NASA/TM-20220008637

Design of Materials With Alchemite™

Joshua Stuckner
Glenn Research Center, Cleveland, Ohio

Thomas M. Whitehead, Robert C. Parini, and Gareth J. Conduit
Intellegens, Cambridge, United Kingdom

Othmane Benafan and Steven M. Arnold
Glenn Research Center, Cleveland, Ohio

July 2022

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NASA Technical Report Server—Registered (NTRS Reg) and NASA Technical Report Server—Public (NTRS) thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers, but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., "quick-release" reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Fax your question to the NASA STI Information Desk at 757-864-6500

- Telephone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Program
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM-20220008637

Design of Materials With Alchemite™

Joshua Stuckner
Glenn Research Center, Cleveland, Ohio

Thomas M. Whitehead, Robert C. Parini, and Gareth J. Conduit
Intellegens, Cambridge, United Kingdom

Othmane Benafan and Steven M. Arnold
Glenn Research Center, Cleveland, Ohio

National Aeronautics and
Space Administration

Glenn Research Center
Cleveland, Ohio 44135

July 2022

Acknowledgments

This work was supported by the NASA Transformational Tools and Technologies project under the Transformative Aeronautics Concept Program within the Aeronautics Research Mission Directorate.

This work was sponsored by the
Transformative Aeronautics Concepts Program.

Level of Review: This material has been technically reviewed by technical management.

This report is available in electronic form at http://www.sti.nasa.gov/ and http://ntrs.nasa.gov/

Design of Materials With Alchemite™

Joshua Stuckner
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135

Thomas M. Whitehead, Robert C. Parini, and Gareth J. Conduit
Intellegens
Cambridge, United Kingdom, CB4 3AZ

Othmane Benafan and Steven M. Arnold
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135

Abstract

Machine learning models that establish the relationships between materials processing and properties can enable inverse design of materials through active learning. Alchemite™ is a commercial software that can perform inverse materials design on sparse data. Here we evaluate Alchemite™'s performance on a dataset of shape memory alloys and a dataset of heat exchangers compared to baseline random forest models. Alchemite™ had higher accuracy when making predictions on sparse data and was more accurate or nearly as accurate as random forests on complete datasets while also quantifying uncertainty. The software was also used to suggest processing steps and design parameters to optimize properties and performance; however, physical validation of the suggested design parameters was beyond the scope of this work. Several useful design insights were gained about the impact of the design parameters on properties and performance including the importance of dopant choice and amount for shape memory alloys and the importance of height and weight on the thermal resistance of heat exchangers.

Introduction

Machine learning and informatics have greatly accelerated the developmental pace of new materials (Refs. 1 and 2). New materials are developed and optimized by fine-tuning processing parameters (e.g., material chemistry, manufacturing steps) to achieve desired properties. Until recently, this has been a forward process where processing parameters, often selected based on theory or computational tools, are tested to determine their properties. Inverse design of materials through active learning allows a practitioner to start with the desired properties and determine optimal processing parameters that will most likely achieve those properties (Refs. 3 to 5). However, applying the inverse design methodology is time consuming and requires expert knowledge from outside the typical materials science domain. Here we evaluate Alchemite™, a commercially available inverse design software by Intellegens, on two datasets. The primary purpose of this document is to report the results of the Alchemite™ software evaluation and not to perform a rigorous scientific study of the datasets used in the evaluation. Results show that Alchemite™ performed similarly well or better than random forest models on both datasets when performing continuous or categorical predictions. On sparse data, Alchemite™ performed significantly better than data imputation with k-nearest neighbors (Ref. 6) and regression using random forest models. Critically, Alchemite™ quantified the uncertainty of predictions, which is necessary for inverse design and allows users to know how much to rely on an individual prediction. Material experiments are expensive and often destructive to the sample, making sparsity a common feature of material's data.

Alchemite™ is a commercial deep learning software tool specifically designed to handle sparse and noisy experimental data. The method was originally developed in the alloy design space (Refs. 7 and 8) and has since seen further applications in the drug discovery (Refs. 9 to 12) and industrial chemistry (Refs. 13 and 14) domains, in each case guiding the design of new formulations and processing conditions to maximize product performance. Alchemite™ is a multi-target method and is capable of handling sparse training data on both the input and output side: the ability to handle sparse data is a key requirement for handling historical materials data that may not be characterized to current standards or may be missing expected information. The multi-target predictions from Alchemite™ are able to leverage nonlinear inter-target correlations, which also requires the ability to manage sparse property measurement data. Other machine learning methods, such as the random forests considered here, require pre-

processing to remove this sparsity, whereas Alchemite™ can handle the sparse data directly. Alchemite™ is accessed either via a server-based application programming interface (API), hosted on-site or in the cloud with an optional web interface, or (as in this work) via a Docker image running directly on the user's machine.

Inverse design of materials can be achieved with active learning (sometimes called sequential learning). Materials discovery through active learning is becoming increasingly popular due to its experimental efficiency (Refs. 1, 7, 8, 15 to 31). Active learning leverages information from previous experiments to recommend an optimal experiment to achieve target properties. It requires some initial data, a machine learning model that captures processing/property relationships with uncertainty, and an optimization algorithm to optimize some acquisition function. Active learning is an iterative process where a surrogate model is trained from the initial data to predict target properties with uncertainty. Then an optimizer is used to find the optimal inputs (processing steps) to the surrogate that maximizes the acquisition function. If the target properties are not achieved, the new experimental observation is added to the dataset and a new surrogate model is trained. This process repeats until the target is achieved or the experimental budget is exhausted. In the latter case, active learning should produce a superior material than would be discovered through trial and error. The initial data can be obtained from public or legacy data, a small design of experiments, or through the Latin hypercube sampling method. The acquisition function can be chosen for exploration (maximum uncertainty) or exploitation (maximum expected improvement), but optimizing for the processing parameters that give the highest likelihood of achieving the target considering uncertainty (maximum likelihood of improvement) has been shown to achieve the target with fewer experiments (Ref. 30). Several open-source packages are available for inverse design of materials (Refs. 32 to 35). By optimally sampling the vast materials design space through active learning, target material properties can be achieved with far fewer experiments, saving time and money.

Methods

Datasets

Two datasets were evaluated in this study. The first was extracted from datasheets of general-purpose heat exchangers sold by Cool Innovations. The datasheets were provided to NASA Glenn Research Center for a machine learning hackathon. The extracted and cleaned dataset contains 12 variables (three categorical and nine continuous) shown in Table 1. The four target variables for prediction are area, thermal resistance, configuration, and material. The configuration is the type of heat exchanger and can be moderate, splayed, sparse, flared, LED fan sinks, standard fan sinks, high power fan sinks, and integrated fan sinks. Figure 1 illustrates the size dimension variables of the dataset. Data sheets and illustrations of all the heat sink configurations can be found on the Cool Innovations website (www.coolinnovations.com).

