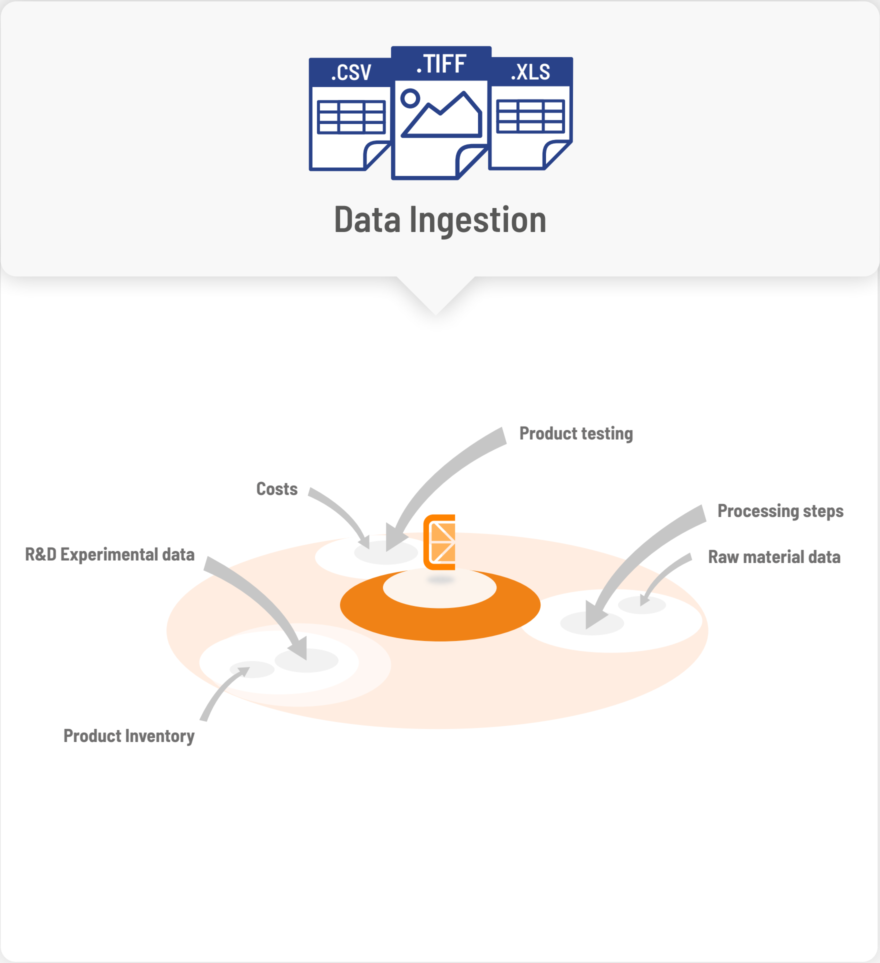
- What data is available and where can it be found?
- How relevant is the data to the project?
- How trustworthy are different data sets?
- Are there any known outliers that should be ignored? (E.g. contaminated samples)
- Which processing and characterization data is most valuable to future projects?

How this helps

This knowledge can identify data sources for new R&D and applications engineering projects and ensure that your scientists and engineers are structuring and sharing the most relevant and trustworthy data.

How the Citrine Platform harnesses this knowledge

Once the data has been digitized, researchers across your organization can find historical, relevant data and understand the context of the data (i.e. who created it, where it came from, how and when it was created, and under what conditions).



DATA CURATION

Data curation is the process of editing data so that it is structured, standardized, linked, and versioned.

