JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Functional Output Regression for Machine Learning in Materials Science

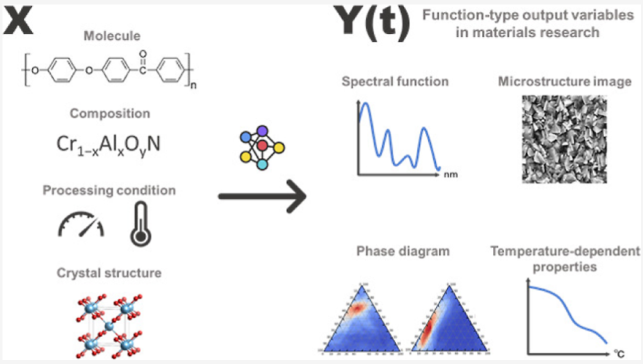
Megumi Iwayama, Stephen Wu, Chang Liu, and Ryo Yoshida*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 4837−4851

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: In recent years, there has been a rapid growth in the use of machine learning in material science. Conventionally, a trained predictive model describes a scalar output variable, such as thermodynamic, electronic, or mechanical properties, as a function of input descriptors that vectorize the compositional or structural features of any given material, such as molecules, chemical compositions, or crystalline systems. In machine learning of material data, on the other hand, the output variable is often given as a function. For example, when predicting the optical absorption spectrum of a molecule, the output variable is a spectral function defined in the wavelength domain. Alternatively, in predicting the microstructure of a polymer nanocomposite, the output variable is given as an image from an electron microscope, which can be represented as a two- or three-dimensional function in the image coordinate system. In this study, we consider two unified frameworks to handle such multidimensional or functional output regressions, which are applicable to a wide range of predictive analyses in material science. The first approach employs generative adversarial networks, which are known to exhibit outstanding performance in various computer vision tasks such as image generation, style transfer, and video generation. We also present another type of statistical modeling inspired by a statistical methodology referred to as functional data analysis. This is an extension of kernel regression to deal with functional outputs, and its simple mathematical structure makes it effective in modeling even with small amounts of data. We demonstrate the proposed methods through several case studies in materials science.

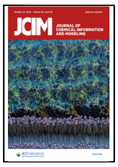


INTRODUCTION

Recently, there has been a growing trend to use machine-learning techniques to accelerate the process of designing and creating new materials in various domains of material science. Conventionally, machine-learning models are used to rapidly perform high-throughput virtual screening across millions or billions of candidate materials that span an enormous search space.[1−5] In general, a model describes physicochemical, electronic, thermodynamic, or mechanical properties as a function of the input materials, which are given in various forms, such as small- or macro-molecules, crystalline systems, chemical or raw material compositions, and their mixtures. To put the task into a machine-learning framework, such a non-numeric variable needs to be transformed into a fixed-length numeric vector called a *descriptor*, which represents the compositional or structural features of the given material.[5−16] Under the supervision of given data, a model is trained to learn the mapping from the vectorized features to their respective properties. In this workflow, the feature representation of the input materials plays a key role in boosting the predictive power.

There is a great deal of prior work on transforming material features into numeric vectors and constructing regression models or classifiers that represent the mapping from vectorized input materials to their output properties. A class of descriptors, referred to as molecular fingerprints, has long been studied in chemical informatics, which converts a chemical structure or molecular graph into an integer-valued vector according to the presence or absence or the number of occurrences of a particular chemical fragment, in which hundreds or thousands of fragments are considered.[10−14] Another type of molecular descriptor employs a quantitative representation of the topological or physicochemical features of a molecular system.[17−20] Chemical composition can be considered as a set variable consisting of a variable number of element species and their contents. There are a large volume of previous studies on the representation of such compositional features.[5−9] A crystal structure is typically vectorized by encoding the local structural environments of each atom and the neighboring relations of constituent atoms in a unit

Received: May 18, 2022
Published: October 10, 2022



https://doi.org/10.1021/acs.jcim.2c00626

cell.[7−9,21] In recent years, there has also been an increasing trend in treating a material structure as a graph and in modeling its properties using graph neural networks (NNs).[22−24] A natural representation of the chemical structure is created on a labeled graph. A periodic configuration of atoms in a crystalline system can also be translated into a graph called a crystal graph, which represents the coordination of constituent atoms in infinitely arranged unit cells.[22] In addition, when predicting the properties of a composite system from its microstructure, it is natural to treat the microstructure as an image. In the study of composite materials, scanning electron microscopy (SEM) and transmission electron microscopy (TEM) are widely used to observe the surface or interfacial structure of the fabricated materials. By treating the microstructure as an image, supervised learning can be performed by regressing real-valued output properties onto the space of the microstructure images, as in computer vision and image recognition.[25,26] Other representation methods have also been investigated for various material systems, such as topological feature representation of disordered material systems using persistent homology,[27] identification of multi-component materials based on the spectral function of powder X-ray diffraction,[28] and prediction of reaction outcomes in organic synthesis based on string representations of product and reactant molecules.[29,30]

As mentioned above, most previous studies have considered the ordinary problem setting of supervised learning, where an input variable is given as a relatively high-dimensional vector encoding material features, and the output is a scalar or low-dimensional real-valued vector, for example, a few sets of physicochemical properties or a class label indicating structural species or the level of physical features. On the other hand, there are many potential problem settings in material science, where the output variable is inherently ultra-high-dimensional or multidimensional (e.g., a functional-type output). However, the methodology of supervised learning in such scenarios has not been well studied. For example, in the task of predicting the ultraviolet−visible (UV−vis) absorption spectra of molecules, the input variable is given by a vectorized molecular structure, and the output variable is given as a function defined on the domain of wavelengths that represents the optical absorbance.[31] In the study of composite materials, it is important to qualitatively and quantitatively understand the influence of processing conditions such as temperature, pressure, and composition on the resulting microstructures. SEM and TEM are commonly used to examine microstructures. If we formulate the problem within a framework of supervised learning, the input is a real-valued vector encoding the processing condition and composition, and the output is given by an intensity matrix representing the grayscale microscopic image. This is a regression problem for multidimensional functional output variables. Alternatively, the problem can be reduced to an image generation task in computer vision. To solve such problems, various types of deep generative models, such as the conditional GAN (cGAN)[32] and encoder−decoder networks,[33] can be applied. In fact, there have been several previous studies in which cGAN was applied to the prediction of microstructures, as described above,[34−36] and an encoder−decoder model was applied to predict the UV−vis absorption spectra of organic molecules.[31] In addition, in statistical science, regression methods for functional output variables have long been studied in the context of functional data analysis,[37] which may also be applicable to solving the aforementioned problems.