The second dataset is a compilation of processing and property variables of shape memory alloys (SMA) extracted from numerous scientific sources and compiled as part of the NASA shape memory materials database and analysis tool (Ref. 36). The SMA dataset used in the work is a cleaned subset of a larger SMA dataset from the NiTi-based alloy system. The dataset contains a total of 15 variables including two categorical variables, 13 continuous variables, twelve input variables, and four target variables. Each SMA in the dataset is composed of three elements including Nickel and Titanium ($Ni_{100-a-b}Ti_aX_b$). Table 1 describes the dataset variables in detail.

TABLE 1.—HEAT EXCHANGER DATASET VARIABLES. BLUE COLORING INDICATES THAT THE VARIABLE IS A TARGET TO BE PREDICTED

Variable	Type	Description
Height	Continuous	Heigh of heat exchanger (mm), see Figure 1
Weight	Continuous	Weight of heat exchanger (g)
Base width	Continuous	Long base dimension (mm), see Figure 1
Base length	Continuous	Short base dimension (mm), see Figure 1
Splay width	Continuous	Width of fins (mm), see Figure 1
Splay length	Continuous	Length of fins (mm), see Figure 1
Airflow	Continuous	Airflow through fins during test (cubic feet per minute, CFM)
Area	Continuous	Area of base (mm^2)
Thermal resistance	Continuous	Resistance to heat exchanging (°C/W)
Base shape	Categorical	Square or Circular
Configuration	Categorical	One of eight including: moderate, splayed, integrated fan
Material	Categorical	Aluminum or Copper

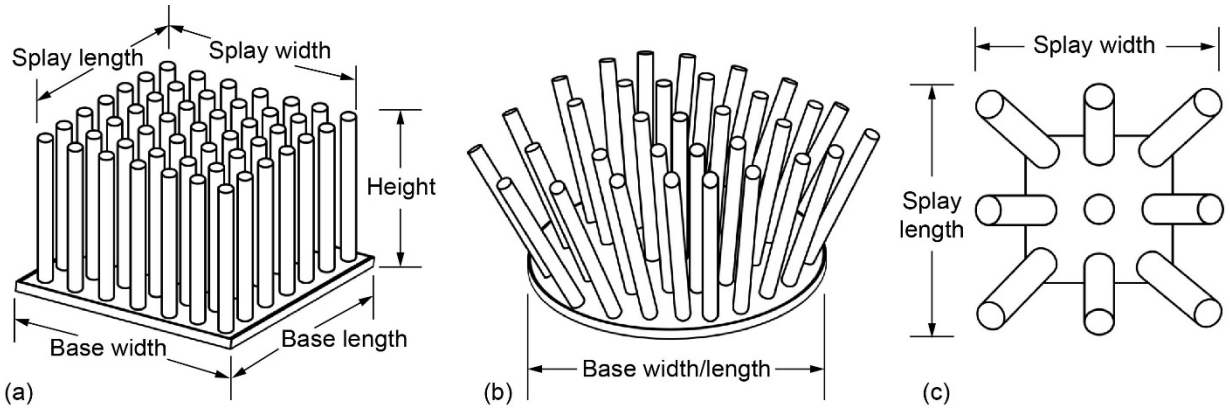

Figure 1.—Illustration of heat exchangers and size variables.

Alchemite™ Training

The Alchemite™ deep learning algorithm is described in detail elsewhere (Refs. 7 and 9); and only a brief description is provided here in the context of the particular heat exchanger and SMA dataset discussed above. As Alchemite™ is a multi-target method, one model was trained for all four target properties of the heat exchanger dataset simultaneously (two continuous variables, thermal resistance and area; and two categorical variables, configuration and material) based on all of the sparse input data available. As described elsewhere (Refs. 9 to 11, 37) Alchemite™ predictions for each of these outputs are based on an ensemble of sub-models, with each element of the ensemble trained on bootstrap samples of the training data, generating distributions of predicted values. Bootstrapping is a resampling method that randomly samples from the dataset with replacement, potentially allowing for individual observations to be repeated. For continuous variables, the Alchemite™ predicted value is the mean of the predictions across this distribution with the uncertainty in the prediction given by the standard deviation across these samples. For categorical variables the modal predicted class across the distribution is returned as the model prediction, with the complement of the probability of this class being selected taken as the uncertainty estimate in the result. The combination of the uncertainty estimates from the ensembling approach with the extrapolatory ability of deep learning methods combines to give an accurate understanding of the confidence in each prediction, including the reduction in confidence due to extrapolating beyond the support of the training set. This enables the probabilistic prioritization of predictions, enabling scientists to focus on only the most confident predictions and prioritize these for experimental validation.

To enable deeper insights into the model behavior, Alchemite™ also reports the relationships that it has identified between input and output variables. This is measured by evaluating the relative weight assigned to each input variable for predicting each output variable, on average across the whole model, using the information gain attributable to each input feature (Refs. 13 and 38).

All data for training and prediction is input to Alchemite™ from comma-separated values (CSV) files. Any categorical variables must be replaced with integer indexes. For example, the heat exchangers have eight categorical configurations (moderate, splayed, sparse, flared, LED fan sinks, standard fan sinks, high power fan sinks, and integrated fan sinks) which must be replaced with the values 0 to 7 in the CSV file. This can be achieved easily in Python using the pandas library factorize command. Factorization should be done before any splitting of the data.

Before model training, hyperparameter optimization was performed with Alchemite™ to determine the best hyperparameters for the model. Hyperparameters are adjustable configurations of the model architecture such as the number of layers in a neural network or the number of decision trees in a random forest. Hyperparameter optimization is performed by training numerous models and selecting the hyperparameters of the model with the least error or highest accuracy. Alchemite™'s accuracy metric is the median coefficient of determination of selected target variables which the trained model should predict well.

After the best hyperparameters are selected, the model is trained using the training data. Alchemite™ performs an internal train/validation data split, therefore a separate validation set is not required. A hold out test set was withheld from Alchemite™ for an unbiased accuracy estimation for comparison with the random forest models. Once a model is trained, the model can be used to make new predictions.

Training Random Forests

Random forests are tree based ensemble models capable of classification and regression (Ref. 39). They are composed of many decision trees that are trained on sub-samples of the

training data. By averaging the predictions of the individual decision trees, overfitting is reduced and the predictive accuracy is increased. Decision trees contain sequential nodes that split the data into increasingly pure class separation using the Gini index (a measure of class purity within a set) until the final node makes a classification or regression prediction (Ref. 40). Each split is made based on a single input variable and the split value and variable is chosen to maximize the class purity after the split. Data normalization was not performed because it does not affect the Gini index and does not benefit the training of random forest models. Random forests were trained in this work using the scikit-learn library in Python (Ref. 41).

Random forests require complete observations in the training data and to make predictions (i.e., there can be no missing values for any of the variables). To evaluate performance on sparse data as a baseline, the sparse datasets were pre-processed to impute the missing values. Data imputation was performed with a k-nearest neighbors approach using the scikit-learn library. The k-nearest neighbors model was fit to the training set and applied to impute missing values on both the training and test sets.

Statistical Metrics

The accuracy of models in predicting continuous variables is measured with the coefficient of determination (R^2), which is the proportion of variance in the target variable that is explained by the input variables (Ref. 42). When $R^2 = 1$, the model is perfectly accurate. A model that always predicts the mean of the target value, has an R^2 of 0. The equation for R^2 is given by:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{1}$$

where y is the true value of the target variable, $\bar{y}$ is the mean of the target value in the data, and $\hat{y}$ is the predicted value for each i^{th} sample.