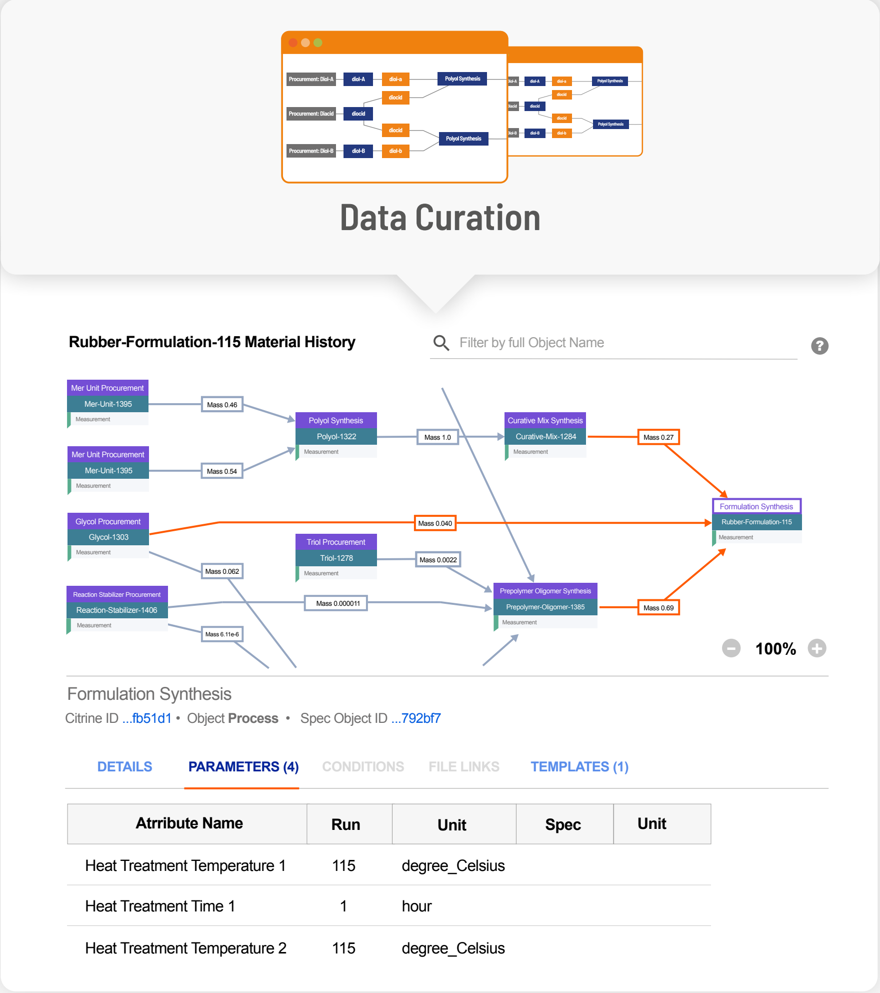
Useful Knowledge in Your Team

- Which metadata is necessary to give proper context to historical data? Which testing standard was used?
- Which data fields represent the same data coming from different sources? (E.g. T, Temp. Temperature)
- How does one data field relate to another? (E.g. sequential processing steps)
- How do test results change depending on raw material or process equipment variation? (E.g. batch variation)
- How are actual measurements and specifications compared?
- Are there distinct clusters in target property data? (E.g. candidate materials could be categorized into sub classes and modelled separately)
- Where does input data apply to all materials, and where does it only apply to a subset of materials?

How this helps

Integrating all this knowledge has three effects:

1. Data is discoverable and trusted by scientists across the organization, which allows a reduction in repeated experiments, and scientists on future projects can leverage past results.
2. Researchers can visualize entire material process histories from procurement to characterization, helping them build a mental model of previous work.
3. Data sets are clean and structured for AI modelling.



Data Curation

Rubber-Formulation-115 Material History

Formulation Synthesis

Citrine ID ...fb51d1 • Object **Process** • Spec Object ID ...792bf7

Atrribute Name	Run	Unit	Spec	Unit
Heat Treatment Temperature 1	115	degree_Celsius		
Heat Treatment Time 1	1	hour		
Heat Treatment Temperature 2	115	degree_Celsius		

How the Citrine Platform harnesses this knowledge

Data curation is supported on the platform through templating and the unique data model, GEMD. Customers can standardize on units and field names, which are enforced with templates to create consistent, interoperable data sets. The GEMD data structure is flexible, in that it allows new data fields to be added as the need arises, ensuring that all the context of complex material processes can be added and linked together. The graphical nature of the data diagrams allows scientists to visualize past experiments.