In this study, we consider two unified frameworks for multidimensional functional output regression that can cover various potential applications in material science. The first approach employs cGAN, inspired by the study of microstructure images in Banko et al.[34] However, because cGAN has training instability in the adversarial learning process and weakness to limited amounts of data, we developed another framework, a statistical modeling that relies on the methodology of functional data analysis for functional output variables. The present model can be viewed as an extension of kernel regression to handle functional output variables and has a simpler mathematical form than the cGAN model architecture. As shown later, the method exhibits outstanding predictive performance even in cases where only a small amount of data is available. We demonstrate these two methods using three case studies. In the first two case studies, the optical absorption spectra of organic molecules in two different regions of UV−vis (170−780 nm) and near-infrared wavelength (NIR: 780−2500 nm) were predicted. The output variable is a spectral function in the wavelength domain. The number of training instances is approximately over 900 for the former case, whereas for the latter, the amount of data is limited as the number of training molecules is approximately 60. The objective of the third example is to predict the electron microscopic image of the microstructure for any given composition and processing conditions in the fabrication of thin-film composite materials. With these applications, we demonstrate the potential predictive ability of the proposed methods on small amounts of training data. We compare ordinary regression, which predicts a scalar output variable with a pre-quantified spectral feature,[38] with the present methods predicting the whole function directly, and show the superiority of the latter and its statistical mechanisms in relation to multitask learning.[39,40] The Python codes used in the case studies were distributed.[41]

■ PRELIMINARY

The present study deals with a supervised learning problem, in which the input variable $X \in \mathbb{R}^p$ is a p-dimensional descriptor vector and the output variable $Y(X, t) \in \mathbb{R}$ is given by a real-valued function of X and an additional argument $t \in \mathbb{R}^q$. Argument t corresponds to a coordinate in image space or to a wavelength at which the spectral function is defined. In the following sections, we describe potential applications in material science.

Spectral Prediction. Molecules undergo temporal transitions from their ground states to higher-energy excited electronic states in response to the absorption of light, such as UV−vis or NIR. The absorption wavelength is proportional to the inverse of energy. The absorbance spectrum, which represents the intensity of optical absorption as a function of wavelength, is determined by the excitation energy levels and the transition probabilities of the electronic states in a molecular system. Accurately predicting molecule-specific absorbance spectra is highly beneficial for various applications, such as the design of organic light-emitting diodes,[1,42] organic photovoltaic cells,[43] and UV filters.[44] Usually, absorption peak wavelengths are predicted from the excited states of electrons obtained ab initio, for example, by performing time-dependent density functional theory calculations.[45] However, owing to the

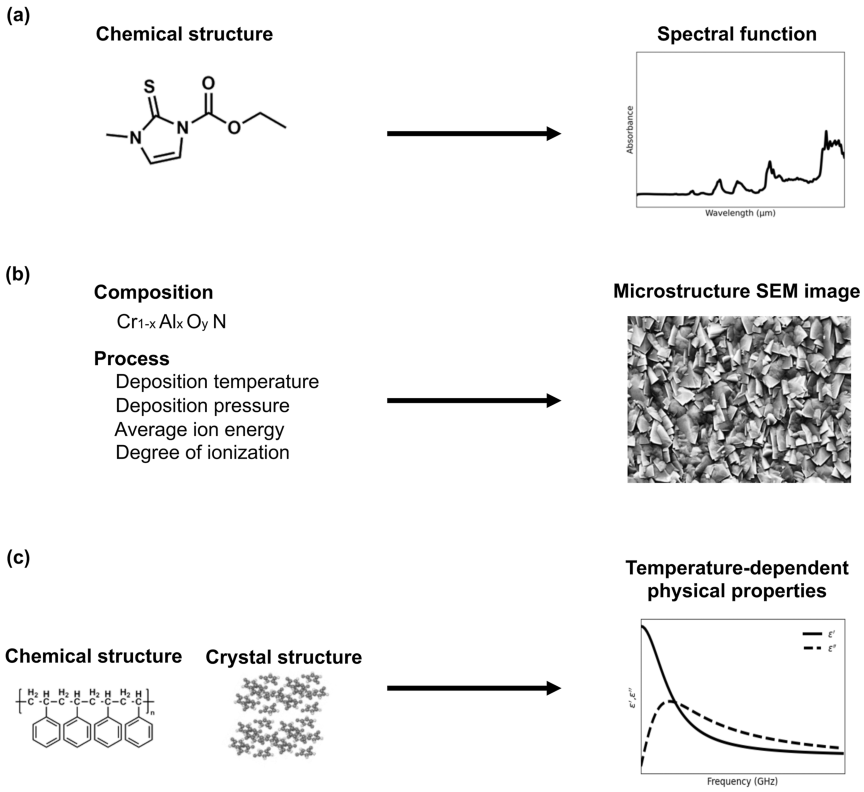


Figure 1. Three potential applications formulated as the problem of functional output regression: (a) prediction of optical absorption spectra based on the chemical structure of an organic molecule, (b) prediction of microstructure images of composite materials based on processing conditions and compositional features, and (c) prediction of frequency-dependent physical properties.

high computational cost, first-principles calculations are not useful for exhaustive molecular screening. Furthermore, while the location and intensity of a peak wavelength can be estimated ab initio, other functional features of the full spectrum, such as the full width at half maximum and absorbance integration in a wavelength interval, cannot be determined.