Two accuracy metrics are reported for categorical variables: precision and F_1 score. Precision is simply the number of correct model predictions divided by the total number of predictions. The F_1 score accounts for precision and recall equally by the harmonic mean of precision and recall:

$$F_1 \text{ score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

where recall is the number of true positive predictions for each class divided by the number of samples in that class (Ref. 43). For multi-label categorical predictions (more than 2) the F_1 score is given by the average F_1 score of each class.

Two metrics are used to quantify the strength of the relationship between two variables. The Pearson's correlation coefficient, r, is a measure of the linear correlation between two variables and is given by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{3}$$

where x and y are two variables, $\bar{x}$ and $\bar{y}$ are the mean of the two variables, and i gives the sample index. Alchemite™'s importance index is a proprietary measure of the importance of each input variable in predicting each target variable.

Results

Shape Memory Alloy Data

Alchemite™ and random forest models were trained to predict the target variables of the SMA dataset. The four target variables are indicated in blue in Table 2.

Figure 2 shows the prediction accuracy of each model on the four targets. In each case the random forest model had a slightly higher R^2 than the Alchemite™ models, but the results were comparable. Alchemite™'s slightly reduced accuracy is likely due to reporting uncertainty in the predictions, which requires training ensembles of sub-models on bootstrap samples of the data. Because each sub-model is trained on fewer unique observations, the overall accuracy is reduced compared to if uncertainty was not calculated. The value of uncertainty prediction is demonstrated in Figure 2(a) where poor predictions given by each model are indicated by blue and orange arrows for Alchemite™ the random forest respectively. The blue error bars around the Alchemite™ predictions show the model's uncertainty about the prediction. Predictions far from the actual value (large error) typically have higher uncertainty. This is extremely valuable from a design perspective. It is impossible to know how confident to be in any given prediction from the random forest or whether the model is extrapolating far outside the training domain. By reporting uncertainty, Alchemite™ allows a user to make an informed decision about the amount of risk associated with relying on a particular prediction.

In order to evaluate the performance of Alchemite™ in imputing missing values, 20 percent of the data values in each attribute were randomly removed. To predict the target variables with the random forest models, the data was pre-processed to impute the missing values using a k-nearest neighbors approach. Alchemite™ imputed the sparse data

TABLE 2.—DESCRIPTION OF SHAPE MEMORY ALLOY DATASET VARIABLES. BLUE COLORING INDICATES THAT THE VARIABLE IS A TARGET TO BE PREDICTED

Variable	Type	Description
Ni %	Continuous	Percent nickel in alloy (implicitly assumed during analysis) (at.%)
Ti %	Continuous	Percent titanium in alloy (at.%)
Element 3 percent	Continuous	Percent element 3 in alloy (at.%)
Heat treatment 1 time	Continuous	Time and temperature of alloy heat treatment steps (°C, h)
Heat treatment 1 temperature	Continuous	
Heat treatment 2 time	Continuous	
Heat treatment 2 temperature	Continuous	
Heat treatment 3 time	Continuous	
Heat treatment 3 temperature	Continuous	
Austenite start temperature	Continuous	Temperature at which the martensite to austenite transformation begins on heating (°C)
Austenite finish temperature	Continuous	Temperature at which the martensite to austenite transformation is completed on heating (°C)
Martensite start temperature	Continuous	Temperature at which the transformation from austenite to martensite begins on cooling (°C)
Martensite finish temperature	Continuous	Temperature at which the transformation from austenite to martensite is completed on cooling (°C)
Element 3	Categorical	Third element type added to the alloy
Processing method	Categorical	Alloy processing (e.g., vacuum arc remelting, sputtering)

directly. Figure 3 shows the prediction accuracy of each model. As expected, both models saw a reduction in accuracy, but the random forest models showed a much larger decrease. For example, when predicting the martensite start temperature Alchemite™ exhibited a 0.15 reduction in R^2 (0.89 to 0.74) while the random forests recorded a 0.35 reduction in R^2 (0.92 to 0.57), which is a 233 percent larger reduction in accuracy. Other target predictions had similar results. Overall, Alchemite™ performed significantly better than random forests with k-nearest neighbor imputation when making predictions with sparse data.

Figure 4 shows the relationships between the variables using two forms of exploratory data analysis. Figure 4(a) shows the Pearson correlation coefficient between all the continuous variables, which is a measure of normalized linear correlation with negative numbers indicating inverse correlation. Categorical variables are not quantified by Pearson correlation. The fraction of titanium or nickel in the alloy is inversely correlated with the fraction of the third element. Depending on the atomic site preference of the third element, some will replace nickel and others will replace titanium. Although increasing the dopant fraction does not always reduce titanium, Pearson's correlation does not consider interaction or non-linear effects between variables and a mild inverse correlation exists across the dataset. Similarly, there is a positive Pearson's correlation between doping with a third element and the austenite and martensite start and finish temperatures. Although this is not always the case and these temperatures may decrease depending on the element and amount of the dopant. Figure 4(b) shows the importance of each input variable (horizontal axis) at predicting each target variable (vertical axis). The chemistry of each sample had the largest impact on the target properties. The Element 3 percent had a greater impact than the Ti %. The third element type had the most impact on predicting the target variables. The first heat treatment time and temperature had a moderate impact and subsequent heat treatments had a reduced impact. This result is likely an artifact of the dataset and preprocessing. Often the first heat treatment is homogenization while the second and third heat treatments affect the properties by changing the microstructure through annealing and aging. Many of the observations in the training data had only one heat treatment and very few had three, thereby reducing the apparent importance of the third heat treatment across the dataset. Feature engineering, perhaps by adding a categorical variable indicating the type of heat treatment performed, could potentially increase the signal for the importance subsequent heat treatments. Pearson's correlations can only account for linear relationships between two variables and cannot quantify non-linear or interaction effects which is very common in materials systems. Higher order and nonlinear importance relationships can be calculated with random forests using Gini importance, permutation importance (Ref. 39), or Shapley Additive exPlanations (SHAP) (Ref. 44) which are not evaluated here. Alchemite™ provides a global view of the importance of each variable in predicting the others and accounts for nonlinear and interaction effects.

