A data-driven scheme to search for alternative composite materials

Michihiro Okuyama,†, Yukihito Nakazawa† and Kimito Funatsu†

†Konica Minolta, Inc, Hachioji, Japan; ‡Nara Institute of Science and Technology, Ikoma, Japan

ABSTRACT

Here, we propose an interaction analysis method, which quantitatively evaluates the interactions among constituent materials in composites that consist of matrix (e.g., resin), fillers, and additives (e.g., fiber-reinforced plastics). Using this method, we developed a data-driven scheme that is capable of identifying alternative constituent materials. The scheme enabled the search for new composites with the same performance as the original composite. The proposed scheme was also found useful for composites with complex interactions, where the features of constituent materials (e.g., the addition ratio) exhibited linear and exponential relationships with the performance. To demonstrate the utility of the proposed scheme, we searched for alternative constituent materials, which are involved in the interaction with the largest contribution to the performance (i.e., bending modulus). The bending moduli of the resulting composites were comparable to the original composite, thereby verifying the utility of the proposed scheme.

1. Introduction

Composite materials, such as fiber-reinforced plastics, are very significant in the development of products in the fields of aerospace, automobiles, and medical equipment, where they have been applied to products as varied as automobile trunks and aircraft gas turbines [1]. Composite materials are compounds of two or more constituent materials, and their performance can far exceed that of a single material. In addition, the importance of environmental sustainability has been increasingly recognized in recent years. This is highlighted by the Sustainable Development Goals adopted by the United Nations General Assembly; thus, environmental sustainability has become a very important issue for modern industry [2]. One goal of the field of composite materials is thus to develop composites that use recyclable materials or biomass and to utilize these materials in fabricating products [3–6]. This effort is expected to reduce both industrial waste and the amount of plastic used. One way to produce such composites for use in products is to substitute the constituent materials that make up the composites with biomass materials or recyclable materials. However, such substitutions may result in performance that is significantly reduced compared to that provided by the composite before the substitution. This occurs because the performance of a composite is affected not only by the features of its individual constituent materials (e.g., addition ratios and physicochemical quantities such as molecular weight and dielectric constant) but also by the interactions among the constituents. Therefore, in order to obtain a new composite material with the same performance as the one before the substitution, it is essential to take both of these issues into account. However, it is difficult to obtain a new composite material that achieves the same performance as the one before the substitution based on human experience and intuition because it takes an enormous amount of time to evaluate countless materials while also taking into...
account the interactions between them. Machine learning has therefore been studied as an approach to solving the difficulty of material exploration. Thanks to this research, a number of methods have been proposed for conducting rapid searches for materials that meet the target performance, and these methods have enabled the discovery of many new materials [7–20]. The methods employed in these previous studies consider both performance and material features, and they represent these relationships as a mathematical formula – or a predictive model – and use machine learning to evaluate the performance of the material. In this way, it is possible to conduct rapid searches among an enormous number of materials for those that fulfill the target performance. However, in developing a composite material, many of the features of the constituent materials that make up the composite are often unknown because the constituents are purchased from manufacturers who generally do not provide this information. This makes it difficult to search for materials using predictive models. These difficulties constitute major barriers to substiutions with biomass or recyclable materials. To overcome this difficulty, it is important (i) to determine the target materials for substitution by evaluating their interactions quantitatively and (ii) to search for new materials to substitute the target material in a way that does not rely on a predictive model.

For such a search, to address point (i), quantifyng and evaluating the interactions among the constituents of the composite is necessary. For the performance of composite materials, this depends on the addition ratio of the constituent materials and their combination. Furthermore, the dependence of performance on the addition ratio is often not linear due to complex interactions among the constituent materials [21,22]. Therefore, the interactions among the components can be represented by a nonlinear term of the form \( x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m} \), where \( n_i \) is an integer greater than or equal to 0 and \( x_i \) is a relevant feature of material \( i \) and \( i = 1, 2, \ldots, m \).

One method for considering such an interaction in a predictive model is to use the Kernel function [23]. However, it is difficult to evaluate the interaction quantitatively in such a predictive model because the Kernel function often is not represented directly by an interaction in such a model. An alternative is to use a predictive model constructed using the quadratic partial least squares (PLS) method [24]. In this method, the latent variable that maximizes the covariance between a performance indicator and the features is constructed using the training data, and the interaction is evaluated by a predictive model in which the relationship between this latent variable and the performance indicator is expressed as a quadratic function. However, since the interaction represented by this predictive model is limited to the form \( x_i x_j \), it is difficult to evaluate interactions of higher-order quantitatively. In order to evaluate higher-order interactions, Suzumura et al. therefore considered an interaction of the form \( x_1 x_2 x_3 \cdots x_m \), which contains the product of two or more features, and they proposed a method for extracting the features and interactions that are correlated with the objective variable [25]. However, as mentioned above, despite the fact that various methods have been proposed for dealing with interactions, there exists no general method for evaluating interactions quantitatively or for considering higher-order interactions such as \( x_1^{n_1} x_2^{n_2} \cdots x_m^{n_m} \).

As mentioned above in point (ii), many methods have been proposed for materials searches, but in order to apply these methods, the following two conditions must be fulfilled: First, feature and performance data must be available for all the training data used in the construction of the predictive model. Second, the feature must be a quantity that can search for unknown materials, which are not included in the training data as input to the predictive model. However, in the case of a composite material, it is extremely rare to be able to use features that allow a search for new materials. Therefore, we are forced to use features that are difficult to search for, such as the addition ratio of the material used in the training data [10]. Consequently, search methods that do not use a predictive model have become important. One such method is a search using unsupervised learning. For example, Zhang et al. applied hierarchical clustering – a method of unsupervised learning – to search for Li-containing materials with high Li-ion conductivities, and they succeeded in predicting as many as 16 Li-containing materials with high Li-ion conductivities among about 3000 Li-containing materials [26]. This search method is unique in that it uses only the feature data for candidate materials and for known high-performance materials through hierarchical clustering. This method is therefore very useful in situations where it is difficult to search using predictive models. However, to the best of our knowledge there is no published example of such an unsupervised-learning-based search method being applied to materials with interactions, such as composite materials.

