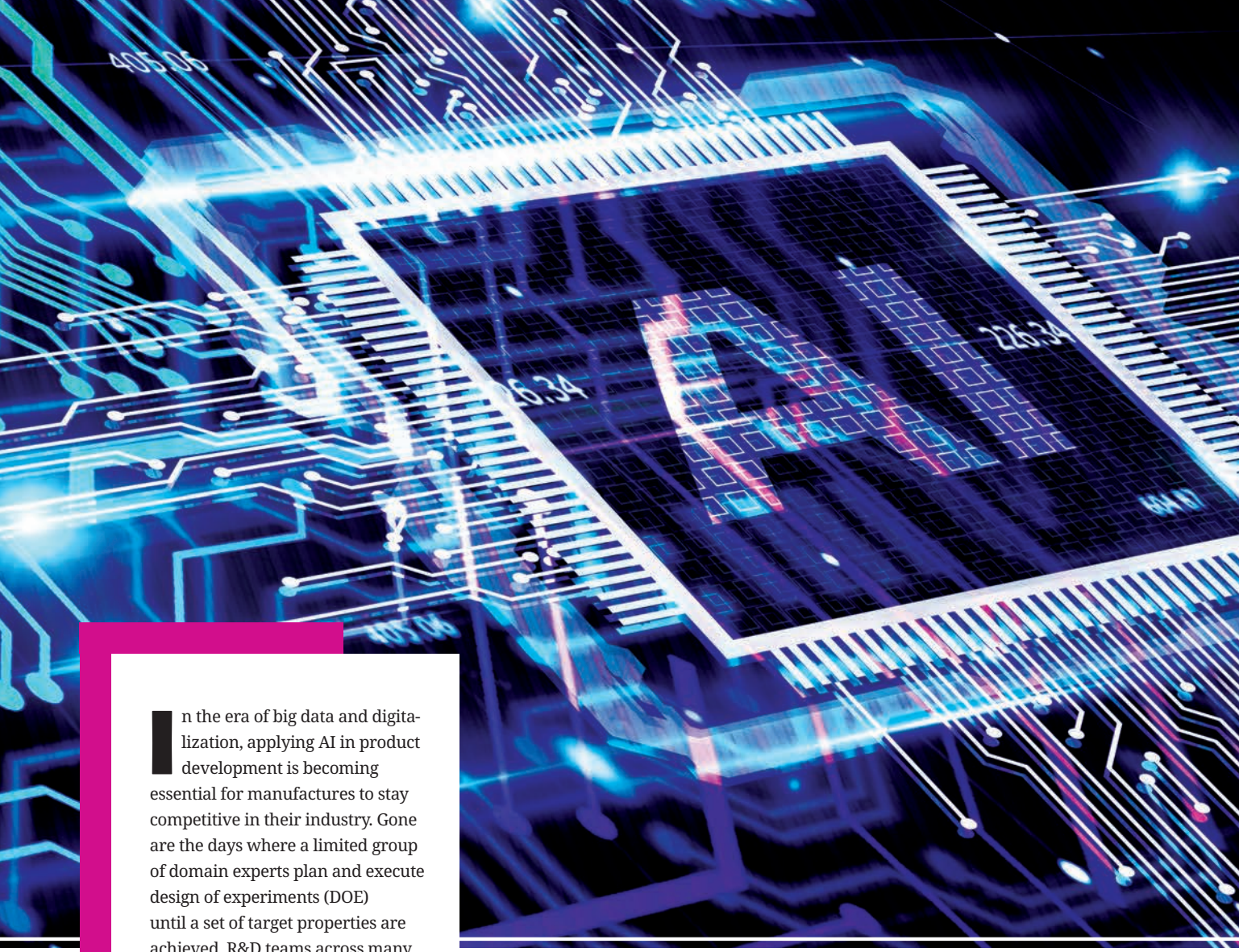


In the era of big data and digitalization, applying AI in product development is becoming essential for manufactures to stay competitive in their industry. Gone are the days where a limited group of domain experts plan and execute design of experiments (DOE) until a set of target properties are achieved. R&D teams across many industries are transforming their workflows by embracing AI initiatives to intelligently glean information from their historical, experimental, and product databases.