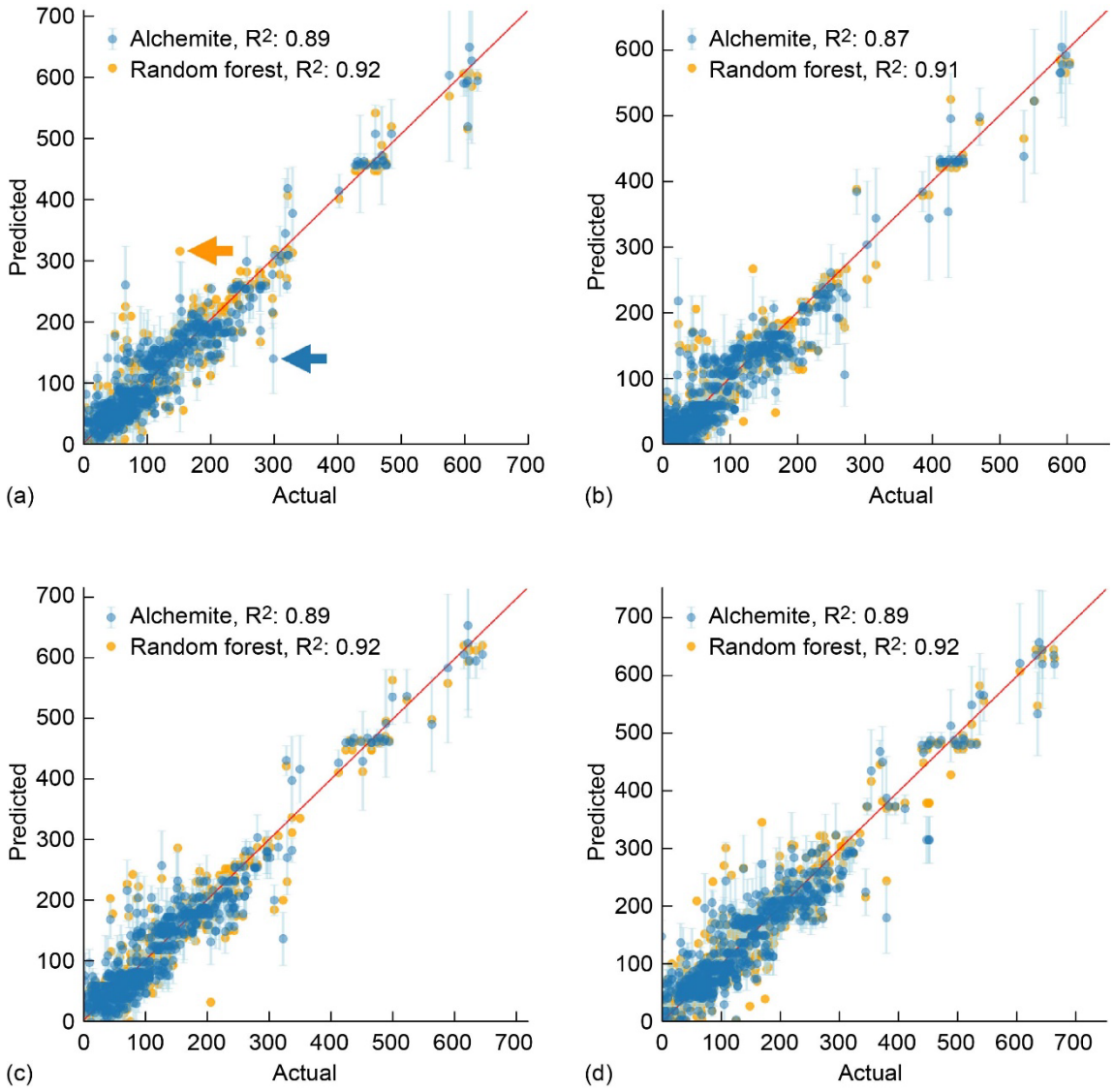

Figure 2.—Prediction accuracy on SMA dataset of Alchemite™ (blue) is compared to random forest models (orange). (a) to (d) show the predicted values versus the actual values of the four target variables in the SMA dataset. Alchemite™ predictions display error bars which represent the models' prediction uncertainty. The blue arrow in (a) indicates an Alchemite™ prediction with uncertainty while the yellow arrow indicates a random forest prediction without uncertainty.

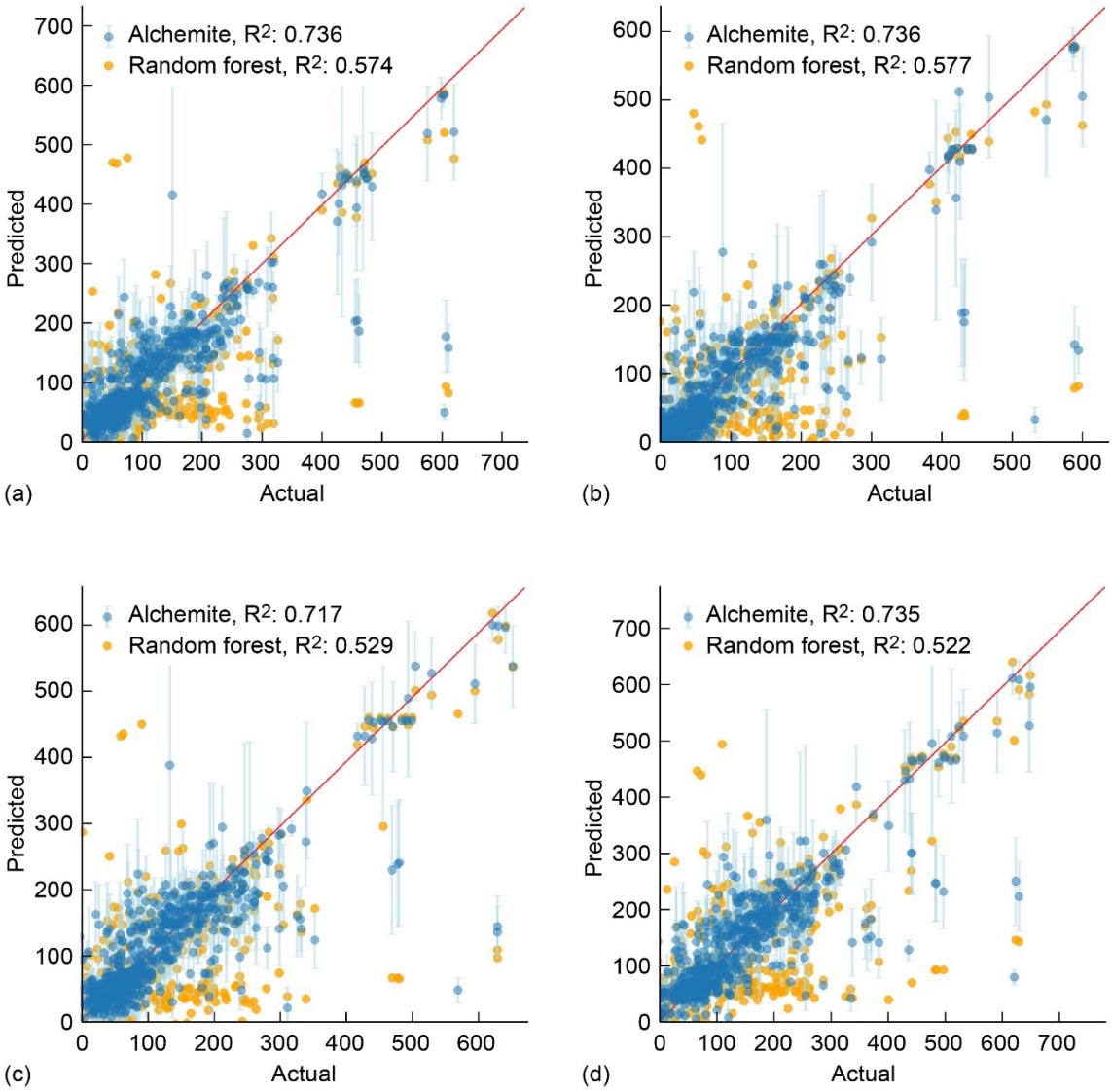

Figure 3.—Prediction accuracy on SMA dataset with 20 percent missing values of Alchemite™ (blue) compared to random forest models (orange). (a) to (d) show the predicted values versus the actual values of the four target variables in the SMA dataset. Alchemite™ predictions display error bars which represent the model's prediction uncertainty.