Here, we address the problem of spectral prediction using a fully data-driven approach that does not rely on ab initio calculations (Figure 1a). The input variable of model f consists of a descriptor vector X and a wavelength $t \in \mathbb{R}^1_+$, where the descriptor X encodes the structural and compositional features of the input molecule. The output variable is the spectral function $Y(X,t)$ of the optical absorbance with respect to varying molecules and wavelengths. In summary, with measurement noise ϵ, the model can be expressed as

$$Y(X, t) = f(X, t) + \epsilon \tag{1}$$

Suppose that for each of the n observed molecules $\{X_i | i = 1, ..., n\}$, the absorption spectrum $Y(X,t)$ is measured over m discretized wavelengths $\{t_j | j = 1, ..., m\}$. In this study, it is assumed that the observation points of the wavelength are common to all molecules: t_j is independent of the index i of the molecule. When the observed wavelengths vary across the molecules, one can obtain a series of complete data with the same observation points by smoothly interpolating the missing data points. The task then comes down to a regression problem for the high-dimensional vector-valued output, which is modeled by

$$y(X) = f(X) + e \tag{2}$$

where $y(X)^{\mathrm{T}} = (Y(X,t_1), ..., Y(X,t_m))$, $f(X)^{\mathrm{T}} = (f(X,t_1), ..., f(X, t_m))$, and $e^{\mathrm{T}} = (\epsilon_1, ..., \epsilon_m)$. As mentioned above, it is assumed that there are no missing data or that the missing data in the direction t have been interpolated beforehand based on a sufficient amount of observed data. However, the functional output kernel regression, which will be shown later, can naturally perform training with no special treatment for missing data, even if the observation points of t are quite sparse.

Machine learning for predicting the optical absorption spectra of molecular systems has not been extensively studied. To clarify the contribution of our work, we consider the recently published work of Urbina et al.[31] that relied on Seq2Seq[33] and its variant encoder−decoder architectures with a built-in attention mechanism. Seq2Seq, which is widely used in natural language processing, was utilized to learn mapping from tokenized SMILES strings[46] or pre-defined molecular descriptors of input chemical structures to the absorption spectra. On the other hand, we introduce a much simpler statistical model, the functional output kernel regression, aimed at stabilizing the learning process by reducing over-parameterization and achieving a high prediction accuracy even in cases where sufficient amounts of data are unavailable for model training. Another distinctive feature of the present method is its high degree of interpretability, which directly describes the occurrence of a peak in a specific wavelength range in relation to the presence or absence of molecular fragments encoded in the descriptor X.

In this study, we focus on the advantages of directly predicting the entire spectral function, rather than predicting a pre-defined univariate functional feature, such as the wavelength of maximum absorbance $\lambda_{\max}$. In the experiments

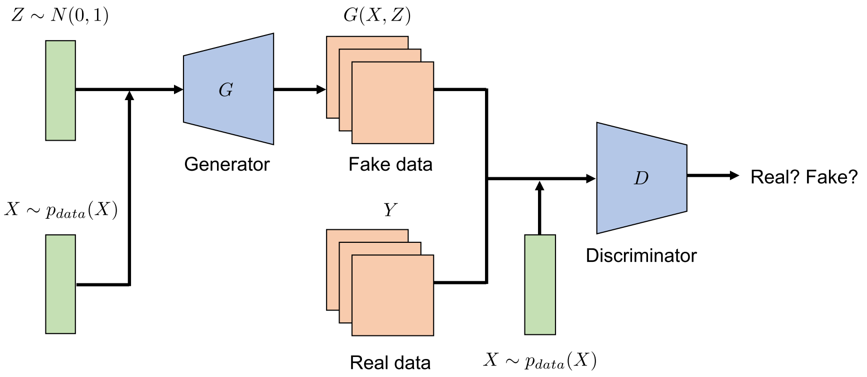


Figure 2. Model architecture of cGAN. The generator (G) represents a mapping from composition/process conditions or molecular descriptor X to a microstructure image or optical absorption spectral function. The additional input Z denotes a Gaussian random noise. The discriminator (D) determines, given an image or spectral function in addition to the conditional variable X, whether it is real or fake. Model architectures are detailed in the Supporting Information Note.

reported later, it was confirmed that the prediction accuracy of $\lambda_{\max}$ calculated from the predicted spectral function often significantly exceeds that of the conventional univariate output regression directly trained with the pre-quantified $\lambda_{\max}$. It should be noted that high-dimensional output variables $y(X) = (Y(X,t_1), ..., Y(X,t_m))$ are closely related. Learning a single model for multivariate outputs simultaneously can be considered a type of multitask learning. In multitask learning, multiple related tasks are learned simultaneously, allowing the model to recognize common mechanisms among target tasks and consequently improve the prediction accuracy of each task.[39] A similar learning mechanism is expected to work in regression with high-dimensional output variables.

Microstructure Image Prediction. Microstructures with varying morphologies, volume fractions, and grain-size distributions can be designed by controlling the composition of the material species and processing conditions.[34−36] Here, we consider the problem of predicting the microstructures for any given compositional and processing parameters. To treat the microstructure as a model output, we use an image obtained by optical or electron microscopy. In the development of composite materials, SEM or TEM has widely been used to analyze the surface and morphologies. Practically, for example, we aim to improve the mechanical properties of a material by controlling the composition and temperature to obtain a finer and more homogeneous grain structure. By defining a microstructure as an image, we can address the machine-learning task by utilizing various well-established techniques in image recognition and computer vision.[34−36] Specifically, the input variable X is given by a real-valued vector representing the compositional and processing parameters, and the output $Y(X,t)$ is defined as a microscopic image that takes a matrix or tensor form for a grayscale or color image, respectively. The variable t represents a two- or three-dimensional image coordinate, and its support is discretized into pixel or voxel positions. The model can therefore be written in the form of a multidimensional vector regression, in the same way as the model for the spectral function described above (Figure 1b). Alternatively, in the context of computer vision research, this task can be regarded as machine learning for conditional image generation.