AI models built on past data can be used to reduce experimentation through simulation and achieve optimal properties much faster using numerical optimization. In this article, a high-performance coating dataset will be used along with ProSensus' FormuSense software. The article will examine the typical steps to effectively utilize R&D datasets including:

1. Data preparation to ensure that the dataset is appropriate for model building. Common pitfalls such as missing data, insufficient variation, and data anomalies are investigated and resolved.
2. Model building using multivariate analysis and its intuitive plots that help subject matter experts interpret their dataset highlighting key correlations and trade-offs.
3. Simulation using the model to predict the outcome of potential experiments without running the actual physical experiment.
4. Constrained numerical optimization to calculate the ideal formulation and processing conditions required to achieve a targeted quality or set of qualities.

## Introduction

Coatings are commonly used in a variety of industries to enhance the performance, durability, and finish of painted surfaces. Coatings manufacturers can address rising raw material costs, unwanted inventory costs, supply chain shortages, and process complexities by reformulating existing products with fewer, cheaper, and more readily available ingredients. However, coating formulations are complex; formulators must carefully select numerous ingredients in hopes of achieving multiple performance targets in robust application conditions, while balancing cost and ingredient availability, and adhering to environmental regulations.

## Product Development Data

Typically, a product development dataset includes three types of input (X) variables:

# Using AI to Rapidly Develop New and Improved High-Performance Coatings

By Marlene Cardin, Kristin Wallace, and Alexander Nguyen, ProSensus

- Formulation ratios—how much of each ingredient is used (measured in percentage, fraction, or other quantity, such as kg)
- Ingredient properties—physical or chemical properties that characterize each ingredient, such as molecular weight or density
- Process conditions—the manufacturing conditions under which the ingredients are combined, such as temperature or mixing speed

The output (Y) variables in a product development dataset include any measured quality or performance properties, such as viscosity or hardness.

While a large number of X and Y variables may exist for one product development dataset, there are typically much fewer degrees of freedom due to the underlying chemistry. In other words, not every X variable can move independently of all other X variables, and not every Y can move independently of all other Y variables.

## Predictive Modeling

Owing to these chemistry complexities, product development activities that follow a conventional approach (involving trial-and-error or design of experiments (DOEs) with numerous physical experiments), tend to become iterative and resource-intensive. By contrast, building a predictive model on historical formulation data, interpreting the model, and performing simulations (applying the model to predict the outcome of new formulations) can drastically reduce the number of required physical experiments and, ultimately, accelerate the time required to develop a new or reformulated product.

### PLS for Product Development

The AI framework for a predictive model should be carefully selected. PLS (partial least squares) is well suited to product development applications because results are visualized with intuitive plots (explainable AI), and the models can be inverted in a constrained optimization framework. In addition, models can be built on both large and small datasets, and missing data and correlated measurements are handled.

The ability to interpret PLS models is a significant advantage over black-box methods because formulators can:

- build trust in the model by verifying existing domain knowledge;
- uncover new learnings from the intuitive plots; and
- use the identified correlations to more intelligently and efficiently plan physical experiments.

## Constrained Optimization

The ability to invert a model in a constrained optimization framework represents a further advantage of selecting PLS for product development applications. In this framework, the PLS model is used in conjunction with a relevant objective function as well as bounds and constraints to determine the required ingredient and process combinations that will achieve specified performance property targets.

Examples of typical constraints for a product development optimization problem include:

- availability of each ingredient
- upper and lower bounds for how much of each ingredient can be used
- limits on the number of ingredients used per formulation

The optimization objective function for a product development application typically includes terms to:

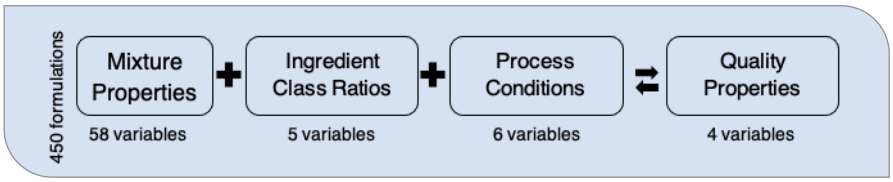
- minimize overall formulation cost
- minimize deviation from performance property targets
- minimize extrapolation from historical design space