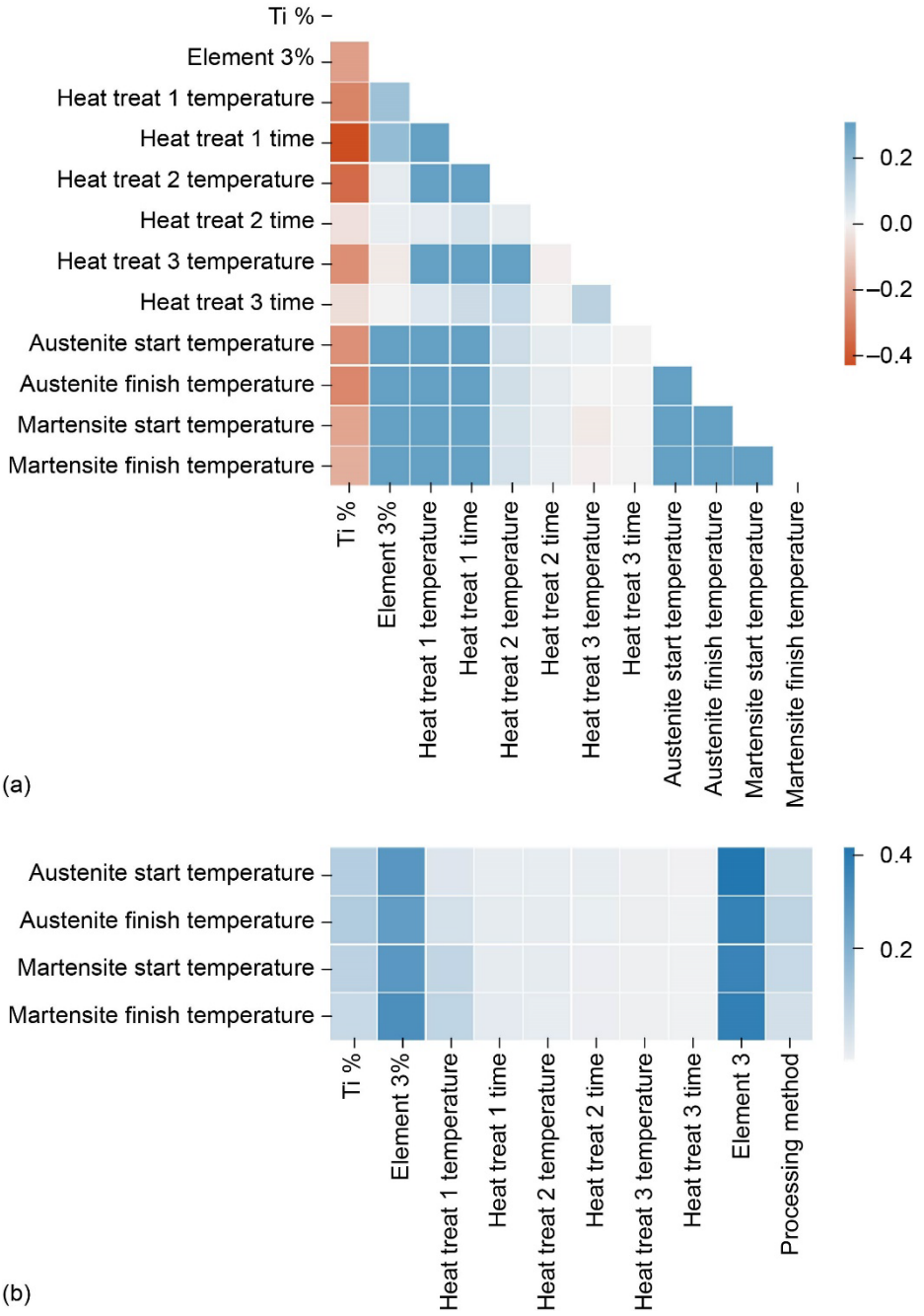

Figure 4.—Plots of relationships between variables in the SMA dataset. (a) shows the Pearson's correlation between the continuous variables. (b) shows Alchemite™'s importance metric of each input variable (horizontal axis) in predicting each target variable (vertical axis).

TABLE 3.—SUGGESTED EXPERIMENT TO PRODUCE AN SMA WITH MAXIMUM AUSTENITE FINISH TEMPERATURE. THE SUGGESTED EXPERIMENTAL SETTINGS AND PREDICTED OUTCOME WITH UNCERTAINTY ARE SEPARATED BY THE BOLD HORIZONTAL LINE IN THE TABLE

Parameter	Value
Ni %	Balance (8.55 percent)
Ti %	49.79 percent
Element 3%	41.66 percent
Heat treatment 1 temperature	822.5
Heat treatment 1 time	75.66
Heat treatment 2 temperature	162.3
Heat treatment 2 time	1685
Heat treatment 3 temperature	322.4
Heat treatment 3 time	3482
Element 3	Au
Processing method	Vacuum arc melting
Austenite start temperature	465.6±54.25
Austenite finish temperature	487.4±51.13
Martensite start temperature	452.5±62.58
Martensite finish temperature	413.7±42.36

Alchemite™ was used to suggest an optimal experiment to iteratively approach the goal of producing a SMA with a maximum austenite finish temperature, targeting a minimum of 250 °C. To ensure a composition summing to 100 percent, the "sum below" target was used in the Alchemite™ configuration to ensure that Ti %, and Element 3 percent totaled less than 100 percent. Ni % was considered the balance. Notably, Alchemite™ chose Au for Element 3. The chosen alloy was $Ni_{8.6}Ti_{49.8}Au_{41.7}$ and had a predicted austenite finish temperature of 487.4±51.13 °C with a probability of hitting the target property of 56.2 percent. The predicted heat treatment 2 and 3 temperatures are too low to have any real effect at any time scale. Since most of the data did not specify heat treatment 2 and 3, it is not surprising that the model didn't make meaningful predictions for those parameters. This chemistry prediction is a reasonable adjustment of data seen by the model to achieve the target austenite finish temperature (Ref. 45). The processing parameters and predicted outcome of the suggested experiment are shown in Table 3.

