CITRINE INFORMATICS

White paper

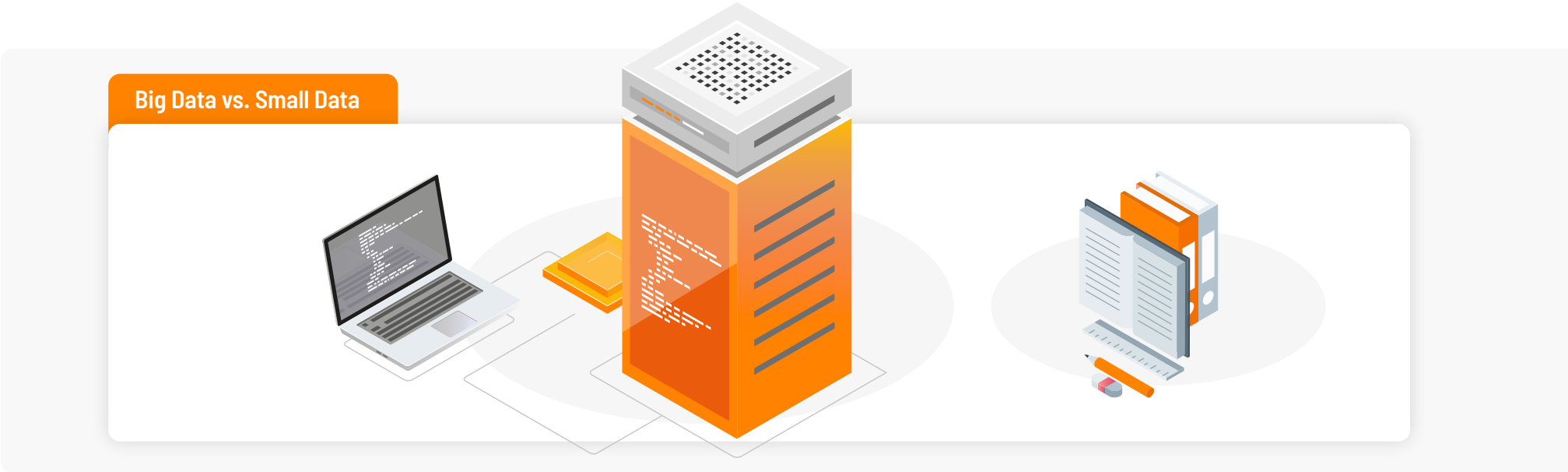
# CHALLENGES IN MACHINE LEARNING FOR MATERIALS AND CHEMICALS *- AND HOW TO OVERCOME THEM*

This white paper summarizes the differences between Machine Learning for Materials and typical AI applications. It explains how Citrine has overcome these difficulties and why off-the-shelf open source AI will require a lot of tailoring to make it work in this space.