In this study, we considered a composite with components, including a matrix (e.g., resin), filler, and additive, and proposed an interaction analysis method that can evaluate the interactions among the component materials quantitatively. It is based on an analogy with the cluster-expansion method of Mayer et al. in statistical physics [27], and it employs a new data-driven search scheme that combines this method with an unsupervised-learning search method. Hereafter, for convenience of explanation, we refer to matrices, fillers, and additives as ‘elements’ and their corresponding materials as ‘elemental materials.’ The search scheme proposed in this paper is shown in Figure 1. The search scheme consists of two parts; (a 1), a quantitative evaluation of the interactions and an identification of the elemental material in the target composite for which a substitute is to be found and (a 2), a search for new elemental materials to be substituted for the identified elemental material.
(See Appendix A. for details). Section 2 describes the interaction analysis method we use to evaluate quantitatively the interactions that occur among materials. In Section 3, we consider as a case study composites consisting of three materials (resin, filler, and additive). In addition we focus on the one interacting material that makes the largest contribution to the performance of the composite, using the bending modulus as the performance indicator, and we search for a new material to substitute it. To demonstrate that the proposed scheme works, we then apply the search scheme proposed in this paper to perform the search in the case study. Finally, we present our conclusions in Section 4.

Although for simplicity we consider in Section 2 a case in which only three kinds of materials constitute a composite material, it is easy to extend this interaction analysis method to composites consisting of four or more materials.

2. Interaction analysis method

2.1. Predictive model decomposed into interactions

Let us consider interactions represented by two and three Features, i.e.

$$x_i^{n_i} x_j^{n_j}, \quad x_i^{n_i} x_j^{n_j} x_k^{n_k},$$  \hspace{1cm} (1)

where $n_i, n_j,$ and $n_k$ are integers greater than 0; $i\neq j\neq k$; and $x_i$ is the feature representing the addition ratio of elemental material $i$. In this section, in order to evaluate the interactions between features quantitatively within the framework of the predictive model represented by Equation (A1), we first derive an equation containing the interactions based on Equation (A1), and we then define indicators to evaluate the interactions represented in that equation. From Equation (A1), the performance indicator $y$ can be expressed as
\[ y = \exp\left( \sum_{i=1}^{N} c_i x_i + c_0 \right) = A \prod_{i=1}^{N} e^{c_i x_i}, \]  

where \( A = e^{c_0} \). Assuming a composite material to be composed of three elements – matrix, filler, and additive—Equation (2) can then be written for a composite material \( a \) as follows:

\[ y_a = A e^{c_{ma} x_m} e^{c_{fa} x_f} e^{c_{aa} x_a}, \]  

where \( c_{ma} \), \( c_{fa} \), and \( c_{aa} \) are respectively the partial regression coefficients for the matrix \( (m) \), filler \( (f) \), and additive \( (a) \) in the composite material \( a \). Here, \( x_m, x_f, \) and \( x_a \) are the features – i.e., the addition ratios of the specific matrix, filler, and additive in the composite material \( a \)—and \( y_a \) is the performance indicator for that composite material. Expanding the exponential function in Equation (3) into a Maclaurin series, we get

\[ y_a = A \sum_{n_{ma}, n_{fa}, n_{aa}=0}^{\infty} \frac{1}{n_{ma}! n_{fa}! n_{aa}!} c_{ma}^{n_{ma}} c_{fa}^{n_{fa}} c_{aa}^{n_{aa}} x_m^{n_{ma}} x_f^{n_{fa}} x_a^{n_{aa}}, \]  

where \( n_{ma}, n_{fa}, \) and \( n_{aa} \) are integers greater than or equal to 0 and we assumed that the features are standardized with mean 0 and variance 1. From Equation (4), it is clear that Equation (2) contains products of features, i.e., interactions. Using Equation (4), we therefore next define indicators for evaluating these interactions. Hereafter, we use the following procedure to define these indicators: (I) From Equation (4), we derive an expression composed of two- or three-factor interactions. (II) As discussed in Section 2.2, below, we define indicators for evaluating those interactions based on the equation derived in (I).

First, let us focus on (I). Using Equation (1), we can express Equation (4) as follows:

\[ y_a = A \left( 1 + \sum_{n_{ma}=1}^{\infty} \frac{1}{n_{ma}!} c_{ma}^{n_{ma}} x_m^{n_{ma}} + \sum_{n_{fa}=1}^{\infty} \frac{1}{n_{fa}!} c_{fa}^{n_{fa}} x_f^{n_{fa}} + \sum_{n_{aa}=1}^{\infty} \frac{1}{n_{aa}!} c_{aa}^{n_{aa}} x_a^{n_{aa}} \right) + \sum_{n_{ma}=1}^{\infty} \sum_{n_{fa}=1}^{\infty} \sum_{n_{aa}=1}^{\infty} \frac{1}{n_{ma}! n_{fa}! n_{aa}!} c_{ma}^{n_{ma}} c_{fa}^{n_{fa}} c_{aa}^{n_{aa}} x_m^{n_{ma}} x_f^{n_{fa}} x_a^{n_{aa}}. \]  

The 2nd–4th terms on the right-hand side of Equation (5) are not interactions but instead are terms consisting of single features. Also, the 5th–7th terms on the right-hand side of Equation (5) are terms consisting of two-factor interactions, and the 8th term contains three-factor interactions. It is therefore essential to calculate these terms in order to define indicators for use in evaluating the interactions. Although each term is described by an infinite sum, we can avoid calculating the infinite sum using the following relation [27]:

\[ \sum_{n=1}^{\infty} \frac{1}{n!} x^n = e^x - 1. \]  