Heat Exchanger Data

The heat exchanger data was used to evaluate Alchemite™'s predictive performance on both categorical and continuous targets. Alchemite™ and random forests were used to predict two categorical variables (material and configuration) and two continuous variables (thermal resistance and area). The results of the predictions are shown in Figure 5. The random forest models were slightly better, but comparable, at predicting the configurations with a precision of 0.99 and F_1 score of 0.87 compared to Alchemite™ with a precision of 0.98 and F_1 score of 0.83. As with continuous variables, Alchemite™ quantified uncertainty for each prediction which typically comes at a slight cost in accuracy. When predicting the material, both Alchemite™ and random forests were 100 percent accurate. Alchemite™ performed significantly better than random forests when predicting the continuous target, thermal resistance, even while accounting for uncertainty. Alchemite™ predictions had an R^2 of 0.79 compared to 0.70 for the random forest model. Both models were 100 percent accurate when predicting area. This was expected because area is easily calculated using simple geometry from the base width, base length, and base shape.

The Pearson's and Alchemite™ importance relationships between the variables are shown in Figure 6. From the Pearson's correlations in Figure 6(a), thermal resistance is inversely correlated to height, weight, widths, and lengths. This is expected because larger heat exchangers can transfer more heat (i.e., less thermal resistance). The area was also heavily correlated to the size features of the heat exchanger including height, weight, and splay size. This is because heat exchangers in the dataset often had increased height and splay dimensions when increasing the base dimensions and is a prime example that correlation is not causation. The Alchemite™ importance metric is shown in Figure 6(b). As expected, the base width and base length were highly important to predicting the area. Interestingly, the base shape was not used to predict the area, which would have been expected to choose between calculating the area of a circle or a rectangle. Instead, the model appears to have made use of the weight information, perhaps to infer the shape. Weight was the primary feature used to predict the material along with height and width (presumably to estimate density). Height, weight, splay width, and base width were the primary features used to predict thermal resistance. Lengths may not have been as important because the heights and widths were highly correlated (i.e., the dataset contained heat exchangers with similar aspect ratios) so the two variables contained a large proportion of redundant information.

To test optimization with constraints, Alchemite™ was asked to produce a heat sink with a thermal resistance below 1 °C/W with the following constraints: a maximum airflow of 800 CFM, max base size of 75×75 mm, and splay size of 100×100 mm. These are realistic constraints when designing a component that may have limited space for the sink or limited power or space for a stronger fan. When limiting the airflow, the optimizer converged on a heatsink with an integrated fan! The full results of the optimization are shown in Table 4. If

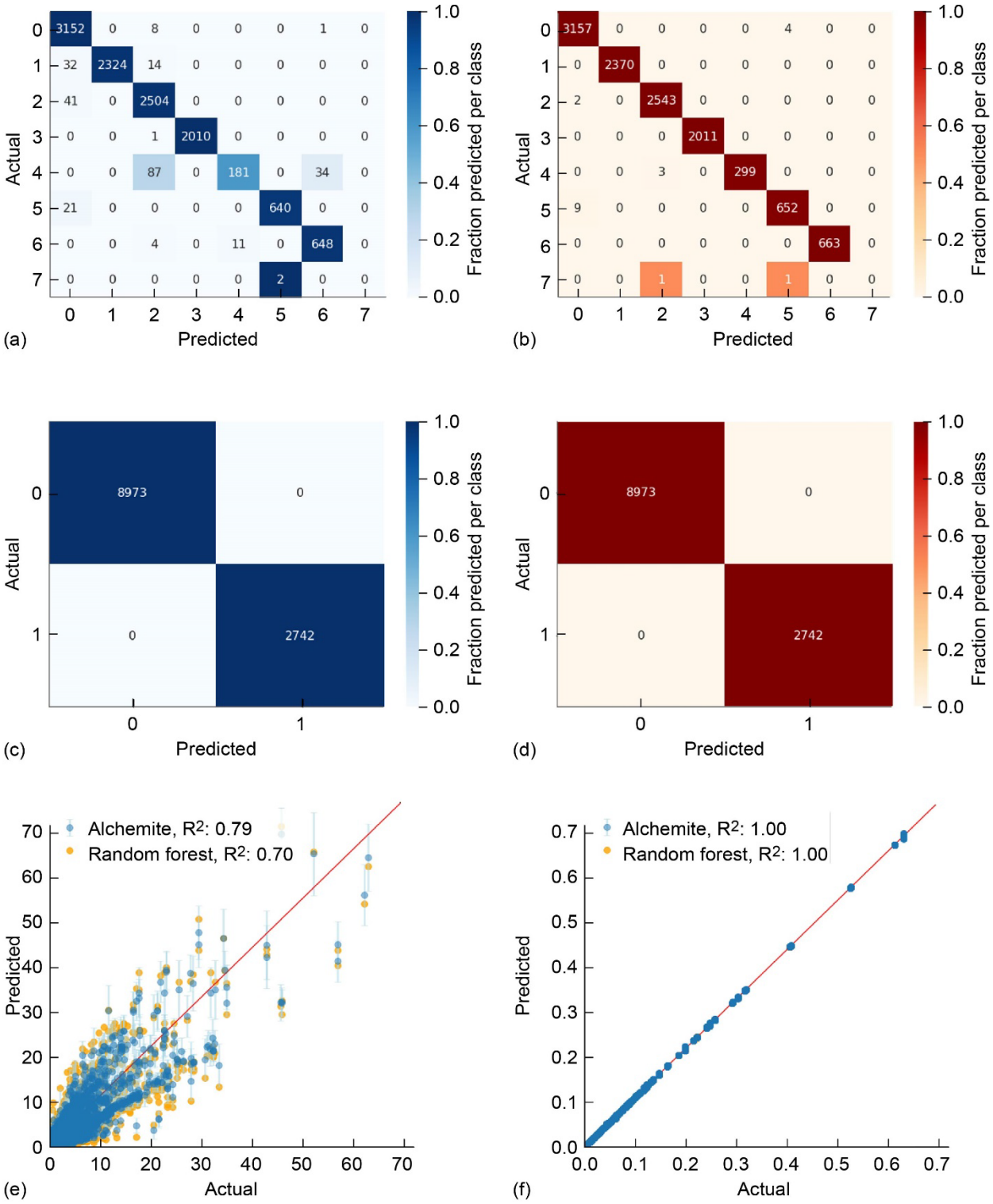

Figure 5.—Prediction accuracy of Alchemite™ and random forest models on the heat exchanger dataset. (a to d) show categorical predictions on the configuration (a, b) and material (c, d) by Alchemite™ shown in blue (a, c) and the random forest shown in orange (b, d). (e and f) show the models' predictions on the thermal resistance l and area (f).