# Challenges In Machine Learning For Materials - *And How To Overcome Them*

## CONTENTS



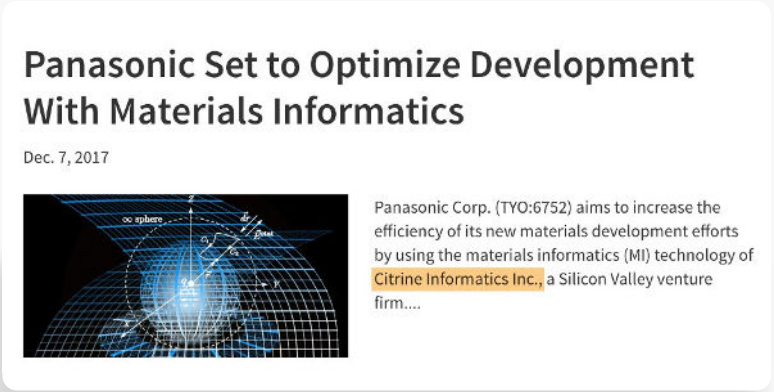


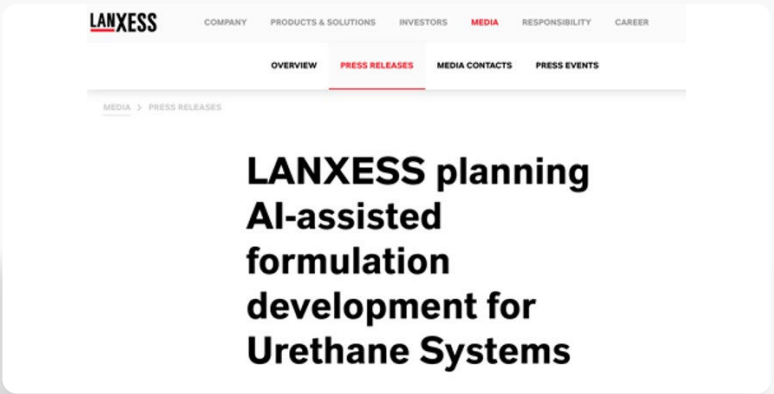
## INTRODUCTION

Companies in many industries are successfully applying machine learning (ML) or artificial intelligence (AI) as an integral part of their digital transformation initiatives. In materials science and chemistry, AI application development began a few years ago with companies such as Toyota and Corning establishing internal AI centers, and other big players such as Panasonic, and Lanxess partnering with Citrine Informatics to pioneer data-driven R&D projects. These companies found that AI can provide a positive return on investment when applied to new product development and are now moving from one-off projects to deploying an enterprise wide AI infrastructure. This shift enables greater autonomy and scalable benefits across an entire organization.

**THIS ISN'T AS STRAIGHT FORWARD AS IT SOUNDS!**

Machine learning for materials and chemicals is more complex than typical applications, and therefore requires a more nuanced approach at many points within the ML workflow. This paper outlines the differences between more mature AI applications (e.g.: business intelligence, consumer insights) and the application of AI in materials and chemicals. It suggests best practices for how to tackle these challenges based on Citrine's work with its partners over the last 5 years.



## DATA TYPE

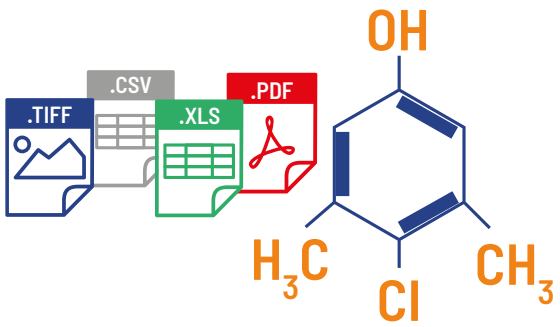
AI applications like consumer behavior and business intelligence obtain data from structured sources like forms, surveys, and website/app analytics tools. In the materials and chemicals industry, data come from many different sources – test data, simulation data, reference data, supplier data sheets – and in different formats – microstructure images, processing equipment, chemical formulas, spectral data, etc. This leads to datasets that are often highly heterogeneous and inconsistently structured. Integrating data from different sources, e.g. the processing lab and the testing lab - requires a deep understanding of materials, unit interoperability, and detailed information on the pedigree of the data.

To use this diverse data, materials and chemicals companies need two things:

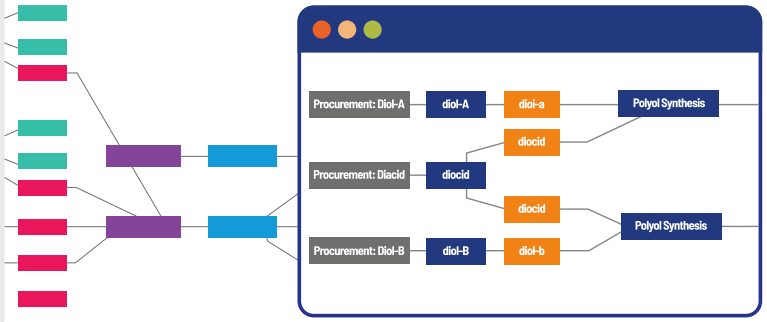
1. An easy way to get legacy data into a centralized database
2. A way to structure the data that has the right mix of flexibility and standardization.

Citrine has also developed a data model – Graphical Expression of Materials Data (GEMD) – designed to capture the history of a material, including raw ingredients, processing steps, testing parameters etc. It is visualized on the platform in a way that helps researchers relate the digitized version of the material to the material itself.