Other Potential Applications. There are many other applications in material science where the prediction of functional outputs is applicable. Many physical properties are determined by temperature, pressure, and frequency in an external electric field (Figure 1c). The dielectric properties of a material, that is, the dielectric constant or dielectric loss tangent, are given as a function of frequency and temperature.[47] In this case, t comprises the two variables. In addition, various polymeric properties, such as the specific volume, linear expansion coefficient, elastic modulus, specific heat, and thermal conductivity, are also dependent on temperature. From the observed transition points in the temperature dependence curve, the glass transition temperature, crystal melting point, and crystallization temperature are calculated. For the imaging analysis of materials, a wide variety of microscopes such as SEM, TEM, and optical microscopes are commonly used, depending on the size of the object to be observed. Various technologies of three-dimensional measurements have been established to analyze the internal structures of materials. For example, three-dimensional TEM allows us to observe the morphology of objects ranging in size from tens to hundreds of nanometers in three dimensions.[48] X-ray computed tomography (CT) is a non-destructive technique that is often used in medical applications. With this method, we can observe microstructures ranging from a few micrometers to millimeters in size without polishing or etching the sample surface.[49] In addition, X-ray CT is a non-destructive inspection method that can be used to measure the fracture process of a material and the change in the material structure in response to heating in four dimensions (three-dimensional space plus time). Furthermore, synchrotron X-ray CT using high-brilliance synchrotron radiation can non-invasively measure the inner structure of materials with a high resolution of several hundred nanometers to several micrometers, even in metals where X-ray penetration is difficult.[50] In principle, the proposed regression method can be applied to a wide range of high-dimensional functional data.

■ METHODS

In this paper, we present two different regression methodologies for multidimensional functional outputs: deep generative modeling using adversarial learning and deep kernel regression for functional outputs. The former, originally developed for image-generation tasks,[34] is introduced to solve the two research subjects. The latter is a newly developed method for overcoming the limited learning performance of the deep generative models.

Conditional GAN. We construct an NN to handle the functional output regression, cGAN, which consists of a

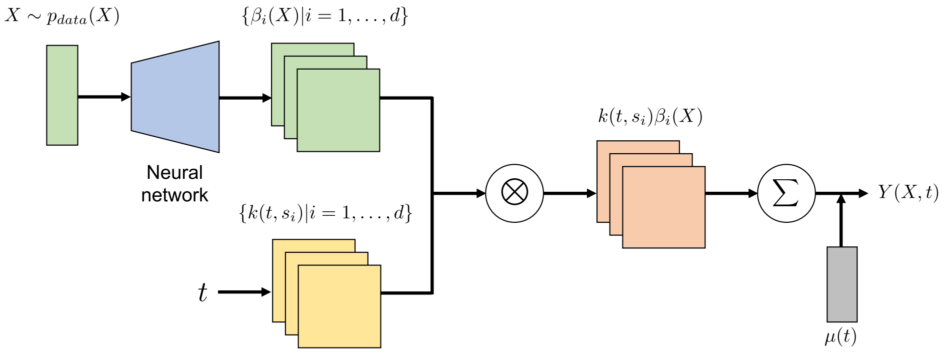


Figure 3. Model architecture of the functional output kernel regression. The coefficient functions $\{\beta_i(X)|i = 1, ..., d\}$ for a set of d pre-defined kernels $\{k(t,s_i)|i = 1, ..., d\}$ depend only on input X. The mapping from X to $\beta_i(X)$ is modeled by a fully connected or convolutional NN. The number d of kernel functions is appropriately controlled to be smaller than the number of observations for t, that is, $d \leq m$.

generative model called a generator (G) and a binary classifier called a discriminator (D). The model structure is summarized in Figure 2 (see the Supporting Information Note for more details). With the generator $Y = G(X,Z)$, the output vector of the m absorption values in eq 2 or the matrix of a grayscale microstructure image is modeled as Y. The input variables consist of X in the regression model (referred to as the conditional variable in cGAN) and random noise Z. The noise $Z \sim N(0,1)$ is assumed to follow a normal distribution with a mean of zero and unit variance. The input (X, Z) is first transformed into an embedding vector by passing through a fully connected embedding layer and is further transformed into a vectorized spectrum or a microstructure image $Y = y(X)$, as in eq 2, by passing a series of differently stacked hidden layers depending on the task. The discriminator $D(X,Y)$ is a binary classifier, in which the conditional variables X and Y are given as inputs. The discriminator judges whether object Y (a spectrum or an image) is real or fake. The discriminator D is modeled as a conventional fully connected NN or a convolutional NN.[51] The model structure, such as the number of layers and neurons in each layer, is determined based on the generalization performance in a separate validation data set, while maintaining the basic form described here. The detailed settings for each problem are described in the Supporting Information Note.

With this composite modeling, G and D are trained alternatively according to the following minmax strategy

$$\min_{G}\max_{D}\mathbb{E}_{(X,Y)\sim p_{\text{data}}(X,Y)}[\log\ D(X,\ Y)] + \mathbb{E}_{Z\sim N(0,1),X\sim p_{\text{data}}(X)}[\log(1-D(X,\ G(X,\ Z)))] \quad (3)$$

The first term becomes larger as $D(X,Y)$ increases, that is, when the discriminator D correctly identifies the input real object as real. The second term becomes large when $1 - D(X,G(X,Z))$ becomes large, that is, when D successfully recognizes the fake $Y = G(X,Z)$ to be fake. The discriminator D is learned such that the classification error is minimized, and G is trained to reduce the second term such that D is misrecognized. By alternately training G and D, we derive G, which can produce high-quality fake spectral functions or microstructure images for any given descriptor X. For the spectral prediction, we generate r random samples $\{Z_i|i = 1, ..., r\}$ ($r = 100$) and use the ensemble $f(X) = \frac{1}{r}\sum_{i=1}^{r} G(X, Z_i)$ of the learned generator to improve the smoothness of the predicted function.

cGAN can be regarded as a supervised learning technique for multidimensional output variables. Because cGANs have been intensively studied, particularly for image-generation tasks, when treating images as the output variable, we can take advantage of the wealth of tips and various extended works that have been accumulated in machine-learning research. However, like other conventional generative adversarial networks, cGANs suffer from instability during the learning process. In particular, vulnerability to small data sets has been pointed out in many previous studies. With limited amounts of training data, the discriminator can easily overfit the data to make a perfect true/false classification, which leads to gradient vanishing and halting of the learning process before a sufficiently accurate generator (the target regression model) is created. The solution to stabilize the adversarial training process is to balance the learning progress of G and D, but this is not an easy task. The primary reason for this is the over-parameterization of the generator caused by the high dimensionality of the vectorized object Y. In the case studies shown below, the dimension of Y is more than 170 or 2000 for spectral function prediction and up to 10,000 for microstructural microscope images with a resolution of 100 × 100. For example, to describe the mapping from X to the high-dimensional image object Y, it is necessary to introduce one or more transposed convolution matrices of large sizes. In applications shown later, the total numbers of model parameters in the trained generators reached the order of 2.7 million or 10 million for the spectral function prediction and 5.8 million for microstructure image prediction.