Using this equation, we can rewrite the 5th–7th terms on the right-hand side of Equation (5) as follows:

\[ \sum_{n_1, n_2=1}^{\infty} \frac{1}{n_1! n_2!} c_{i_1}^{n_1} c_{i_2}^{n_2} x_{i_1}^{n_1} x_{i_2}^{n_2} = (e^{x_{i_1}} - 1)(e^{x_{i_2}} - 1). \]

In the same way, the 8th term on the right side of Equation (5) is

\[ \sum_{n_1, n_2, n_3=1}^{\infty} \frac{1}{n_1! n_2! n_3!} c_{i_1}^{n_1} c_{i_2}^{n_2} c_{i_3}^{n_3} x_{i_1}^{n_1} x_{i_2}^{n_2} x_{i_3}^{n_3} = (e^{x_{i_1}} - 1)(e^{x_{i_2}} - 1)(e^{x_{i_3}} - 1). \]

Thus, Equation (5) becomes:

\[ y_a = A \{ 1 + (e^{c_{ma} x_m} - 1) + (e^{c_{fa} x_f} - 1) 
\]

\[ + (e^{c_{aa} x_a} - 1) + (e^{c_{ma} x_m} x_f - 1)(e^{c_{fa} x_f} - 1) 
\]

\[ + (e^{c_{ma} x_m} x_a - 1)(e^{c_{aa} x_a} - 1) 
\]

\[ + (e^{c_{fa} x_f} x_a - 1)(e^{c_{aa} x_a} - 1)(e^{c_{ma} x_m} x_f x_a - 1) \}, \]

where the 2nd–4th terms on the right-hand side of Equation (9) correspond to the 2nd–4th terms on the right-hand side of Equation (5); the 5th–7th terms on the right-hand side of Equation (9) correspond to the 5th–7th terms on the right-hand side of Equation (5); and the 8th term on the right side of Equation (9) corresponds to the 8th term on the right side of Equation (5). Hence, Equation (9) is numerically computable and its physical meaning is clear. The derivation of Equation (9) accomplishes procedure (I) above. However, the 2nd–4th terms on the right-hand side of Equation (9)—i.e., the 2nd–4th terms on the right-hand side of Equation (5)—also include the linear terms \( c_i x_i \). To define and describe the indicators in Section 2.2, it is convenient to show the equation in which the 2nd–4th terms on the right-hand side of Equation (9) are separated into linear and nonlinear terms. Thus,
Equation (10) consists of the sum of a constant term, a linear term, and a nonlinear term. This is the equation comprising two/three factors that was being sought in procedure (1).

2.2. Indicators for evaluating the interactions

To accomplish procedure (II), we next define indicators for evaluating the interactions quantitatively based on Equation (10). The purpose of defining these indicators is to use them to determine the elemental materials that are subject to substitution among the elemental materials that make up the target composite, as described in (a 1) of Section 1. To achieve this purpose, we define indicators that can address the following four questions quantitatively: (β 1): How much does the nonlinear term contribute to the performance of the target composite material? (β 2): If the contribution of the nonlinear term to the performance is not negligible, how much do the interactions contribute to the performance of the target composite material? (β 3): If the contribution of the interactions to the performance is not negligible, how large are the contributions of two- or three-factor interactions to the performance? (β 4): If the contribution of two-factor interactions to the performance is not negligible, how great are the contributions of the interactions between any two elemental materials (matrix–filler, matrix–additive, or filler–additive) to the performance?

First, we define an indicator for evaluating (β 1) as follows, using the linear and nonlinear terms identified in the previous section:

\[
R_n = \frac{\tilde{y}_n}{|\tilde{y}_n| + |\tilde{y}_l|},
\]

where \(\tilde{y}_n = \tilde{y}_{n1} + \tilde{y}_{n2} + \tilde{y}_{n3}\), and

\[
\tilde{y}_l = c_{m_n}x_{m_n} + c_{l_f}x_{l_f} + c_{a_n}x_{a_n},
\]

\[
\tilde{y}_{n1} = (\varepsilon^{m_n}x_{m_n} - 1) + (\varepsilon^{l_f}x_{l_f} - 1) + (\varepsilon^{a_n}x_{a_n} - 1) - \tilde{y}_l,
\]

\[
\tilde{y}_{n2} = (\varepsilon^{m_n}x_{m_n} - 1)(\varepsilon^{l_f}x_{l_f} - 1) + (\varepsilon^{m_n}x_{m_n} - 1)(\varepsilon^{a_n}x_{a_n} - 1)
\]

\[
+ (\varepsilon^{l_f}x_{l_f} - 1)(\varepsilon^{a_n}x_{a_n} - 1),
\]

\[
\tilde{y}_{n3} = (\varepsilon^{m_n}x_{m_n} - 1)(\varepsilon^{l_f}x_{l_f} - 1)(\varepsilon^{a_n}x_{a_n} - 1).
\]

In Equation (11), \(\tilde{y}_n\) is the linear term represented by the sum of the 2nd–4th terms on the right-hand side of Equation (10), and \(\tilde{y}_n\) is the nonlinear term represented by the 5th and subsequent terms on the right-hand side of the same equation. The absolute value of Equation (11) thus represents the proportion of the nonlinear term in the total of nonlinear and linear terms. Note that Equation (11) itself is not an absolute value; it can be negative because the value of \(\tilde{y}_n\) can be negative. This can occur because \(\tilde{y}_n\) contains terms of the form \(e^x - 1\), where \(x\) is the addition ratio, which can be negative when the partial regression coefficient \(c\) becomes negative. If \(\tilde{y}_n\) is negative, it decreases the performance indicator. On the other hand, if \(\tilde{y}_n\) is positive, it increases the performance indicator. Note that all the indicators defined in this section express the proportion of the quantity that contributes to increasing or decreasing the performance indicator (the quantity in the numerator) among the quantities contained in the denominator.