### Quickly ingests diverse file types




### Visually represent the entire material history



## DATA VOLUME

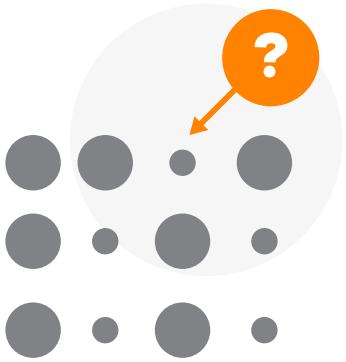
"BIG DATA" is a phrase commonly associated with AI and machine learning. Companies with millions of customers or thousands of sensors can quickly and cheaply amass billions of data points, in a controlled way, with datapoints consistently populated. In Materials and Chemicals, each "row" of data can cost months of time and tens of thousands of dollars to obtain. "Complete" datasets are often small (100+ rows in a data rich scenario) and incomplete. Suppliers, for example, may not provide a data point that is unfavorable for their product, leaving a hole in the data set. A machine learning approach for materials therefore has to work with small, sparse datasets.

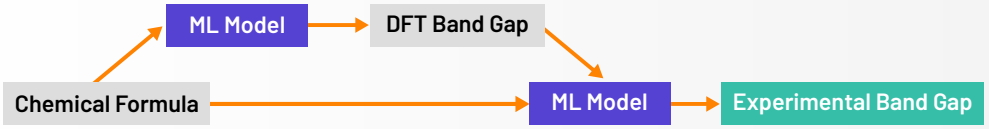
**Transfer learning, where information from one dataset is used to inform a model on another, can be an effective tool for bridging sparse data while preserving the differences in the underlying measurements. Citrine scientists explore different transfer learning techniques in the following paper:**

**Overcoming data scarcity with transfer learning**

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., & Meredig, B. (2017).
arXiv preprint arXiv:1711.05099.

**MI Models Can Be Nested – One used to predict a property which is then used as an input to another model**



## ESTABLISHED DOMAIN KNOWLEDGE

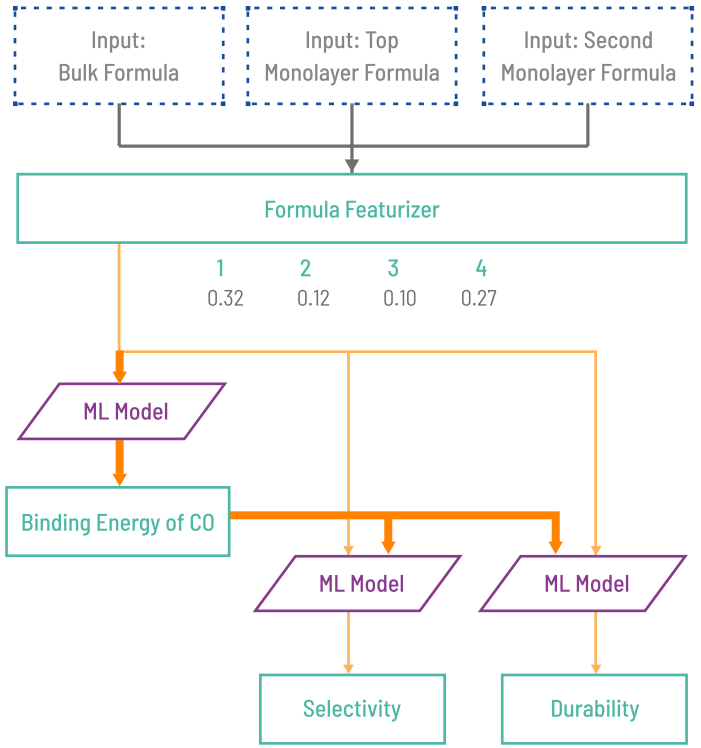
The laws of physics and chemistry are not applicable in typical machine learning applications. Data scientists in consumer or business intelligence applications can rely on large datasets and "out of the box" algorithms to analyze data and uncover new insights. In the materials and chemistry space, with sparse datasets, it is critical to integrate domain knowledge into machine learning workflows. The first place this happens is when choosing which candidate materials to predict properties for. This can be based on a database of candidates from previous work, or a set of rules can be created that reflect scientific knowledge of which candidates are physically possible and likely to be good. Domain knowledge can also be built directly into an ML model. Known analytical relationships and correlations can be built into a model to increase accuracy, leading to better predictions and quicker R&D results.