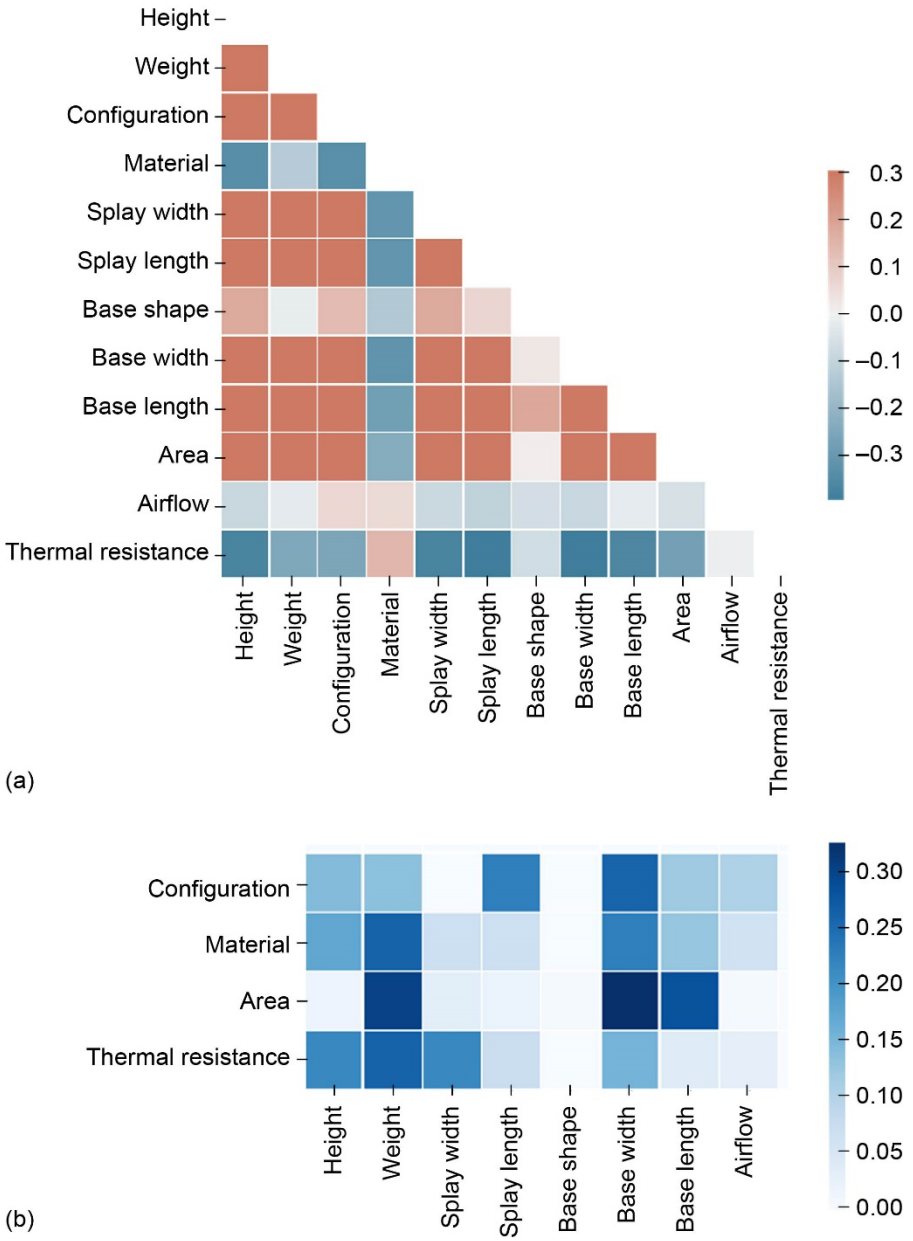

Figure 6.—Plots of relationships between variables in the heat exchanger dataset. (a) shows the Pearson's correlation between the continuous variables. Correlations involving categorical variables are not true Pearson correlations. (b) shows Alchemite's™ importance metric of each input variable (horizontal axis) in predicting each target variable (vertical axis).

TABLE 4.—RESULT OF OPTIMIZATION TO PRODUCE A HEAT EXCHANGER WITH A THERMAL RESISTANCE BELOW 1 °C/W (TARGET INDICATED IN ORANGE). SEVERAL CONSTRAINTS TO THE OPTIMIZATION ARE SHOWN IN BLUE. THE OPTIMAL PARAMETERS ARE SHOWN IN BLACK TEXT

Parameter	Value
Height	63.04 mm
Weight	3729.0 g
Base width	37.80 mm (max: 75 mm)
Base length	55.96 mm (max: 75 mm)
Splay width	49.18 mm (max: 100 mm)
Splay length	58.47 mm (max: 100 mm)
Airflow	513.2 CFM (max: 800 CFM)
Base shape	Rectangle
Material	Aluminum
Configuration	Integrated Fan Sink
Thermal resistance	**0.261±0.037 °C/W (Target: < 1 °C/W)**

the design could not afford power or an additional power cable for the heat sink fan, the optimization could be constrained to not use an integrated fan sink. In the training dataset the base and splay width was always larger than the lengths by design. However, this optimal design had the lengths larger than the width which may therefore suggest that the model is extrapolating outside of the training data. This could be fixed by adding a constraint that forces the lengths to be less than the widths. Admittedly, the optimization produces a couple of dubious results which may stem from this extrapolation. For example, 8.1 lb (3.7 kg) is quite heavy for a 2.2×1.5 in. (37.8×55.7 mm) aluminum heat exchanger. It is also unexpected that the optimizer didn't use to the maximum allowed constraints, to reduce the thermal resistance. This accentuates the need for sound statistical and data science principals when setting up the Alchemite™ model and optimization to avoid strange results which may stem from allowing the model to extrapolate, especially in nonphysical ways.

Conclusion

Alchemite™ was successful in predicting desired targets of a materials and a non-materials dataset. Alchemite™ was able to handle inputs and predictions of both continuous and categorical variables. As a result of initial analysis during this work, improvements were made to Alchemite to increase categorical prediction accuracy. Compared to random forest baseline models, Alchemite™ was far superior when operating on sparse data and better than or nearly equivalent to random forests on the full datasets while also quantifying uncertainty. Alchemite™ was also used to suggest optimal experiments for inverse design, perform design optimization, and rank important features. Physical validation of suggested parameters was not performed. Alchemite™'s uncertainty quantification is useful to know when the model may be extrapolating and how much to rely on a given prediction and is critical for performing active learning.

Several insights about the datasets were gained from this analysis. The thermal resistance of the heat exchangers was highly correlated with height; however, the Alchemite™ importance index showed that weight was a stronger predictor. Surprisingly, airflow was not a strong predictor of thermal resistance. In order to maximize thermal resistance when airflow was limited in the design, a heat exchanger with an integrated fan was suggested as being optimal. The choice and amount of dopant in NiTi SMAs had the largest impact on the austenite and martensite start and finish temperatures, which was not apparent from Pearson's correlation alone. To maximize the austenite finish temperature, Alchemite™ suggested replacing a large amount of Ni with Au.

References

1. T. Lookman, P.V Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," npj Comput. Mater., vol. 5, no. 1, pp. 1–17, 2019.
2. J. Schmidt, M.R.G. Marques, S. Botti, and M.A.L. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater., vol. 5, no. 1, pp. 1–36, 2019.
3. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature, vol. 559, no. 7715, pp. 547–555, 2018.
4. Y. Tian et al., "Role of uncertainty estimation in accelerating materials development via active learning," J. Appl. Phys., vol. 128, no. 1, p. 14103, 2020.
5. A.G. Kusne et al., "On-the-fly closed-loop materials discovery via Bayesian active learning," Nat. Commun., vol. 11, no. 1, pp. 1–11, 2020.
6. O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, no. 6, pp. 520–525, 2001.
7. B.D. Conduit, N.G. Jones, H.J. Stone, and G.J. Conduit, "Design of a nickel-base superalloy using a neural network," Mater. Des., vol. 131, pp. 358–365, 2017, doi: 10.1016/j.matdes.2017.06.007.
8. B.D. Conduit, N.G. Jones, H.J. Stone, and G.J. Conduit, "Probabilistic design of a molybdenum-base alloy using a neural network," pp. 8–11.