Next, we define an indicator for evaluating (β 2):

\[
\tilde{R}_l = \frac{\tilde{y}_l}{|\tilde{y}_n| + |\tilde{y}_l|},
\]

where \(\tilde{y}_l = \tilde{y}_{n2} + \tilde{y}_{n3}\). Equation (16) represents the proportion of the interaction to the combination of nonlinear terms comprising single features (\(\tilde{y}_{n1}\)) and interactions (\(\tilde{y}_l\)) that contributes to increasing or decreasing the performance.

Similarly, we define an evaluation indicator for (β 3) by the following equation:

\[
\tilde{R}_{nt} = \frac{\tilde{y}_{nt}}{|\tilde{y}_{n2}| + |\tilde{y}_{n3}|},
\]

where \(a = 2, 3\). This index represents the proportion of two- or three-factor interaction that contributes to an increase or decrease in performance.

Finally, we define an indicator for evaluating (β 4) by the following equation:

\[
\tilde{R}_{2\beta} = \frac{\tilde{y}_{2\beta}}{|(\varepsilon^{m_n}x_{m_n} - 1)(\varepsilon^{l_f}x_{l_f} - 1)| + |(\varepsilon^{m_n}x_{m_n} - 1)(\varepsilon^{a_n}x_{a_n} - 1)| + |(\varepsilon^{l_f}x_{l_f} - 1)(\varepsilon^{a_n}x_{a_n} - 1)|}.
\]
where \( y, \delta = m_a, f_a, a_a \). Equation (18) represents the proportion of the interaction between two elemental materials (matrix–filler, matrix–additive, or filler–additive) that contributes to increasing or decreasing the performance indicator among the two-factor interactions. Using the four indicators defined above, it is possible to address quantitatively the questions \((\beta 1)\) through \((\beta 4)\).

This method for determining which of the elemental materials are to be substituted in the target composite material using the aforementioned indicators to answer questions \((\beta 1)\) through \((\beta 4)\) in order is the interaction analysis method defined in this paper.

### 3. Case study

In this section, we consider as a case study a search for alternative materials in a composite material consisting of three elements – matrix (resin), filler, and additive – and we apply the search scheme described in Appendix A, with the bending modulus as the performance indicator. The purpose of this section is to demonstrate that the proposed scheme works using this case study.

We begin by stating the following two points: First, for composites with interactions, we identify the elemental material involved in the interaction that makes the largest contribution to the performance indicator using the interaction analysis method described above, and we then search for a new elemental material to substitute it. Second, we demonstrate experimentally that the composite material obtained by substituting the elemental material obtained through this search has the same bending modulus as the composite material before the substitution. The search scheme for this case study follows the procedure described in Appendix A. In Appendix A, we did not present detailed determinations of the specific elemental materials to be substituted because they depend on the problem under investigation, but we describe them here.

We assembled data for the composites that consist of performance measurements and the features defined in Equation (A1), and we then determined the composites with interactions contained in these data by the following procedure. First, we divided the data for the composites into training data and test data, and we constructed two types of predictive models using the training data: a predictive model using the natural logarithm of the performance indicator \( y \) as the objective variable and the features \( x \) as the explanatory variables (Model A) and a predictive model using the performance indicator \( y \) itself as the objective variable and the features \( x \) as the explanatory variables (Model B). Next, we predicted the performance from the test data using these two predictive models, calculated the absolute values of the residuals, and compared the predicted and measured values. In this evaluation, if the absolute value of the residual for Model A is lower than that for Model B using the test data, the nonlinear term cannot be ignored. The composite materials in the data that fulfill this condition are therefore subject to evaluation of the interactions. The reason for this is that the interactions belong to the nonlinear term in Model A [Equation (10)]. Finally, we identify the elemental materials involved in the interactions that make the greatest contribution to the performance by evaluating the interactions quantitatively using the interaction analysis method described in Section 2. The search scheme for the case study described above is shown in Figure 2.

#### 3.1. Collection of bending-modulus data

For each of the three elements, we obtained nine different kinds of elemental materials. From these elemental materials, we fabricated composites using one selected from each of the three types of elements, and we measured their bending moduli using the following procedure. We weighed each of the 27 different matrices, fillers, and additives to obtain the specified addition ratios, and we mixed them in a micro-compounder (MC15, Xplore Instruments) to obtain the composite compound. The temperature at the time of compounding was 230 °C, the rotation speed was 80 rpm at the time of materials feeding, and 130 rpm after the materials were fed, and the materials were mixed for 5 seconds. Then, we used a small injection-molding machine (IM12, Xplore Instruments) to fabricate a molded part from the compound in order to obtain a composite specimen in the shape specified in ISO 527-2-1 BA (2012). The injection temperature was set to 230 °C, the temperature of the mold was set to 60 °C, the injection pressure was set to 10 bar, and the molding time was set to 18 seconds. We measured the bending modulus using a universal materials-testing machine (Tensilon RTC-1250, A&D Co., Ltd.), utilizing the specimens of composite materials fabricated as described above. The test conditions were as follows: the distance between fulcrums was 40 mm, the indenter and the support radius were 5 mm in diameter, and the test speed was 2 mm/min. The bending modulus was calculated from the stress gradient in the strain interval 0.05 % – 0.25 %. We measured the bending moduli for a total of 38 composites with different kinds of elemental materials and different addition ratios using the fabrication conditions described above.