## Coatings Case Study

The goal of this project was to reformulate existing high-performance coatings to use fewer ingredients while maintaining existing performance property (i.e., quality). The available data included:

- 450 historical formulations
- 81 unique ingredients from 5 ingredient classes (resin, solvent, catalyst, additives, isocyanates)
- 116 ingredient properties (not all available for each ingredient)
- 6 process conditions (such as ambient conditions and film thickness)
- 4 performance (quality) properties, including viscosity, hardness, glass-transition temperature, and gel fraction

**FIGURE 1**
**PLS model structure.**




## Assemble Dataset

An important first step is to assemble the dataset. The goal is to evaluate the suitability of the structured data for predictive modeling and to calculate advanced input features such as mixture properties.

Examples of data assembly tasks include:

- correcting common anomalies (such as outliers, inconsistent measurement units, inconsistent nomenclature, etc.)
- calculating ingredient class use (summation of formulation ratios for all ingredients in a single ingredient class)
- assessing variation in ingredient properties, ingredient use, process conditions, and quality properties
- evaluating the amount of missing data

### *Mixture Properties*

Mixture properties are a key concept of the PLS model structure. Mixture properties are calculated by combining ingredient properties and formulation ratios using appropriate mixing rules.

When mixture properties are available and sufficiently characterize the ingredients, formulation ratios can be excluded from a model. This variation of the model structure is very powerful, as it allows for new ingredients to be considered in future formulations, so long as the ingredient properties are known.

Mixture properties can be calculated per ingredient class or globally (across all ingredient classes). This flexibility is helpful in cases where some ingredient classes have more missing data than others, when the ingredient property measurements differ between ingredient classes, or when the formulator only wishes to consider new ingredients in specific ingredient classes.

In this dataset, 4 of the 5 ingredient classes contain ingredient properties; therefore, mixture properties were calculated per class, using a linear weighted-average mixing rule. This resulted in 58 new input (X) variables that contained sufficient variation for use in the model.

## Build & Interpret PLS Model

With the assembled dataset, the next step is to build the PLS model. The PLS model structure for this dataset is shown in **Figure 1.**

In addition to selecting X and Y variables, the 450 formulations are separated into a training group and a validation group. In this case, every fifth formulation was added to the validation group. The purpose of creating a validation group is to evaluate model fit on "unseen" data.

A number of plots can be analyzed from the PLS model to assess model fit, identify outliers, and understand key correlations. Some examples of these results and interpretation follow in **Figures 2 and 3.**

As evident, some of the quality variables in this coatings dataset have a better model fit than others. Gel Fraction has the poorest fit, which was expected, owing to measurement noise.

As evident, one of the resin mixture properties ("Res_Tg," colored white in the coefficients plot) has a strong, positive correlation with Glass Transition Temperature. This finding was confirmed by reviewing a conventional *observed versus predicted* plot alongside a *line plot* of this particular resin mixture property **(Figures 4 and 5).**

Logical and expected results, such as the previous example, can be further validated by consulting with domain experts. When a PLS model reveals unexpected results, running physical experiments to confirm the findings is recommended.

## Simulate New Experiments

The developed PLS model was then used to predict the outcome of replacing a key ingredient (Resin #8), because it would no longer be available to the manufacturer due to supply chain issues. This ingredient had been used in 35% of all historical formulations.

Several digital simulations were completed to predict the outcome of replacing Resin #8 with 4 alternative resins (Resin #21, #52, #50, #40) for one key product.

**FIGURE 2**
**R2Y-This plot summarizes how well the model captures variation for each quality variable. Note that "Final Model" includes training and validation.**