Kernel Regression with Functional Outputs. In addition to cGAN, we present another model with a simple, naturally interpretable model structure. The design concept is inspired by regression models for functional outputs, which have been studied in the context of functional data analysis. The functional output $Y(X,t)$ is modeled as follows

$$Y(X,\ t) = \sum_{i=1}^{d} k(t,\ s_i)\beta_i(X) + \mu(t) + \epsilon \quad (4)$$

The first term is the weighted sum of the d kernel basis functions $\{k(t,s_i)|i = 1, ..., d\}$. The kernel centers s_i are equally spaced in the domain of wavelengths or image coordinates. In this study, we use the Gaussian radial basis function (RBF) kernel as

$$k(t, s_i) = l \, \exp\left(-\frac{\|t - s_i\|^2}{2\sigma^2}\right) \tag{5}$$

The variance $\sigma^2 > 0$ and length scale $l > 0$ are hyperparameters adjusted based on the evaluation of the generalization performance. Alternatively, one may predetermine the hyperparameters based on the empirically known resolution of a measurement system or the inherent variation of $Y(X,t)$ in varying t for the physical system of interest. The regression coefficient $\beta_i(X)$ depends only on the input variable X, which is modeled by a NN, as described below. $\mu(t)$ is a baseline function estimated by imposing smoothness as a regularizer. ϵ denotes the noise term.

The overall model represents a system in which each kernel function pre-arranged in the domain of variable t is activated or deactivated depending on the input variable X, for example, the presence or absence of a specific fragment in a fingerprinted chemical structure X. One advantage of this modeling is that the number of parameters can be reduced by controlling the number of kernels d placed in the domain t. In the cGAN generator, the number of neurons in the output layer inevitably increases because it is constrained by the dimensionality of Y.

The regression coefficients $\{\beta_i(X)|i = 1, ..., d\}$ are modeled differently using NNs for the two tasks (Figure 3, Supporting Information Note). We use NNs with a structure similar to that of the generator in cGAN. In the prediction of the optical absorption spectra, the input X is given by a 1024 binary vector that encodes the chemical structure of a molecule using the extended connectivity fingerprint[10] with a radius of 3. The mapping from X to the d output coefficients is modeled by multiple blocks of stacked hidden layers, including a fully connected layer, batch normalization layer, and leaky ReLU activation function.[52,53] For the microstructure image prediction, as detailed later, the input X includes six processing parameters and a Gaussian noise. The mapping from X to the d output coefficients consists of multiple blocks of stacked layers, including a transposed conventional layer, batch normalization layer, and leaky ReLU activation function. Hyperparameters such as the number of convolutional layers and neurons are tuned based on a separate validation data set. See the Supporting Information Note for the procedure for hyperparameter tuning.

In model training, the following objective function is minimized with respect to the parameters in the model of $\{\beta_i(X)|i = 1, ..., d\}$ and the baseline function $\mu(t)$

$$L(\beta, \mu) = \sum_{(X,t)\in\mathcal{D}_{obs}} C(Y(X, t), \hat{Y}(X, t)) + \lambda \sum_{i=1}^{m} \sum_{j\in\mathcal{A}_i} (\mu(t_i) - \mu(t_j))^2 \tag{6}$$

The first term involves the discrepancy C between an observed $Y(X,t)$ and its prediction counterpart $\hat{Y}(X, t) = \sum_{i=1}^{d} k(t, s_i)\beta_i(X) + \mu(t)$. In the experiments reported later, for both spectral prediction and microstructure prediction, C is defined by an ordinary squared loss, which is summed over all observations of X and t. The second term is the regularization term for the baseline function. Regularization induces a smooth transition between $\mu(t_i)$ and $\mu(t_j)$ for an observation point t_i and its neighborhood $t_j \in \mathcal{A}_i$. In this formulation, the observed series of $Y(X,t)$ is allowed to have some missing values for t.

■ RESULTS AND DISCUSSION

We highlight the potential prediction capability and learning mechanism of the proposed methods by presenting three application examples. The objective of the first two examples is to predict the function of the UV−vis or NIR absorption spectrum of an organic molecule, taking its chemical structure as the input. In the first case, the number of samples was approximately 2200 or less, whereas in the second case, the number of samples was only 68. The objective of the third example is to predict electron microscopy images of microstructures from the composition and processing conditions of the fabricated thin metal films.

Prediction of UV−Vis Absorption Spectrum. Urbina et al.[31] produced two experimental data sets of UV−vis spectra that encompass different chemical spaces. In the paper, these data sets were referred to as Dataset I and Dataset II. Dataset I consists of the absorption spectra of 949 different commercial compounds measured using a high-performance liquid chromatography system. Dataset II contains the spectra of 2222 different commercially available pharmaceutical molecules, which were measured with a spectrophotometer in a multi-well plate format. The absorption spectra were measured at 181 and 171 wavelengths equally spaced in the UV−viz (220−400 nm) for Dataset I and Dataset II, respectively. Compared to Dataset I, Dataset II has a larger amount of data and a larger diversity of chemical structures. For model training and evaluation of prediction performance, approximately 70% of the compound set was randomly selected as the training set and the remaining approximately 15% as the test set. Partitioning of the data set was performed according to the compound species to avoid multiple spectral profiles of the same compound leaking into the test set. In Urbina et al.,[31] the results of applying two encoder−decoder architectures with LSTM cells and an attention mechanism, respectively, were reported, which were compared with the performance metrics of the present methods.

For the kernel regression, the observed wavelength range 220 or 230−400 nm was divided into 128 equally spaced segments, and the kernel centers s_1, ..., s_d ($d = 128$) were placed there. The set of hyperparameters consisted of the variance σ^2, length scale l in the RBF kernel, and the number of hidden layers and neurons in the NN. For each σ^2 and l, three and four grid points were set as candidates. The number of hidden layer blocks was set between 1 and 4. Once the number of blocks was determined, the number of neurons in each layer was determined, as shown in the Supporting Information Note (Figure S2 and Table S3). Correspondingly, the total number of candidates for the hyperparameter search was 48 (=4 × 4 × 3). The generalization performance of the trained model for each candidate was measured using the root mean square error (RMSE) on the validation data set, and the best combination was identified.