#### 3.2. Construction of predictive models

To select data from the collected data for composite materials, for which the nonlinear term cannot be neglected, we grouped the datapoints into 38 datasets
using the following procedure. From the histogram of the measured bending-modulus values shown in Figure 3, we selected three data points from the region of high bending modulus – over 15 GPa – and three selected randomly from the region of bending modulus smaller than 15 GPa. We created 6 and 32 test and training datasets, respectively. Using the information in the training dataset, we then constructed two predictive models – namely, Model A and Model B—using the PLS regression method [28]. In order to construct these two predictive models, it is necessary first to define the objective and explanatory variables. As described in Section 3, we used the natural logarithm of the bending modulus as the objective variable for Model A and the bending modulus itself as that for Model B. As the explanatory variables for both predictive models we used the addition ratios of the 27 different kinds of elemental materials. It is necessary also to determine the optimal number of latent variables in the construction of the predictive models using the PLS regression method. We therefore considered a predictive model constructed using the PLS regression method consisting of one to ten latent variables, and after performing leave-one-out cross-validation (LOOCV) and calculating the root-mean-square error (RMSE), we focused on the minima in the dependence of the RMSE on the number of latent variables, and we determined the optimal number of latent variables to be the smallest number that minimized the RMSE. The results of using LOOCV within the framework of the PLS regression method to determine the optimal number of latent variables are shown in Figure 4. This figure shows that the optimal number

Figure 2. Alternative-Materials search scheme for case studies. This scheme follows the search scheme described in Appendix A. Model A and Model B are predictive models described in the text. The quantities \( y_a \) and \( y_c \) are the performance indicators for the composite material before and after substitution by the candidate material determined using ISOMAP, respectively. The quantities \( |\varepsilon_i| (i = A, B) \) are the absolute values of the residuals computed for model i using the test data. The brackets labeled (a'1) and (a'2) in this figure correspond to (a 1) and (a 2) in Figure 1, respectively. The applied interaction analysis method is indicated by the red box.
of latent variables was four for both Model A [Figure 4(a)] and Model B [Figure 4(b)]. We used these two predictive models constructed using the PLS regression method to compare the prediction accuracy for the test data. Finally, using these two predictive models, we obtained the predicted values of the bending moduli for the test data, and we calculated the absolute values of the residuals from these values. Figure 5 shows the absolute values of the residuals for test data C1–C6 (See Appendix B, for the composition of C1–C6.). This figure shows that for the three composites (C4, C5, and C6) in the region of bending modulus over 15 GPa, the absolute values of the residuals for Model A were smaller than those for Model B. That is, these test data were predicted more accurately by Model A than by Model B.

The test data C4–C6—which have bending moduli of 15 GPa or higher—are thus composites for which the nonlinear term cannot be ignored. Therefore, we used the composite materials corresponding to these data (i.e., test data C4–C6) as targets for evaluating the interactions. However, for convenience, we chose only the data for composite material C4 and used it as the target for evaluating the interactions.

3.3. Identification of the materials targeted for substitution using the proposed interaction analysis method

In this section, in order to determine which elemental material is to be substituted in composite material C4, we use the indicators defined in Section 2.2, to identify the interaction that makes the largest contribution to the bending modulus.

We made this decision in the following way. First, we calculated the indicators $R_a$, $R_i$, $R_{ia}$, and $R_{iag}$ using the results predicted for composite material C4 by Model A. From the results of this calculation, we identified the interaction that makes the largest contribution to the bending modulus. We then determined the elemental materials involved in this interaction. Finally, from the elemental materials so identified, we selected the ones to be substituted.

We calculated the four indicators for composite material C4, and the results are shown in Figure 6 (See Appendix C, for the quantities required for calculating the indicators.). The value of $R_a$ was 0.51 [Figure 6(a)]. This shows that in this predictive model, the nonlinear term increases the bending modulus, and it accounts for about 51% of the total of both the linear and nonlinear terms. Here, the nonlinear term in composite C4 is calculated to be 5.63 GPa, which is not a small amount. We then calculated $R_i$, and found this value to be 0.44 [Figure 6(a)]. This interaction increases the bending modulus, and this contribution accounts for 44% of the total nonlinear term. The value of the interaction for composite C4 is 2.46 GPa, which also is not a small amount. To examine the proportions of two- and three-factor interactions quantitatively, we show the calculated results for $R_{i2}$ and $R_{i3}$ in Figure 6(b), which shows that the value of $R_{i2}$ was 0.82 and $R_{i3}$ was $-0.18$. This implies the following: The two-factor interactions increase the bending modulus by about 82% of the total (two- and three-factor) interactions. On the other hand, the three-factor interactions decrease the bending modulus.
modulus, accounting for about 18 % of the total interaction. This result shows that the two-factor interactions make the largest contribution to the bending modulus. Finally, to identify which of the two-factor interactions – i.e., those among the matrix–filler, matrix–additive, and filler–additive interactions – make the largest contribution to the bending modulus, we show the calculated results for $r_{2mf}$, $r_{2ma}$, and $r_{2fa}$ in Figure 6(c). These values were $-0.16$, $-0.08$, and $0.76$, respectively. Thus, both the matrix–filler and matrix–additive interactions act to decrease the bending modulus, and their contributions account for 16 % and 8 %, respectively, of the total of the two-factor interactions. On the other hand, the filler–additive interaction acts to increase the bending modulus, and it accounts for 76 % of the total contribution of the two-factor interactions. The contribution of the filler–additive interaction to the bending modulus is thus much higher than that of either the matrix–filler or matrix–additive interactions.

By applying the proposed interaction analysis method to composite C4, we thus found that the filler–additive interaction makes the largest contribution to the bending modulus. Therefore, in this case study, since the chemical structure of the additive in composite C4 is known, we decided to search for an elemental material that can be used to substitute this additive used in composite C4.

### 3.4. Materials search for alternatives and measurements of bending moduli

For simplicity, we call the additive used in composite material C4 additive a4. The purpose of this section is to search for a new additive to substitute additive a4 and to verify experimentally that a composite with the new additive has the same bending modulus as composite C4; thus, demonstrating the use of the proposed search scheme.

In order to search for a new additive to substitute additive a4, we obtained candidate compounds using the following procedure. We used the simplified molecular input line entry system (SMILES) to obtain data from PubChem [29] for compounds from TCI (Tokyo Chemical Industry), and we extracted those compounds that did not contain metal atoms and which had molecular weights greater than 100 and less than 900. Then, using the program alvaDesc [30], a tool to calculate and analyze molecular descriptors and fingerprints, we calculated 3885 two-dimensional (2D) molecular descriptors from the SMILES for the extracted compounds and for additive a4. In addition, we predicted the decomposition temperature for each additive using a predictive model for the decomposition temperature based on 2D molecular-descriptor data for the TCI compounds. In this way, we selected 1309 compounds with a probability greater than 80 % of having a decomposition temperature above 240 °C as candidate compounds for the search for new additives.