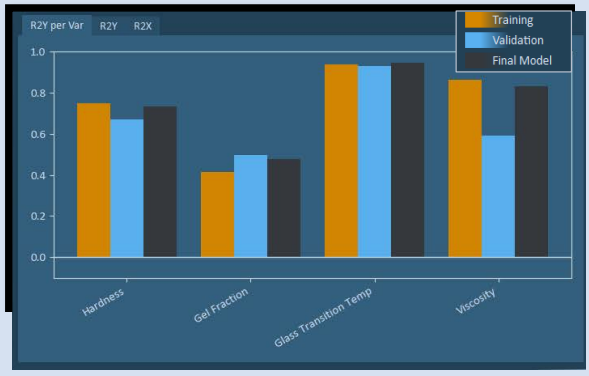
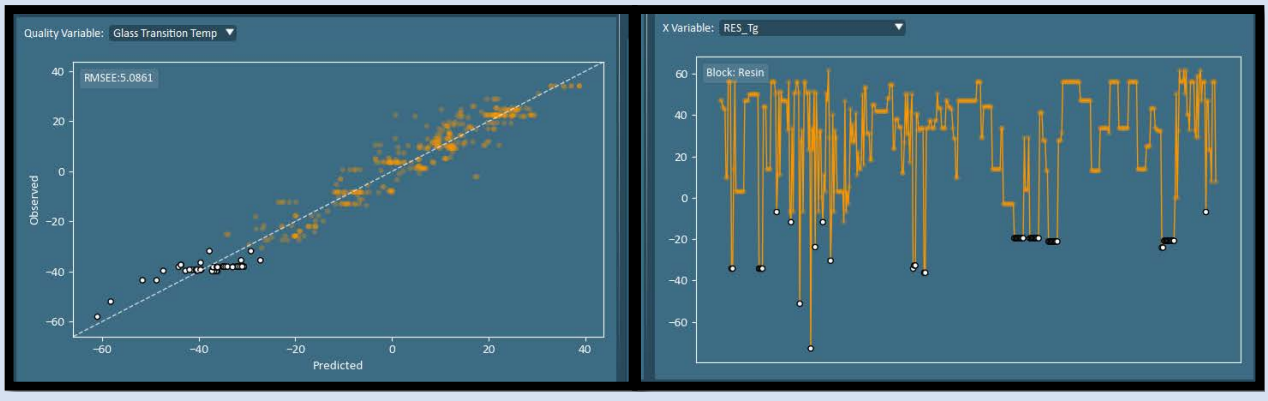



**FIGURE 3**
**Coefficients-This plot summarizes the multivariate impact all X variables on each Y variable.**


**FIGURES 4-5**
**Observed versus predicted plot alongside a line plot.**




To assess which alternative resin was best suited to replace Resin #8, several factors were considered, including predicted quality, scores, model validity, and cost.

### *Predicted Quality*

The plot in **Figure 6** compares each alternative resin to the original in terms of predicted quality of the final coating product (as the number of standard deviations from model average). Because the goal was to reformulate an existing product, the alternative resins that result in similar quality to the original product are preferable.

Another consideration is to ensure that the scores (latent variables) of the simulation scenarios do not diverge too far from those of the original formulation. In PLS, the number of variables in a dataset are

**FIGURE 6**
**Alternative resins compared to the originals in terms of predicted quality of the final coating product.**

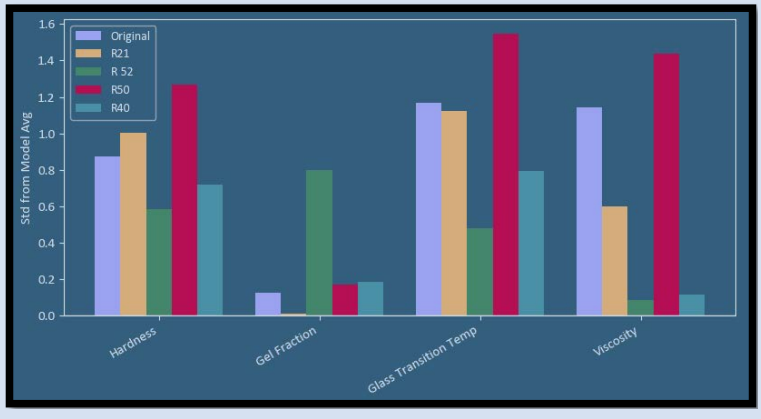
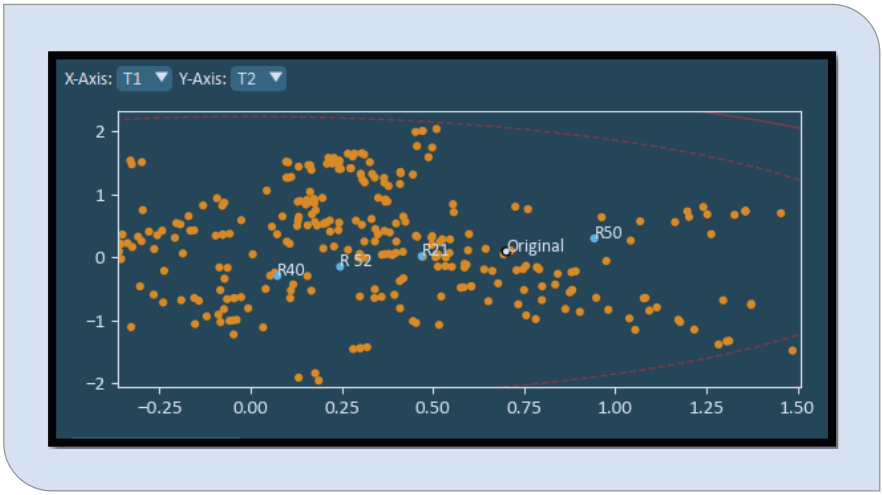