The hyperparameter of cGAN was given by the network structure of the generator and discriminator. As in the kernel regression, both network structures consisted of a series of stacked hidden layers, with each block consisting of a fully connected layer, a batch normalization layer, and a leaky ReLU activation function. The difference from the kernel regression was that the output variables were directly formed by the m absorbances of the different wavelengths in the spectral

Table 1. Comparison of the Prediction Accuracy for Optical Absorption Spectral Functions[a]

		RMSE	R^2	MAE	RMSE derivative
Dataset I	LSTM	0.169 ± 0.132	0.626 ± 1.166	0.119 ± 0.106	0.013 ± 0.010
	attention	0.154 ± 0.144	0.680 ± 1.230	0.091 ± 0.120	0.018 ± 0.020
	kernel	**0.111 ± 0.009**	**0.798 ± 0.043**	0.075 ± 0.006	**0.012 ± 0.001**
	cGAN	0.112 ± 0.014	0.786 ± 0.057	**0.071 ± 0.010**	0.053 ± 0.002
Dataset II	LSTM	0.064 ± 0.062	0.710 ± 0.472	0.047 ± 0.075	0.008 ± 0.006
	attention	**0.055 ± 0.071**	0.699 ± 0.259	**0.044 ± 0.052**	**0.006 ± 0.007**
	kernel	0.093 ± 0.004	0.655 ± 0.007	0.066 ± 0.001	0.009 ± 0.000
	cGAN	0.085 ± 0.002	**0.718 ± 0.022**	0.058 ± 0.003	0.016 ± 0.001
USGS	kernel	**0.076 ± 0.024**	**0.602 ± 0.200**	**0.053 ± 0.009**	**0.022 ± 0.003**
	cGAN	0.074 ± 0.003	0.493 ± 0.102	0.058 ± 0.009	0.528 ± 0.043

[a] Dataset I and Dataset II cover the UV−vis spectra, and the USGS spectral library covers data in the NIR domain. The performance metrics shown for the LSTM and attention-based encoder−decoder models are those reported in Urbina et al.[31]

function. The number of blocks to be searched was between 1 and 4, and the number of neurons in each layer was determined, as shown in the Supporting Information Note (Figure S1 and Table S1). The generalization performance of the models was measured using the RMSE of the validation data set to identify the best hyperparameter combination.

Through validation, in Dataset I, the numbers of stacked hidden layers for the NNs in the kernel regression, generator, and discriminator of cGAN were selected as 4, 4, and 2, respectively. In Dataset II, the number of stacked hidden layers for the NNs in the kernel regression, generator, and discriminator of cGAN were selected as 3, 4, and 3, respectively. The value of σ^2 was selected as 0.0005 for both data sets, and l was selected as 0.5 or 5 for Dataset I or Dataset II, respectively. To evaluate the performance of these models, we calculated the RMSE, coefficient of determination (R^2), and mean absolute error (MAE) between the m predicted spectral values and their observations for each test molecule and compared the median of each performance metric for the 150 and 342 test molecules. We also considered the differences in the spectral series and evaluated the gradient-level prediction performance using the same procedure. These performance measures are detailed in the Supporting Information Note.

Table 1 summarizes the means and standard deviations of the performance measures for the three independent numerical experiments; the prediction accuracy of the LSTM-based and attention-based encoder−decoder models reported in Urbina et al.[31] is also given. In Dataset I, the kernel regression outperformed the other three, but the difference with cGAN was not pronounced. The reason for the low prediction accuracy of the difference spectrum of cGAN is due to the use of randomly sampled Z in the calculation of the predicted spectra, which resulted in a loss of smoothness in the gradient of the spectrum. However, this problem can be solved by smoothing the predicted spectrum in the post-processing step. As the code for the encoder−decoder models is not distributed, we could not go further into the comparison of the models, but it should be concluded that in this case study, they have almost the same performance.

Here, the predicted spectral functions and their observed values are presented exhaustively to obtain a view of the predictive capability of the kernel regression and cGAN. Figures 4 and 5 show the prediction outcomes for 27 randomly selected test molecules in Dataset I and Dataset II, respectively. To obtain a more comprehensive view of the prediction accuracy, Supporting Information Note (Figures S3 and S5) also provides the results of 60 randomly selected test molecules in each data set (see also Figures S4 and S6 for the results of training). According to a careful visual inspection, it can be seen that the models retain a surprisingly high prediction accuracy. For a significant number of molecules, the positions of multiple peaks in the absorption spectrum and the shape of the function were almost perfectly predicted. In some cases, even features that are not visually noticeable, such as plateaus or tiny peaks, were captured appropriately. This observation suggests that the presence or absence of chemical substructures is a major factor in the optical absorption spectra of molecules. There was no significant difference between the kernel regression and cGAN in terms of capturing the broad trend of the spectral function. However, as mentioned above, cGAN requires some effort to detect the peak position because of the noise fluctuations of Z in the prediction equation. Even when smoothing is applied, unexpected false peaks can occur. Therefore, we conclude that the kernel regression has an advantage in terms of the spectral prediction.

Spectral Prediction with Limited Data in the USGS Spectral Library. The U.S. Geological Survey (USGS) Spectral Library Version 7[54] contains reflectance optical spectra of over 1000 different molecules, each of which was measured in three different sets of environments: laboratory, field, and airborne spectrometers. Of these samples, 68 organic compounds that were measured in the laboratory at room temperature were pre-extracted. The compiled data set contained no case where multiple spectra were assigned to the same compound. All spectral values were recorded at 2151 common discrete points approximately equally spaced in the wavelength range 0.35−2.5 μm from the ultraviolet to the far infrared. Compared to the number of UV−vis data set shown above, the sample size of 68 was quite small. Note that when representing a function using a sum of RBF kernels, the shape of the function to be predicted is required to be smooth. However, the reflectance spectra of the original data showed multiple sharp downward spikes, making it difficult to represent them as a weighted sum of RBFs. Therefore, a smooth spectral function was defined as the output variable using the conversion equation $A = \log_{10} 100/R$ from reflectance R to absorbance A.