There are two points to note here. First, the reason for considering the decomposition temperature of the additive is because the composite specimens for which the bending moduli are to be measured are prepared using high-temperature heating (230 °C) in a compounding machine. If the matrix, filler, or additive – which are put into a kneading machine to...
fabricate the composite material – decompose at high temperature, the composite material will not consist of the specified elemental materials. The second point concerns the construction of a predictive model for the decomposition temperature. This was done in the following way. We prepared 20 compounds with known decomposition temperatures from the SIMLES data. We then constructed a predictive model for the decomposition temperatures using a Gaussian-process regression method with an radial basis function kernel [31]. The objective variable for this model was the decomposition temperature. The explanatory variable was a 2D molecular descriptor that was calculated from alvaDesc. In addition, we used the probability of the decomposition temperature exceeding 240 °C – i.e., the value integrated over the region where the decomposition temperature exceeds 240 °C in the probability distribution of decomposition temperature obtained from the predictive model, for this prediction.

Next, in order to evaluate the similarity between two compounds quantitatively, we need a feature value. We therefore performed principal-component analysis using the 2D molecular descriptors for a total of 1310 compounds consisting of additive a4 and candidate materials, and we then used as features the 1st to 400th principal components obtained from this analysis, which made a cumulative contribution of 100 %. Using these candidate compounds and features, we used the complete isometric feature mapping (ISOMAP) to search for compounds that are similar to additive a4 [32]. Figure 7 shows the ISOMAP results. The feature data for the 1309 candidate compounds were distributed as shown by the green dots in Figure 7, while additive a4 is indicated by the red dot (a4). From this distribution, we selected candidate compounds that are very similar to additive a4 — i.e., candidate compounds A1 and A2, which are very close to the red dot in Figure 7 — as experimental candidates. On the other hand, if a compound far from the red point is used as a substitute material, the bending modulus of the composite is expected to be much different from that of composite C4. To confirm this, we also selected compounds far from the red point – i.e. those indicated by N1 and N2 — as candidates for the experiment. We then fabricated four new composites by substituting the additive in composite C4 with the four additives selected above, and we measured their bending moduli. The fabrication conditions for these composites were the same as those described in Section 3.1. The addition ratio of each elemental material in the new composite is the same as that in composite C4; i.e. 40 % for the matrix, 40 % for the filler, and 20 % for the additive.

Table 1 shows the experimental results for the bending moduli of the four composites with different additives. This table shows that the bending moduli of the composites with A1 and A2 as additives – i.e. Composite 1 and Composite 2 — are 18.51 GPa and 18.80 GPa, respectively. On the other hand, those of the composites with additives N1 and N2—namely, Composite 3 and Composite 4 — were 20.10 GPa and 12.86 GPa, respectively. In order to see how much the bending moduli of Composites 1–4 differ from that of composite C4, we calculated the absolute values of the difference between the bending modulus of each Composite and composite C4, as shown in Table 1. The values of Δ for Composite 1 and Composite 2 are 0.62 GPa and 0.91 GPa, respectively; i.e. the difference between them and composite C4 is less than 1.0 GPa. On the other hand, the values of Δ for Composite 3 and Composite 4 are 2.21 GPa and 5.03 GPa, respectively, so the difference between them and composite C4 is more than 2.0 GPa. That is, the difference in bending modulus between C4 and Composite 1 or Composite 2 is small, while that between C4 and Composite 3 or Composite 4 is large. Thus, the composite materials in which additive a4 is substituted by a candidate compound with features similar to those of additive a4 has a bending modulus approximately equivalent to that of composite material C4.

In summary, we applied the search scheme proposed in this paper to search for alternative materials for composites, using the bending modulus as the performance indicator. We then verified experimentally that the composites fabricated using the additives obtained from the ISOMAP search had similar performance to the original composites. We therefore conclude that the search scheme proposed in this paper works. In addition, through the results presented so far, we have been able to use the proposed interaction analysis method to

![Figure 7. ISOMAP visualization of the feature values of 1310 compounds comprising candidate compounds and additive a4. Here, the features are the 1st–400th principal components obtained from principal-component analysis. The vertical and horizontal axes are axes determined by ISOMAP. The red dot represents additive a4, and the blue dot (A1) and purple dot (A2) represent compounds with features similar to those of additive a4. The orange dot (N1) and gray dot (N2) are compounds that differ significantly from the features of additive a4.](image-url)
Table 1. Measured bending moduli of composites made from additive A4 and from four different additives (unit: GPa). The quantity $\Delta$ is the absolute difference between the bending modulus of each composite and composite C4 (unit: GPa).

| Composite | Additive | Bending modulus | $\Delta$ |
|-----------|----------|-----------------|---------|
| C4        | A4       | 17.89           | 0.0     |
| 1         | A1       | 18.51           | 0.62    |
| 2         | A2       | 18.80           | 0.91    |
| 3         | N1       | 20.10           | 2.21    |
| 4         | N2       | 12.86           | 5.03    |

evaluate quantitatively how much the complex interactions between the elemental materials of a composite contribute to its performance. The present search scheme is therefore useful in searching for alternative materials for composites with complex interactions.

In this search scheme, only the chemical structures of the elemental materials to be substituted and of the candidate substitution materials need to be known; that is, it is not necessary to know performance values such as the bending moduli or chemical structures of all the materials in the training data used in the predictive model. We therefore expect the search scheme proposed in this paper to be very effective even in cases where it is difficult to search for materials using a predictive model, such as when the predictive model must be expressed by the addition ratio of a kind of material whose features are unclear.