Read more about the GEMD data model here

FEATURIZATION

In a materials and chemicals context, featurization allows scientists to augment existing data sets with known physical quantities and relationships that may not be explicitly captured in existing data sets (e.g. bulk modulus, electron density, atomic polarizability, ionization energy, etc.). These quantities often improve the predictive performance of AI models. The Citrine Platform contains default feature libraries and allows teams to incorporate their own features into the AI workflow.

Useful Knowledge in Your Team

- What scientific information is correlated to material property measurements or performance targets? (E.g. particle size in a micrograph)
- How can we convert information to data? (E.g. converting the letters and numbers in a chemical formula to atomic weight or interatomic distance)

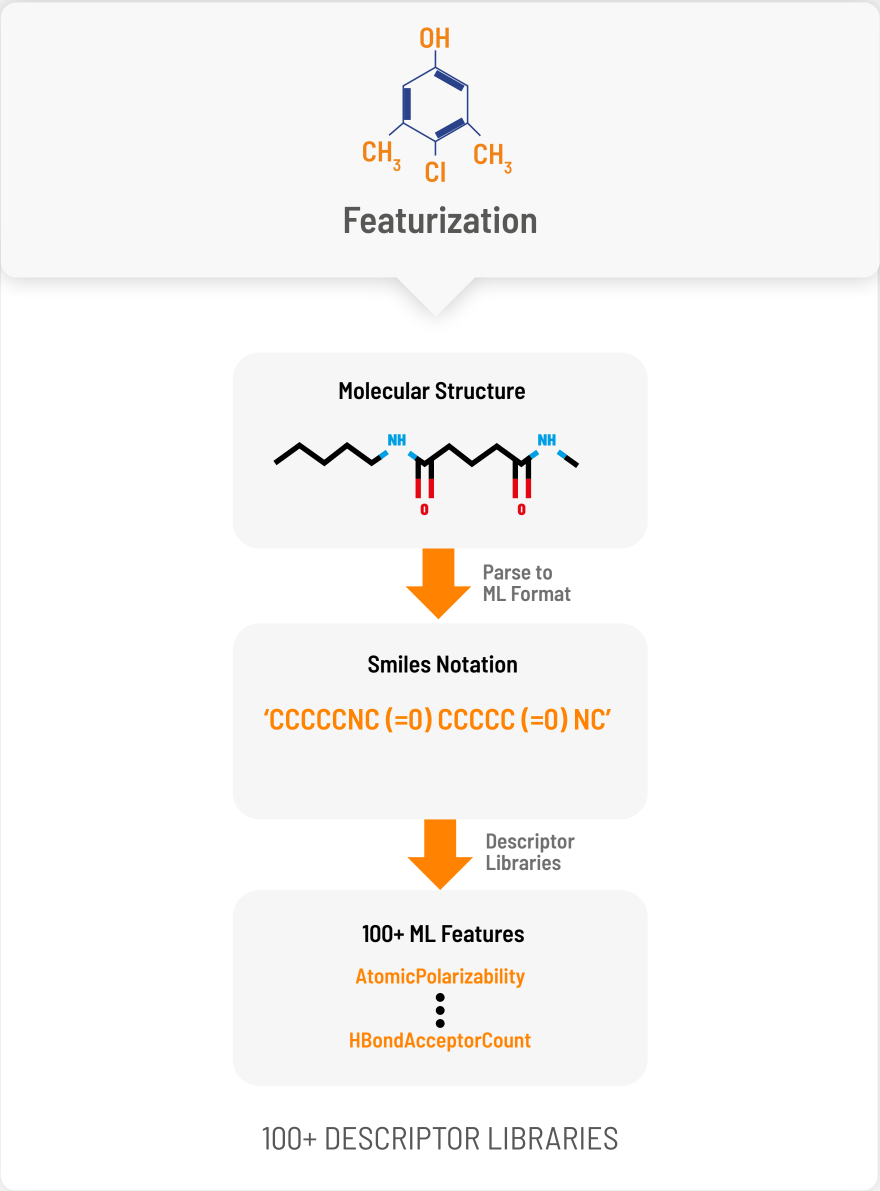
How this helps

Materials data sets are often sparse and it's expensive to produce more data through experimentation and characterization. Featurization improves model accuracy by the addition of known physical quantities to existing data sets, thus adding detail to data sets without additional lab resources. More accurate models mean fewer experiments to achieve your desired outcome.

How the Citrine Platform harnesses this knowledge