The model structure and procedure for selecting hyperparameters for the cGAN and kernel regression were exactly the same as those for the analysis of the UV−vis spectra described above. Randomly selected instances of 80, 10, and 10% from the entire data set were used for model training, hyperparameter validation, and performance evaluation, respectively. In addition, data splitting was independently

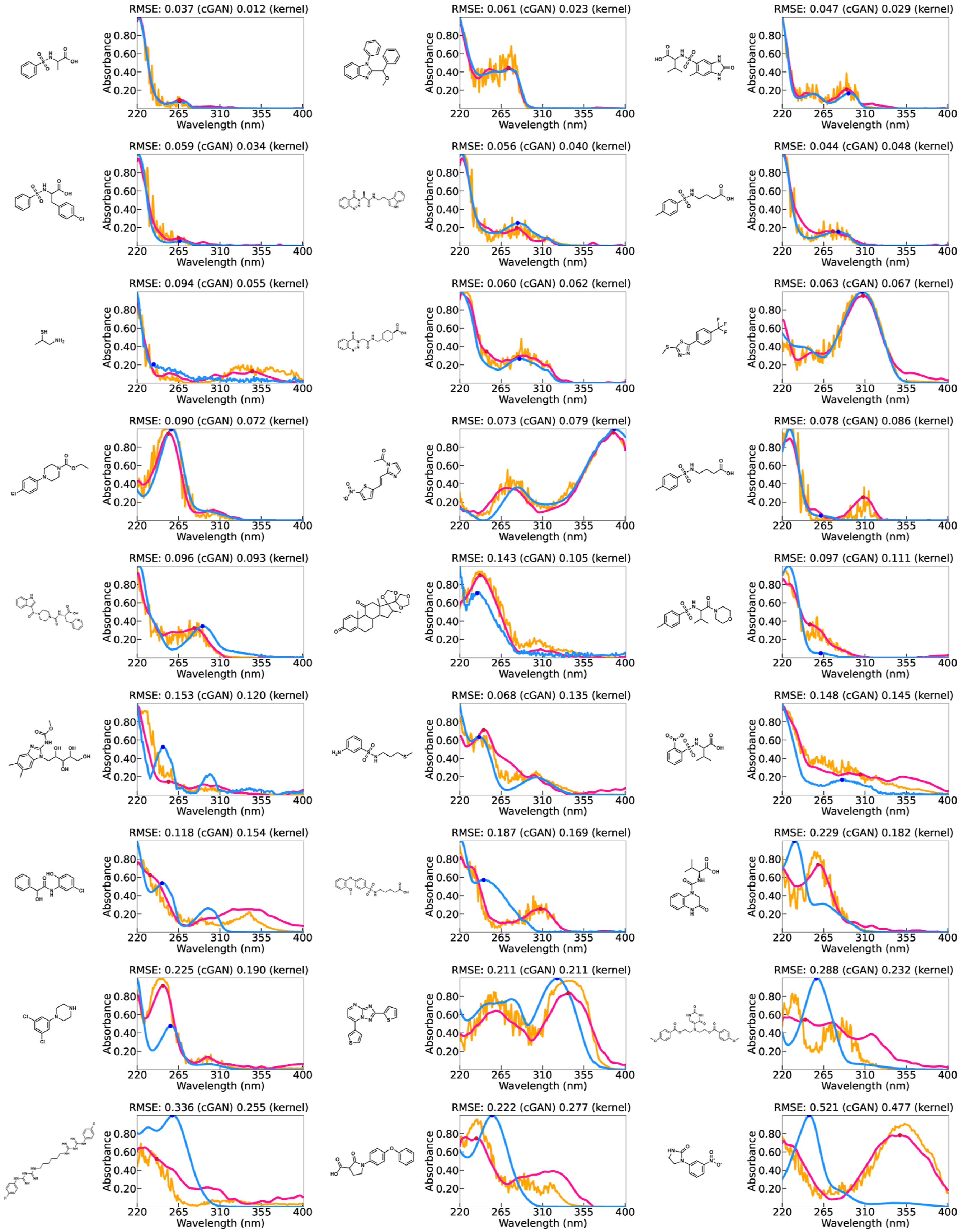


Figure 4. Prediction results of (a) the functional output kernel regression (pink) and (b) cGAN (orange) with observed UV−vis spectra (blue) in Dataset I.