4. Conclusion

In this paper, we considered a composite material comprising three types of elemental materials and evaluated a case where the performance depends on the features of each elemental material and the interaction among the elemental materials. For this case, we have proposed an interaction analysis method and data-driven search scheme for finding alternative materials. The proposed search scheme assumes that the predictive model and the features used in it are given by Equation (A1) and that at least features such as the chemical structures and physicochemical properties of the elemental materials to be substituted are known. The proposed search scheme consists of the following two points: The first is to determine which material is to be substituted by evaluating the interactions quantitatively, and the second is to search for a new material to substitute it in a way that does not rely on a search using a predictive model.

We proposed the interaction analysis method, which evaluates the interaction at the first point. In this method, the four indices, which can quantitatively evaluate the interaction, are defined based on the equation where the performance obtained from applying the Maclaurin expansion to Equation (2) is expressed by the linear and nonlinear terms. By applying these indices to the alternative material search scheme, it is possible to identify the elemental materials to be substituted in the composite materials with complex interactions.

As a case study, we considered a search for an alternative material in a composite consisting of three elements: a matrix (resin), a filler, and an additive. By applying the interaction analysis method to the composite material C4, we found that the interaction between filler and additive provided the largest contribution to the bending modulus. Next, we considered alternatives to the additives involved in this interaction. Using ISOMAP, two types of additives among the candidate compounds with similar features of the additives used in C4 were found. We measured the bending moduli of the composites obtained by substituting the additives used in C4 with the candidate compounds determined by this search, and we found the bending moduli of the substituted composites to be comparable to that of the unsubstituted composite C4. These results indicate that the search scheme proposed in this study works in the search for alternative materials for composites with interactions.

Using the proposed interaction analysis method, the search scheme described in this paper can evaluate quantitatively how much the complex interactions among materials contribute to the performance of a composite. In addition, it can search for alternative materials using unsupervised learning methods such as ISOMAP, even when it is difficult to use a predictive model, as long as only the chemical structure and physicochemical properties of the target material are known. We therefore conclude that the search scheme proposed in this paper provides a useful scheme for developing composite materials with complex interactions.

Acknowledgement

We thank H. Kita, Ph.D, T. Arai, T. Oshiyama, Ph.D., K. Kasahara, K. Murata, Ph.D., Y. Ikeda, Ph.D., and Y. Araki, Ph.D., for valuable discussion. The author would like to thank Enago (www.enago.jp) for the English-language review. All figures in this paper were generated using python Matplotlib [33].

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Michihiro Okuyama  http://orcid.org/0000-0003-1864-4317
Yukihito Nakazawa  http://orcid.org/0000-0002-8346-805X
Kimito Funatsu  http://orcid.org/0000-0002-9368-0302
References

[1] Rajak DK, Pagar DD, Kumar R, et al. Recent progress of reinforcement of composite materials: A comprehensive overview of composite materials. J Mater Res Technol. 2019;8(6):6354–6374. DOI:10.1016/j.jmrt.2019.09.068

[2] Sustainable Development Goals [Internet]. New York: United Nations; [cited 2021 Dec 12]; Available form: https://sdgs.un.org/goals

[3] Vold J, Ulven C, Chisholm B. Torrefied biomass filled polyamide biocomposites: mechanical and physical property analysis. J Mater Sci. 2015;50(2):725–732.

[4] Karunaratna MS, Lauer MK, Thiounn T, et al. Valorisation of waste to yield recyclable composites of elemental sulfur and lignin. J Mater Chem A. 2019;7(26):15683–15690. DOI:10.1039/C9TA03222C

[5] Ji A, Zhang S, Bhagia S, et al. 3D printing of biomass-derived composites: application and characterisation approaches. RSC Adv. 2020;37 (10):21698–21723. DOI:10.1039/D0RA03620J

[6] Chang BP, Gupta A, Muthuraraj R, et al. Bioreosourced fillers for rubber composites sustainability: Current development and future opportunities. Green Chem. 2021;23:5337–5378.

[7] Miyao T, Kaneko H, Funatsu K. Inverse QSPR/QSA analysis for chemical structure generation (from y to x). J Chem Inf Model. 2016;56 (2):286–299.

[8] Ikebata H, Hongo K, Isomura T, et al. Bayesian molecular design with a chemical language model. J Comput Aided Mol Des. 2017;31(4):379–391. DOI:10.1007/s10822-016-0008-z

[9] Griffiths RR, Hernández-Lobato JM. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. Chem Sci. 2020;11 (2):577–586.

[10] Ikeda Y, Okuyama M, Nakazawa Y, et al. Materials informatics approach to predictive models for elastic modulus of polypropylene composites reinforced by fillers and additives. J Comput Chem Jpn Int Ed 2021;7:2020–2007.

[11] Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. Phys Rev Lett. 2015;115(20):205901. DOI:10.1103/PhysRevLett.115.205901

[12] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–276. DOI:10.1021/acscentsci.7b00572

[13] Xue D, Balachandran P, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design. Nat Commun. 2016;7:11241.

[14] Terayama K, Sumita M, Tamura R, et al. Black-Box Optimization for Automated Discovery. Acc Chem Res. 2021;54(6):1334–1346. DOI:10.1021/acs.accounts.0c00713

[15] Khakurel H, Tausifq MFN, Roy A, et al. Machine learning assisted prediction of the Young’s modulus of compositionally complex alloys. Sci Rep. 2021;11 (1):1–10. DOI:10.1038/s41598-021-96507-0

[16] Kanno S, Imamura Y, Hada M. Alternative materials for perovskite solar cells from materials informatics. Phys Rev Mater. 2019;3(7):075403.

[17] Sawada R, Iwasaki Y, Ishida M. Boosting material modeling using game tree search. Phys Rev Mater. 2018;2(10):103802.

[18] Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat Mater. 2016;15(10):1120–1127. DOI:10.1038/nmat4717

[19] Wu S, Kondo Y, Kakimoto M, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput Mater. 2019;5:66.

[20] Honda T, Muroga S, Nakajima H, et al. Virtual Experiments by Deep learning on Tangible Materials. Commun Mater. 2021;2:88.