Citrine works with customers' experts to codify their knowledge into the Citrine Platform by:

- Building physical, chemical, and commercial constraints into design spaces
- Encapsulating domain knowledge into ML models through featurization, empirical, or analytical calculations.

As the platform is modular and can be shared by different R&D teams or business units in an enterprise, captured domain knowledge can be reused across projects.

**Known relationships between binding energy and durability have been incorporated**



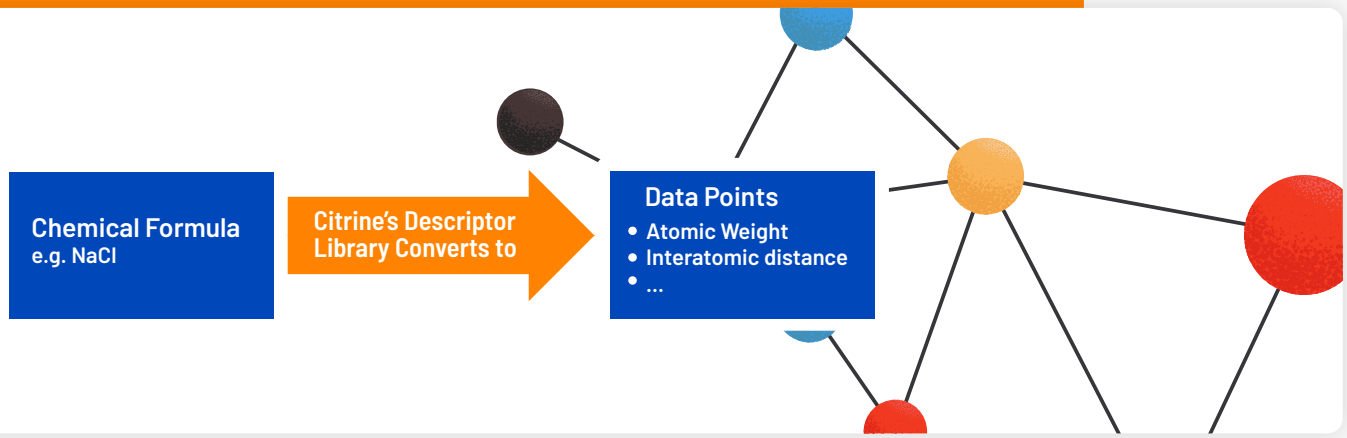
## DATA REPRESENTATION

In traditional applications, data types are often very simple and can be easily interpreted using algorithms. For example, the interval between two date fields can be converted into a length of time. Materials data, on the other hand, is more complex. Letters and numbers are not important in a chemical formula (e.g., "H" "2" "O") but understanding what they represent physically (two atoms of the element hydrogen bonded to one atom of the element oxygen in a molecule) is critical to analysis. A microstructure image is not a parameter in and of itself; the interesting feature of it may be the fineness of the lamella structure, or the average size and density of precipitates.

Data must be converted into a series of "descriptors" that represent the features of the data that are important before they can become a parameter in a machine learning model. Defining and expressing these descriptors requires deep domain knowledge and is critical to the success of a project.

Citrine has developed a set of descriptor libraries that turn common materials and chemical data types into a set of descriptors. For instance, it can understand chemical formulas and convert them into descriptors such as DFT energy density or interatomic distance. The platform enables users to define key features for their materials applications.