Over the last 7 years Citrine has developed a library of descriptors that convert materials- and chemical-specific information to data, and the Citrine Platform allows scientists to create custom features based on their domain or business need.

Read a case study where polymer featurization helped our customer improve customer responsiveness



Featurization

Molecular Structure

Parse to ML Format

Smiles Notation

'CCCCCNC (=O) CCCCC (=O) NC'

Descriptor Libraries

100+ ML Features

AtomicPolarizability

HBondAcceptorCount

100+ DESCRIPTOR LIBRARIES

DESIGN SPACE GENERATION

A design space is the set of feasible combinations of ingredients, recipes, and processing parameters for a given project. Design spaces can be generated by simply importing a list of materials that you would like to investigate, or by creating that search space through a set of rules or constraints. Product developers can use their understanding of the design space constraints – including ingredient compatibility, availability and cost, and processing parameters – to programmatically define the search space.

Useful Knowledge in Your Team

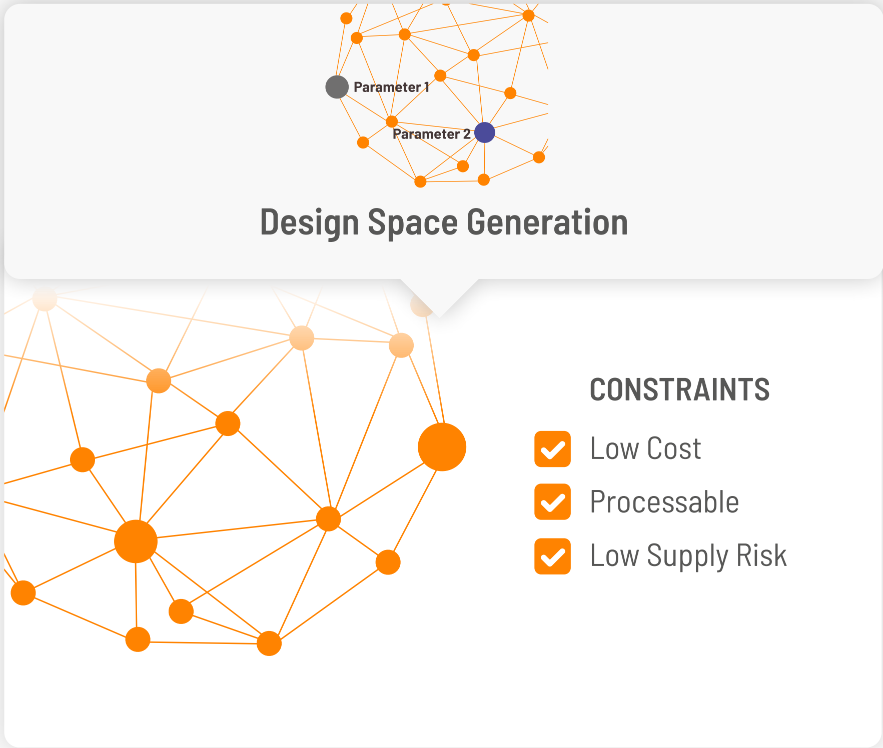
- Are you restricted to a specific set of ingredients or processing parameters based on cost or internal capabilities?
- What rules can be used to generate the design space?