**FIGURE 7**
**A score plot identifies similarities in formulations.**




reduced to a small number of independent factors—called latent variables or scores—that explain the most variance. A *score plot* is thus used to identify how similar formulations are to one another, as shown in **Figure 7.** Formulations that cluster together on the score plot have X and Y data that are very similar.

Much like quality, alternative resins that result in scores similar to the original coating formulation are preferable.

### *Model Validity*

To ensure that the PLS model is not extrapolating beyond the known design space, it is important to calculate two model validity metrics called Hotelling's $T^2$ ($HT^2$) and squared prediction error (SPE). Predictions resultant of simulation scenarios that violate either model validity metric (≥100% of the limit) cannot be trusted without performing physical experiments to confirm the predictions and update the PLS model.

### *Cost*

Finally, it is important to compare the overall cost of each simulated formulation with the original. Of course, lower cost is more desirable.

**Table 1** compares both the model validity metrics and formulation cost.

After reviewing all the considerations, Resin #21 and Resin #50 were selected as the most promising alternatives to the original Resin #8.

Depending on the product development goal, the most promising candidate may be different. For example, if higher viscosity was prioritized over cost, then Resin #50 would be the clear candidate.

## Optimization

The developed model can be inverted using constrained mathematical optimization. The goal in this case was to use fewer than 10 ingredients in formulations that typically used significantly more ingredients. As can be seen in the score plot in **Figure 8a,** historical formulations using many ingredients clustered towards the bottom right. Five score targets for the optimization were set with constraints on:

- the number of ingredients per class;
- ensuring the ratios summed to 100%; and
- ensuring the resulting formulations remained within the model validity metrics.

The scores of the resulting optimization formulations are shown on the right in **Figure 8b**. These fell within the target regions and obeyed the constraints. The number of ingredients used per class is shown in **Table 2.**

**TABLE 1**
**Model Validity Metrics and Formulation Cost**

| Resin Used | Original (Resin #8) | Resin #21 | Resin #52 | Resin #50 | Resin #40 |
|---|---|---|---|---|---|
| HT2 Model Validity (% of limit) | 47.8 | 65.7 | 76.5 | 73.6 | 111.1 |
| SPE Model Validity (% of limit) | 42.9 | 56.6 | 88.6 | 52.8 | 77.4 |
| Overall cost of formulation | 0.96570 | 0.95878 | 1.0259 | 1.0259 | 1.0551 |