**Descriptor libraries can convert information e.g a chemical formula into data e.g. atomic weight**



## PREDICTION TASK

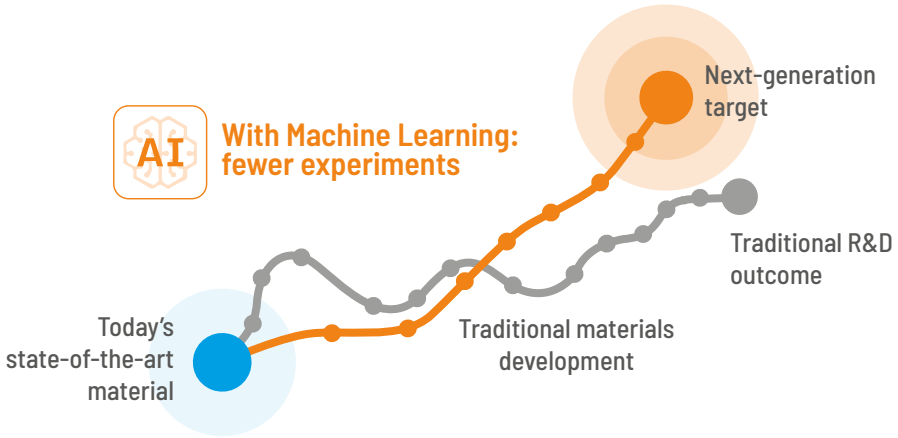
Artificial intelligence is great for pattern recognition. Typically, a model is trained on a large dataset of known information, and the model accurately predicts into which category a new example fits. For example, given a large set of cat and dog photographs, can a model accurately predict whether the next photo is a dog or a cat? R&D scientists are not interested in finding existing, common cases, but rather exploring higher performing "outlier" materials that push the limits of existing properties.

The Citrine Platform tackles this challenge with a method called sequential learning. An initial (and oftentimes sparse) dataset is used to train an ML model that suggests which candidates to make and test next. New experimental candidates iteratively improve the accuracy of the model and get closer to the desired property/performance profile.

The model enables extrapolation by taking advantage of uncertainty estimates for each property prediction. These uncertainty estimates allow the model to make calculated bets about where to explore. This process ensures that experiments are selected with data-driven insights rather than trial and error, reducing the number of experiments, time, and cost to meet R&D goals. Learn more about the sequential learning methodology in this paper:

**High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates**
Ling, J., Hutchinson, M., Antono, E., Paradiso, S., & Meredig, B. (2017). Integrating Materials and Manufacturing Innovation, 6(3), 207-217.

**Data-driven research requires fewer experiments**



## SAMPLE BIAS

Sample bias in datasets is common across all AI applications, including materials and chemicals. However, there are two unique forms of sample bias in Materials and Chemistry data that must be overcome:

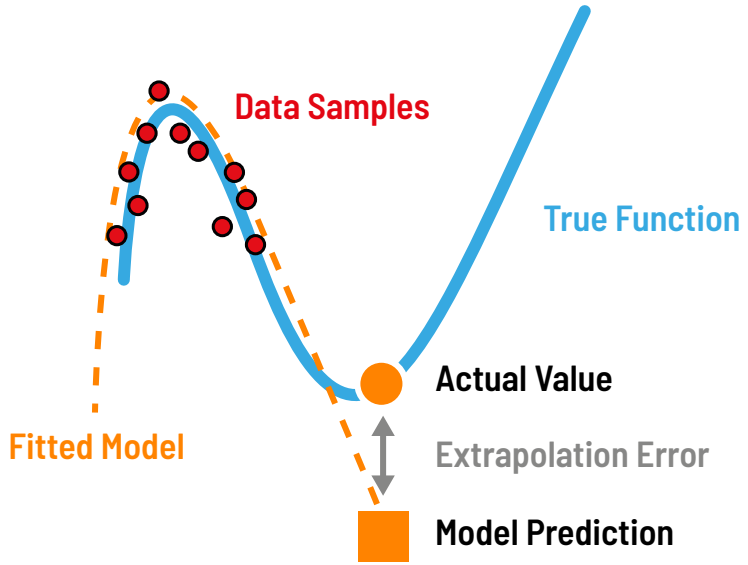
1. Lack of "failed" data – Machine learning models require data that includes a range of measurements – failures as well as successes. No matter how advanced the algorithm, it cannot evaluate whether a candidate material will meet performance criteria if it does not have a distribution of successes and failures in the training data. R&D data and scientific publications tend to bias towards successful results, and data from previous failed experiments may not be captured.
2. Process and measurement variations – Processing equipment can cause slight variations in sample measurements, even at the same composition. These variations can be skewed by equipment choice, measurement technique, or the scientist performing the experiment.

Materials and chemicals companies can guard against sample bias by:

- Testing and measuring reference samples to normalize results
- Utilizing materials-specific cross validation techniques when training ML models (link to paper) to avoid overfitting to biased data
- Treating data from successful AND failed experiments as corporate assets to encourage the dissemination of valuable information for future R&D efforts

**Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery.**
Meredig, B., Antono, E., Church, C., Hutchinson, M., Ling, J., Paradiso, S., ... & Mehta, A. (2018). Molecular Systems Design & Engineering, 3(5), 819-825.