How this helps

By setting sufficient constraints in your design space, you can guarantee that suggested candidates meet the physical, market, environmental, or customer needs of your project. Knowledgeable team members can prevent future problems by properly constraining the design space to materials that can be produced at-scale, not just on lab equipment. The size of the design space doesn't significantly affect the cost or timescales of the AI project, so you can broaden the design space if you're pursuing a more exploratory R&D project. This opens up the possibility of finding exceptional new candidate materials.

How the Citrine Platform harnesses this knowledge

On the Citrine Platform, design spaces are a reusable asset for future projects across your organization. Design spaces can also be visualized to give your scientific team the opportunity to understand how project constraints impact the possible set of new materials. Reusing a design space can prevent constraints being missed and work carried out on unfeasible material candidates.



Read a case study where Panasonic used Citrine's design space capabilities to patent new semiconductor molecules

DESIGN SPACE GENERATION (cont.)

Easy to interpret Design Space Visualizations

PREFERRED INGREDIENT LIST

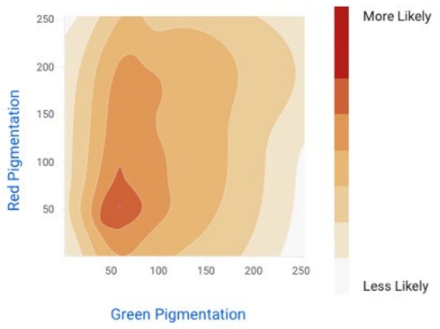
- Low cost
- Bulk bought
- Low supply risk
- Low EHS concerns

SPECIAL INGREDIENT LIST

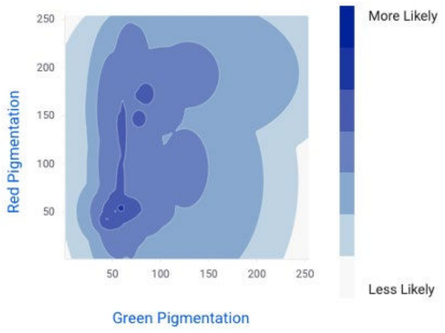
CORE DESIGN SPACE

SPECIAL DESIGN SPACE

Core design space



Special design space



Probability of achieving property with one or more of the candidate materials

AI MODEL DEVELOPMENT

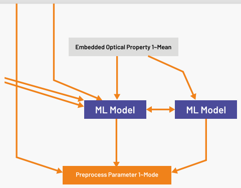
AI model development involves selecting the best machine learning algorithm(s) to use, which data the model should train on, and linking different models together to provide the best predictive power for the problem at hand. For example, you can use two models in conjunction to represent how a property is related to both micro- and nano-scale features. With Citrine s hierarchical modelling approach, you can also incorporate known physical or analytical relationships via equations, so machine learning does not have to relearn known physics.

Useful Knowledge in Your Team

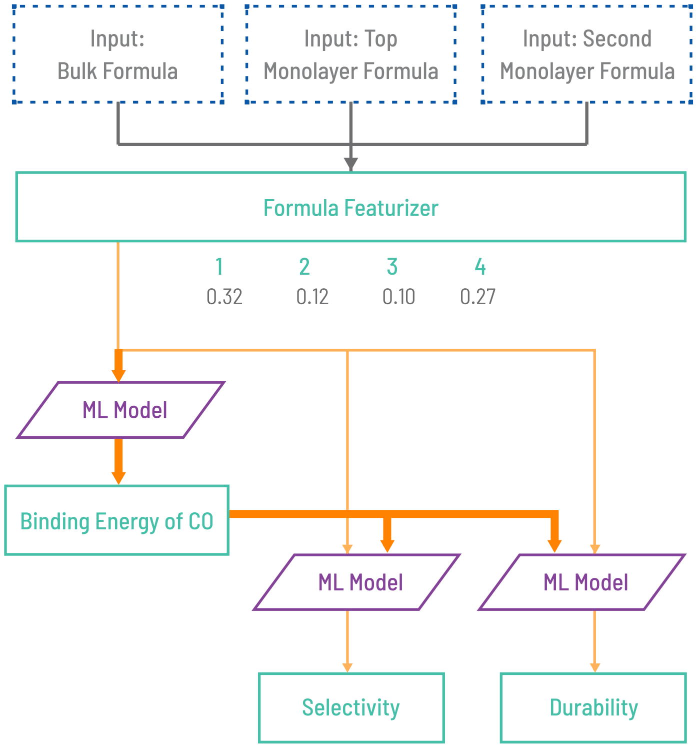
- Which data fields are correlated?
- Can we exclude extraneous data from our modelling approach to reduce noise?
- Are there intermediate relationships between your inputs and the target properties?
- Can your materials be categorized into groups that behave similarly?
- Are there equations we can incorporate into our modelling approach that represent known structure <> process <> property relationships?
- Are there distinct mechanisms that impact target properties depending on certain ingredients or process parameters?