9. T.M. Whitehead, B.W.J. Irwin, P. Hunt, M.D. Segall, and G.J. Conduit, “Imputation of Assay Bioactivity Data Using Deep Learning,” J. Chem. Inf. Model., vol. 59, no. 3, pp. 1197–1204, 2019, doi: 10.1021/acs.jcim.8b00768.
10. B.W.J. Irwin, T.M. Whitehead, S. Rowland, S.Y. Mahmoud, G.J. Conduit, and M.D. Segall, “Deep imputation on large-scale drug discovery data,” Appl. AI Lett., vol. 2, no. 3, p. e31, 2021.
11. B.W.J. Irwin, J. Levell, T.M. Whitehead, M.D. Segall, and G.J. Conduit, “Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data,” pp. 1–17, 2020, doi: 10.1021/acs.jcim.0c00443.
12. E.G. Tse et al., “An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials,” J. Med. Chem., vol. 64, no. 22, pp. 16450–16463, 2021, doi: 10.1021/acs.jmedchem.1c00313.
13. T.M. Whitehead, F. Chen, C. Daly, and G.J. Conduit, “Accelerating the Design of Automotive Catalyst Products Using Machine Learning,” Johnson Matthey Technol. Rev., vol. 44, no. 0, pp. 0–17, 2021, doi: 10.1595/205651322x16270488736796.
14. S.X. Drakopoulos et al., “Formulation and manufacturing optimization of lithium-ion graphite-based electrodes via machine learning,” Cell Reports Phys. Sci., vol. 2, no. 12, p. 100683, 2021.
15. S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda, and J. Shiomi, “Designing nanostructures for phonon transport via Bayesian optimization,” Phys. Rev. X, vol. 7, no. 2, p. 21024, 2017.
16. T. Yamashita, N. Sato, H. Kino, T. Miyake, K. Tsuda, and T. Oguchi, “Crystal structure prediction accelerated by Bayesian optimization,” Phys. Rev. Mater., vol. 2, no. 1, p. 13803, 2018.
17. S. Kiyohara, H. Oda, K. Tsuda, and T. Mizoguchi, “Acceleration of stable interface structure searching using a kriging approach,” Jpn. J. Appl. Phys., vol. 55, no. 4, p. 45502, 2016.
18. M. Yamawaki, M. Ohnishi, S. Ju, and J. Shiomi, “Multifunctional structural design of graphene thermoelectrics by Bayesian optimization,” Sci. Adv., vol. 4, no. 6, p. eaar4192, 2018.
19. L. Bassman et al., “Active learning for accelerated design of layered materials,” npj Comput. Mater., vol. 4, no. 1, pp. 1–9, 2018.
20. P.V Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, “Adaptive strategies for materials design using uncertainties,” Sci. Rep., vol. 6, no. 1, pp. 1–9, 2016.
21. A. Solomou et al., “Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling,” Mater. Des., vol. 160, pp. 810–827, 2018.
22. A. Talapatra, S. Boluki, T. Duong, X. Qian, E. Dougherty, and R. Arróyave, “Autonomous efficient experiment design for materials discovery with Bayesian model averaging,” Phys. Rev. Mater., vol. 2, no. 11, p. 113803, 2018.
23. A.M. Gopakumar, P.V Balachandran, D. Xue, J.E. Gubernatis, and T. Lookman, “Multi-objective optimization for materials discovery via adaptive design,” Sci. Rep., vol. 8, no. 1, pp. 1–12, 2018.
24. A. Dave et al., “Autonomous discovery of battery electrolytes with robotic experimentation and machine learning,” Cell Reports Phys. Sci., vol. 1, no. 12, p. 100264, 2020.
25. R. Dehghannasiri et al., “Optimal experimental design for materials discovery,” Comput. Mater. Sci., vol. 129, pp. 311–322, 2017, doi: 10.1016/j.commatsci.2016.11.041.
26. D. Xue, P.V Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, “Accelerated search for materials with targeted properties by adaptive design,” Nat. Commun., vol. 7, no. 1, pp. 1–9, 2016.
27. D. Xue et al., “An informatics approach to transformation temperatures of NiTi-based shape memory alloys,” Acta Mater., vol. 125, pp. 532–541, 2017.
28. W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, “Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals,” 2021, [Online]. Available: http://arxiv.org/abs/2102.06191.
29. B. Rouet-Leduc, C. Hulbert, K. Barros, T. Lookman, and C.J. Humphreys, “Automatized convergence of optoelectronic simulations using active machine learning,” Appl. Phys. Lett., vol. 111, no. 4, p. 43506, 2017.
30. J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, “High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates,” Integr. Mater. Manuf. Innov., vol. 6, no. 3, pp. 207–217, 2017.
31. P.V Balachandran, B. Kowalski, A. Sehirlioglu, and T. Lookman, “Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning,” Nat. Commun., vol. 9, no. 1, pp. 1–9, 2018.
32. T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, and K. Tsuda, “COMBO: An efficient Bayesian optimization library for materials science,” Mater. Discov., vol. 4, pp. 18–21, 2016.
33. M. Hutchinson, S. Paradiso, and L. Ward, “Citrine informatics lolo.” 2016, [Online]. Available: https://github.com/CitrineInformatics/lolo.
34. T.M. Dieb, S. Ju, K. Yoshizoe, Z. Hou, J. Shiomi, and K. Tsuda, “MDTS: automatic complex materials design using

Monte Carlo tree search," Sci. Technol. Adv. Mater., vol. 18, no. 1, pp. 498–503, 2017.
35. K. Kandasamy et al., "Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly," J. Mach. Learn. Res., vol. 21, no. 81, pp. 1–27, 2020.
36. O. Benafan, G.S. Bigelow, and A.W. Young, "Shape memory materials database tool—a compendium of functional data for shape memory materials," Adv. Eng. Mater., vol. 22, no. 7, p. 1901370, 2020.
37. B. Efron, "Bootstrap methods: another look at the jackknife," in Breakthroughs in statistics, Springer, 1992, pp. 569–593.
38. B. Frénay, G. Doquire, and M. Verleysen, "Is mutual information adequate for feature selection in regression?," Neural Networks, vol. 48, pp. 1–7, 2013.
39. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
40. C. Gini, "Measurement of inequality of incomes," Econ. J., vol. 31, no. 121, pp. 124–126, 1921.
41. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
42. C.R. Rao, C.R. Rao, M. Statistiker, C.R. Rao, and C.R. Rao, Linear statistical inference and its applications, vol. 2. Wiley New York, 1973.
43. L.R. Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.
44. S.M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, 2017.
45. L. Casalena et al., "Mechanical behavior and microstructural analysis of NiTi-40Au shape memory alloys exhibiting work output above 400 C," Intermetallics, vol. 86, pp. 33–44, 2017.