[21] Nimanpure S, Hashmi SAR, Kumar R, et al. Mechanical, electrical, and thermal analysis of sisal fibres/kenaf fibre hybrid polyester composites. Polym Compos. 2019;40(2):664. DOI:10.1002/poc.24706

[22] Prasad KE, Das B, Maitra U, et al. Extraordinary synergy in the mechanical properties of polymer matrix composites reinforced with 2 nanocarbons. Proc Natl Acad Sci. USA. 2009;106(32):13186. DOI:10.1073/pnas.0905844106

[23] Bishop CM. Pattern Recognition and Machine Learning. New York (NY): Springer; 2006.

[24] Wold S, Kettaneh-Wold N, Skagerberg B. Nonlinear PLS modeling. Chemomt Intell Lab Syst. 1989;7(1–2):53–65.

[25] Suzumura S, Nakagawa K, Umezue Y, et al. Selective Inference for High-order Interaction Features Selected in a Stepwise Manner. IPSJ Trans. 2021;14:1–11.

[26] Zhang Y, He X, Chen Z, et al. Unsupervised discovery of solid-state lithium ion conductors. Nat Commun. 2019;10(1):1–7. DOI:10.1038/s41467-019-13214-1

[27] Mayer JE, Montroll E. Molecular distribution. J Chem Phys. 1941;9(1):2–16.

[28] Wold S, Sjostrom M, Eriksson L. PLS-regression: A basic tool of chemometrics. Chemomt Intell Lab Syst. 2001;58(2):109–130.

[29] Kim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021 Jan 8;49(D1):D1388–D1395. DOI: 10.1093/nkx/gkaa971.

[30] Mauri A. alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In: Ecotoxicological QSARs. Methods in Pharmacology and Toxicology. Humana, New York (NY): Springer; 2020.

[31] Williams CK, Rasmussen CE. Gaussian processes in machine learning. Cambridge (MA): MIT press; 2006.

[32] Tenenbaum JB, Silva VD, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290(5500):2319–2323.

[33] Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90–95.
Appendix

Appendix A. Alternative-materials search scheme

The search scheme proposed in this paper targets composite materials consisting of three elements—matrix, filler, and additive—and it assumes that the predictive model can be represented by the following equation:

\[
\ln y = \sum_{i=1}^{N_x} c_i x_i + c_0, \tag{A1}
\]

where \( y \) is the performance indicator, \( c_i \) \((i = 0, 1, \cdots, N_x)\) are the partial regression coefficients, \( x_i \) \((i = 1, 2, \cdots, N_x)\) are the addition ratios of the matrices, fillers, and additives used in the training data [10], and \( N_x \) is the number of features. The predictive model in Equation (A1) does not include features such as the molecular structures and physicochemical properties of the elemental materials. Therefore, although this information is not necessary for all elemental materials in the training data used in Equation (A1), we assume that at least features such as the chemical structure and the physicochemical quantities of the material to be substituted are known.

In order to achieve (a1) in Figure 1, it is important to quantify the interactions among the elemental materials. This is performed using the interaction analysis method described in section 2. The results identify the elemental material to be targeted for substitution from among the elemental materials that make up the target composite material. The identification of this elemental material depends upon the specific problem being addressed. In Section 3, we describe this point for our case study.

The next step (α2) in Figure 1 is to search for a new elemental material to substitute the one identified in (α1). The objective of this search is to obtain a new composite material that has the same performance as the composite material before substitution by substituting another elemental material for the one identified in (α1). To accomplish this, we begin with a number of candidate materials for which the chemical structures are known. These candidate materials are then explored for prospective new materials. However, as mentioned in Section 1, the performance of composite materials not used in the training data cannot be predicted by Equation (A1). Therefore, it is difficult to search for new materials using the predictive model. We therefore introduce the features of the elemental material to be substituted and of the candidate material. If the feature data of the candidate material are similar to those of the elemental material to be substituted, we assume that the performance of the composite material after substitution by the candidate material will be equivalent to the performance of the composite material before substitution. Then, we search for candidate materials that fulfill this assumption by visualizing high-dimensional feature data using unsupervised learning methods such as ISOMAP. Next, we measure the performance of the composite material created by substituting the candidate material determined from the unsupervised learning search. If the measured performance of the composite material after substitution is in fact equivalent to that of the composite material before substitution, the search is terminated. However, if the performance of the substituted composite material is not equivalent to that of the composite material before the substitution, we return to the unsupervised learning search and search for another candidate material. This is the search scheme proposed in this paper.

Appendix B. Composition of composite materials in C1–C6

Table B1. Composition of composite materials in C1–C6. The column of elemental materials (matrix, filler, and additive) consists of the materials used as elemental materials and their addition ratio (unit: %). For example, m3(70) in the Matrix column means that 70 % of material m3 was added as matrix, and a hyphen in the filler column means that no filler was added. The bending modulus column represents the actual measured value of bending modulus (unit: GPa) in each composite material.

| Composite | Matrix | Filler | Additive | Bending modulus |
|-----------|--------|--------|----------|-----------------|
| C1        | m3(70) | f1(20) | a1(10)   | 3.25            |
| C2        | m4(80) | f2(15) | a1(5)    | 5.83            |
| C3        | m4(65) | f3(5)  | a2(25)   | 10.54           |
| C4        | m1(40) | f4(40) | a4(20)   | 17.89           |
| C5        | m2(40) | f4(40) | a4(20)   | 19.23           |
| C6        | m3(40) | f4(40) | a4(20)   | 21.22           |

Appendix C. Quantities used in the calculation of the indices of the interaction analysis method for C4

Table C1. The quantities used to calculate Equation (11) and Equations (16)–(18) for C4.

| \( y_i \) | \( x_{iA} \) | \( x_{iB} \) | \( y_{iB} \) | 1st term of Eq. (14) | 2nd term of Eq. (14) | 3rd term of Eq. (14) |
|-----------|-------------|-------------|-------------|----------------------|----------------------|----------------------|
| 1.30472586| 0.77821398  | 0.77999810  | −0.17625964 | −0.22963579          | −0.11978717          | 1.12942106           |