How this helps

By leveraging this information, the AI models predict only on the unknown parts of the problem, improving model accuracy and predictive ability.



AI Model Development



GRAPHICAL MODELLING

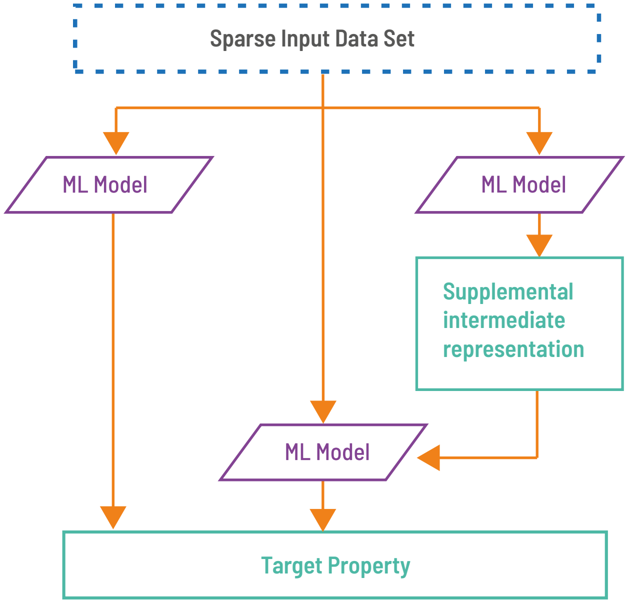
AI MODEL DEVELOPMENT (Cont.)

How the Citrine Platform harnesses this knowledge

Citrine uses a graphical modelling architecture which has two advantages:

1. It unlocks the ability to combine analytical formulas and machine learning models in the workflow. In the example on the right, each purple shape represents an AI model or an expert-provided formula. Each link represents a known relationship or correlation. This hybrid approach means that you can build accurate models with less data, efficiently using all the information available to you.
2. Researchers can partner with data scientists to ensure that the model is using all available domain knowledge, because they can see and understand graphically how the model is constructed. This also aids new members of the team to get up to speed quickly.

The other way that the Citrine Platform can harness this knowledge is through transfer learning. When a data set is sparse, it might be possible to improve model accuracy by first modelling a related intermediate property. For example, a target structural property could be known to be related to another structural property for which there is enough data to train an accurate model.



TRANSFER LEARNING

TARGET SETTING AND STRATEGY

Target setting describes the set of properties or performance that the end customer or R&D team is looking for in the new materials. This is often a combination of constraints and objectives, but should always lead to a logical scoring system by which candidate materials can ultimately be ranked. From there, a set of candidates are selected to be tested to validate the model's accuracy and provide more data. These are chosen using a selection strategy.

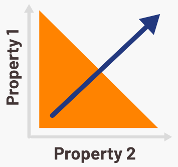
Strategy in this context relates not just to how to use the platform in the best way, but also using information provided by the Citrine Platform to guide both R&D strategy and customer negotiations.