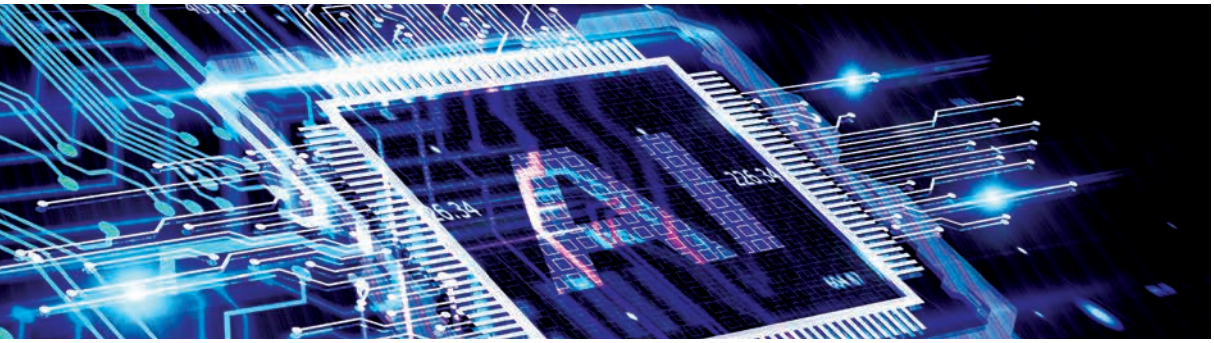
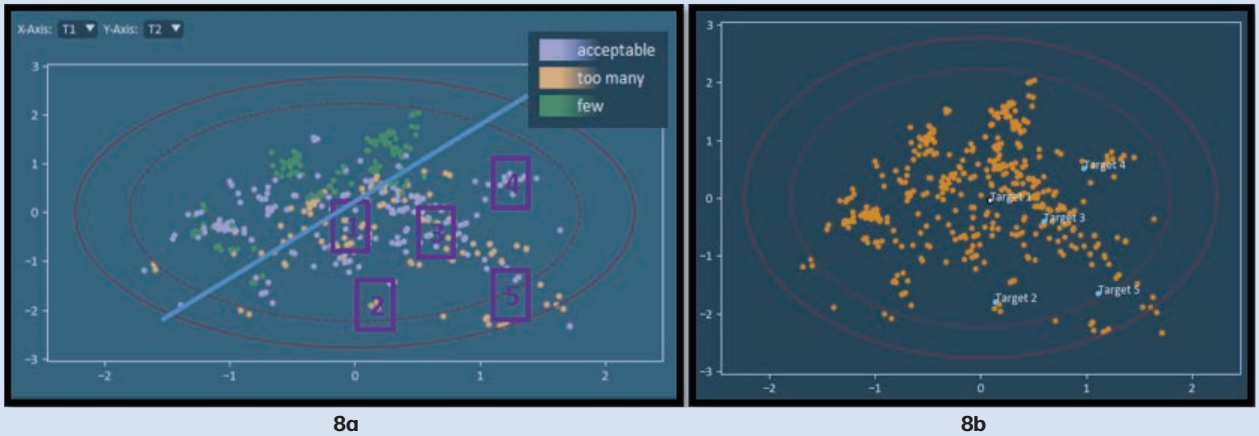


**FIGURE 8**
**8a. Historical formulations using many ingredients are clustered toward the bottom right.**
**8b. Scores of the resulting optimization formulations are shown.**




8a 8b

**TABLE 2**
**Ingredients Used Per Class**

| Ingredient Limits | Historical | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 |
|---|---|---|---|---|---|---|
| Additive | 1 | 1 | 1 | 1 | 1 | 1 |
| Catalyst | 1-4 | 1 | 1 | 2 | 3 | 3 |
| Isocyanates | 1-2 | 1 | 1 | 1 | 1 | 1 |
| Resin | 1-8 | 3 | 2 | 2 | 2 | 2 |
| Solvent | 3-8 | 4 | 4 | 4 | 3 | 3 |
| Total | 8-22 | 10 | 9 | 10 | 10 | 10 |

## Conclusion

Applying AI to product development datasets for accelerated innovation is quickly becoming a requirement for manufacturers that wish to remain competitive in their industry. The framework outlined in this article can be applied to a variety of datasets that range in size and industry, and it can help manufacturers identify weaknesses in their dataset while developing a predictive model. The predictive model can be used in a forward direction (e.g., simulation) or backward direction (e.g., constrained mathematical optimization) to reduce physical experimentation to achieve a variety of product development goals such as reformulating existing products or developing new products.

A coatings dataset was used to illustrate the usefulness of an AI model in reformulating a product first by replacing one key ingredient, and then by reducing the total number of ingredients. ❊

**Marlene Cardin** *is director of projects at ProSensus. Email: marlene.cardin@prosensus.ca.*