**Biased data can lead extrapolation astray – sequential learning can be used to explore gaps**

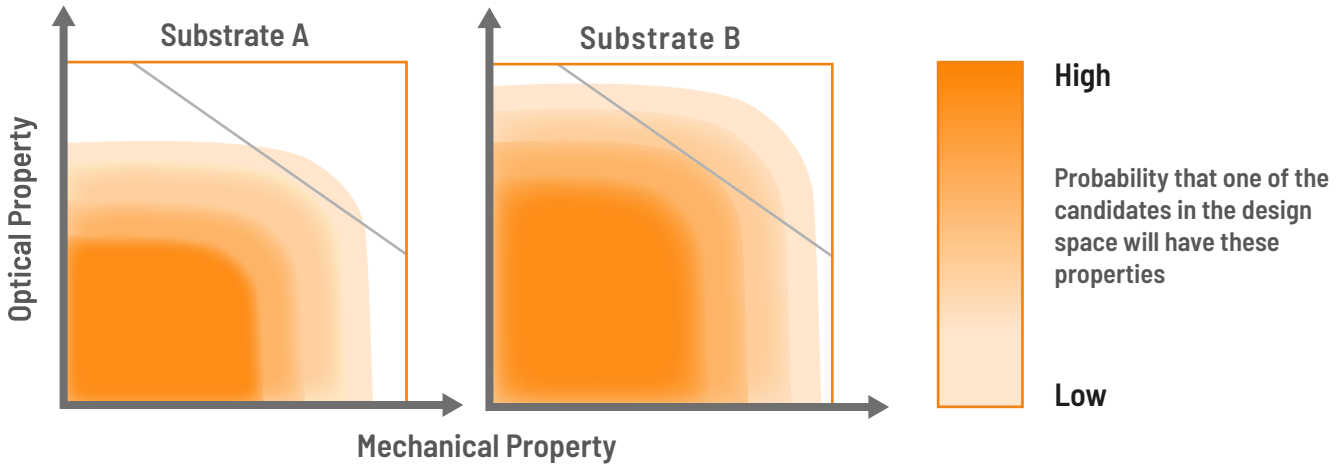


## UNCERTAINTY IN DATA AND MODELS

In many commercial machine learning applications, prediction uncertainty has little consequence. For example, if there is a +/- 0.75-point uncertainty in the 0-5 scale used to recommend movies for a movie streaming service, this would have little impact on what movie a user would watch next. However, a 15% uncertainty in a material's property could determine the next trial for a costly experiment. Uncertainty helps scientists understand, interpret, and act on model predictions.

**The Citrine Platform includes uncertainty quantification in each model and each prediction to help scientists interpret ML models and choose the best experiments. Visualizations showing the likelihood of achieving performance targets can guide R&D direction and project selection.**

**Visualization of probability of success helps researchers decide on their research direction**



## INTERPRETABILITY

As R&D is digitized in the materials and chemicals industry, it is important for scientists to scrutinize, sanity-check and learn from their models. In other words, a typical "black box" machine learning software is not fit for purpose.

In order to improve model interpretability, Citrine's Platform:

- exposes relationships between input parameters and predicted values so that scientists can see how the model arrived at its predictions.
- visualizes data to display how input data, analytic relationships, computational data, featurization and machine learning techniques are used to build a model.

Exposing these aspects of a machine learning workflow ensures that researchers understand and act on data-driven insights.

**The importance of different features is displayed**

Feature Importance — See all (15)

| | Name | Importance |
|---|---|---|
| 1. | Elemental polarizability | 0.32 |
| 2. | Shear modulus | 0.12 |
| 3. | Trouton's Ratio | 0.10 |
| 4. | Ratio of electron affinity to electronegativity | 0.07 |

## SUMMARY

Materials and chemicals data present many unique challenges not found in traditional AI applications. These challenges demand a tailored approach. Citrine has developed and proven its materials and chemicals-specific AI approach across many materials and chemicals classes, demonstrating that the Citrine Platform can help its partners accelerate product development, use data to guide R&D strategy, and facilitate knowledge transfer and AI assets across an organization.

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