Useful Knowledge in Your Team

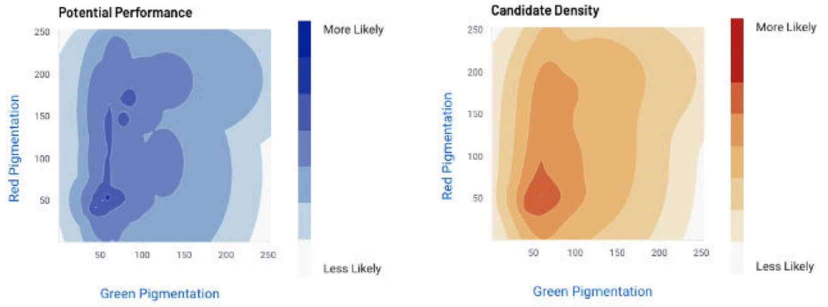
- What is the full set of constraints and objectives demanded by the end customer or the market?
- What is the relative importance of these objectives? Are there intermediate relationships between your inputs and the target properties?
- What is the deadline of the project?
- How many candidates can sensibly be tested in the same batch?

How this helps

This knowledge ensures that the right properties are targeted, and the right selection method is used. Early on, it may be important to test a diverse selection of candidates. Later on in the process, it could be valuable to focus on more promising, higher-certainty candidates. It also makes sure that the number of candidates chosen to test fits the batch size capabilities of testing equipment.



Target Setting and Strategy

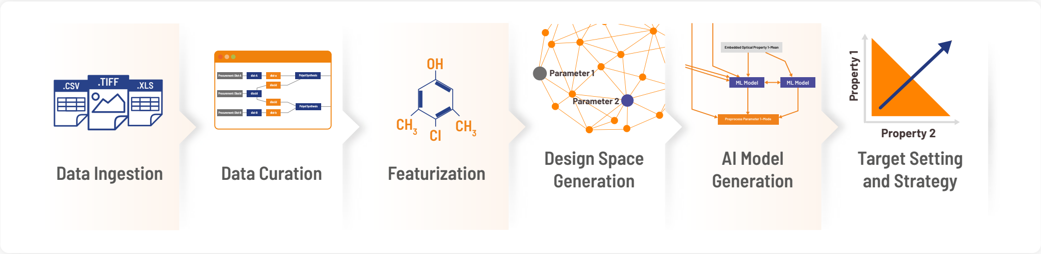


How the Citrine Platform harnesses this knowledge

The Platform allows the researchers to define sophisticated multi-objective targets and choose the candidate scoring strategy from a drop-down list.
The Citrine Platform also uses patent-pending technology to visualize both the number of candidate materials that are predicted to hit certain properties and also the likelihood that one or more of the candidates will hit those properties. If lots of candidates are predicted to hit target properties, this should be a more promising research direction. If there is only one candidate that is predicted to hit target properties, but the likelihood of it doing so is high, then it may also be reasonable to invest in synthesizing and validating that candidate. (see diagram)

This objective data can be compared across research strategies and across portfolios of research projects, to ensure that valuable resources are being assigned to projects with the highest likelihood of success.

SUMMARY



Data Ingestion → Data Curation → Featurization → Design Space Generation → AI Model Generation → Target Setting and Strategy

The Citrine Platform captures, leverages, and shares domain knowledge throughout the full AI workflow. This functionality was consciously designed so that the valuable experience of your team could be exploited to its fullest extent. Once integrated into the platform, your data and your team s knowledge are digital assets that provide you with a competitive advantage. They can be reused on future adjacent projects and should be protected like your other IP assets.

Make sure that whatever Materials Informatics System you use, it captures and leverages all your team's knowledge.

Arrange an online meeting

Listen to our podcasts

Download more case studies

Subscribe to our newsletter

Citrine Informatics Inc.
2629 Broadway St
Redwood City, CA 94063

citrine.io
info@citrine.io