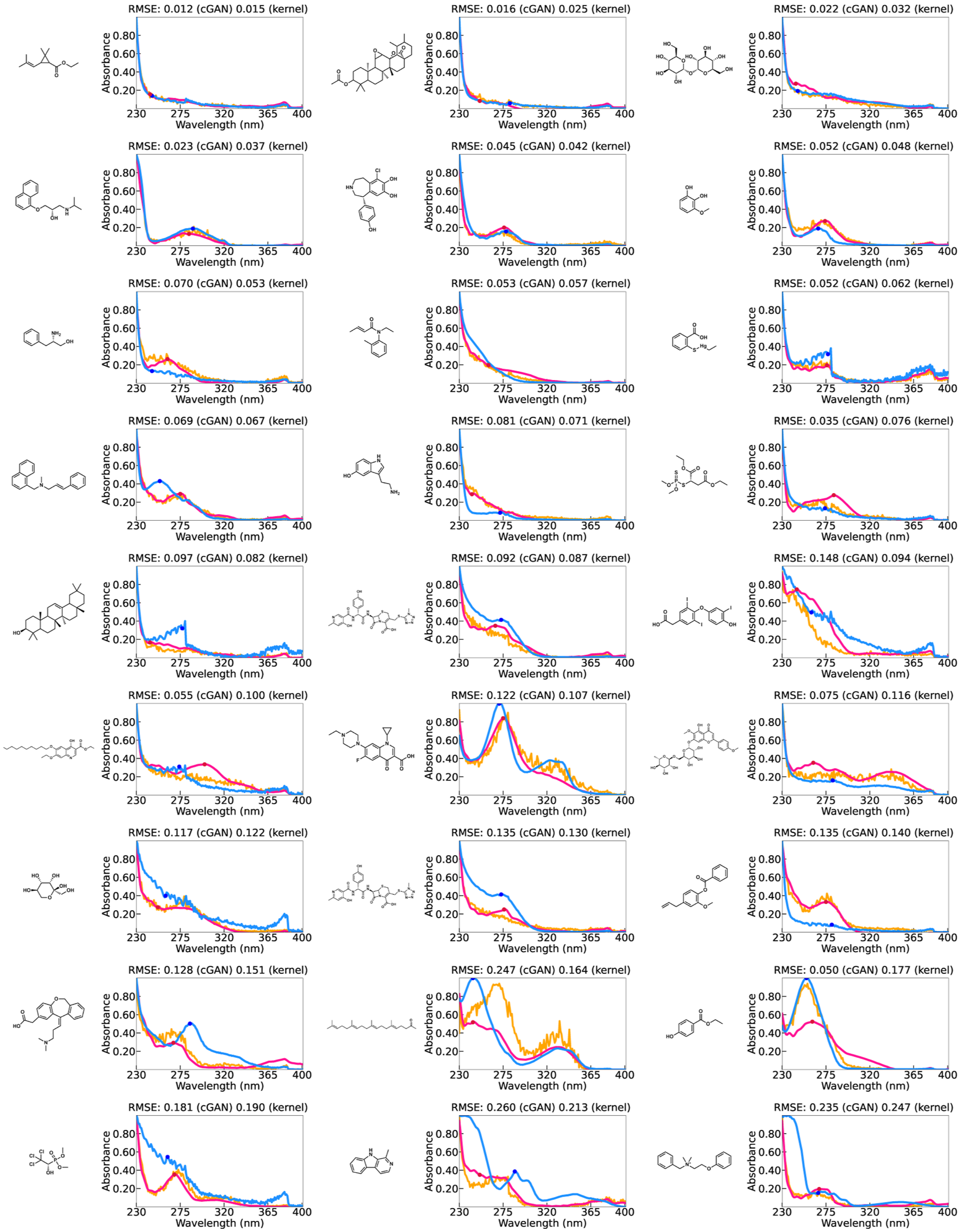


Figure 5. Prediction results of (a) the functional output kernel regression (pink) and (b) cGAN (orange) with observed UV−vis spectra (blue) in Dataset II.

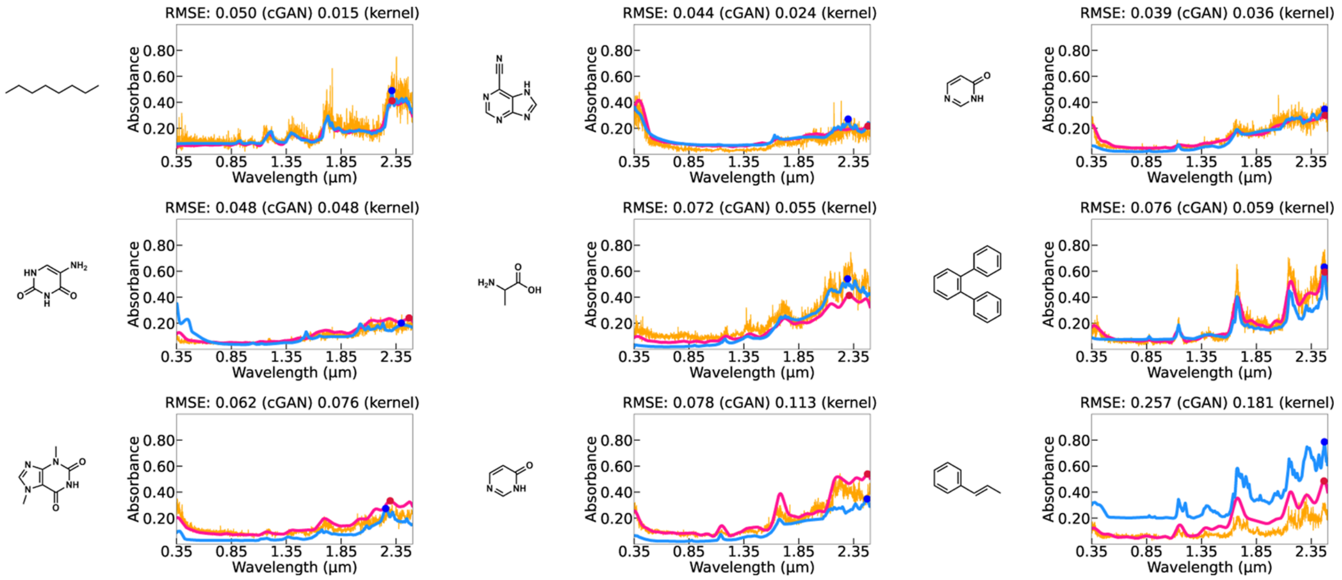


Figure 6. Prediction results of (a) the kernel regression (pink) and (b) cGAN (orange) for the optical absorbance spectra of nine test molecules with experimental profiles (blue) in the USGS spectral library.

performed three times to calculate the mean and variability of the performance measures. The numbers of stacked hidden layers for the NNs in the kernel regression, generator, and discriminator of cGAN were selected as 3, 3, and 2, respectively. The values of σ^2 and l were selected as 0.0005 and 1, respectively.

The prediction accuracy of cGAN and kernel regression for the test instances, which were trained on the optimized hyperparameters, are summarized in Table 1 (see the Supporting Information Note for details on the performance measures). Unlike the results of the UV−vis spectra, the kernel regression overwhelmingly outperformed cGAN in accuracy. For example, the R^2 values for cGAN and kernel regression were 0.493 ± 0.102 and 0.602 ± 0.200, respectively, and the MAEs were 0.058 ± 0.009 and 0.053 ± 0.009, respectively. It is likely that for the given small data set, the over-parameterization of the generator in cGAN caused degradation in the predictive performance such as getting stuck in a poor local optimum. However, the kernel regression reached a sufficiently high prediction accuracy despite being trained on only 54 samples. Figure 6 shows the predicted and true spectra for several cases (see Figure S7 for a more comprehensive visualization of the prediction results and Figure S8 for the result of fitting to the training data). As in the UV−vis cases, the peak positions and functional features of the observed spectra can generally be predicted. We also observed that, in many cases, the plateau and minor peaks could be properly captured. The kernel regression has a strong tolerance for a limited sample size, which will be further investigated later.

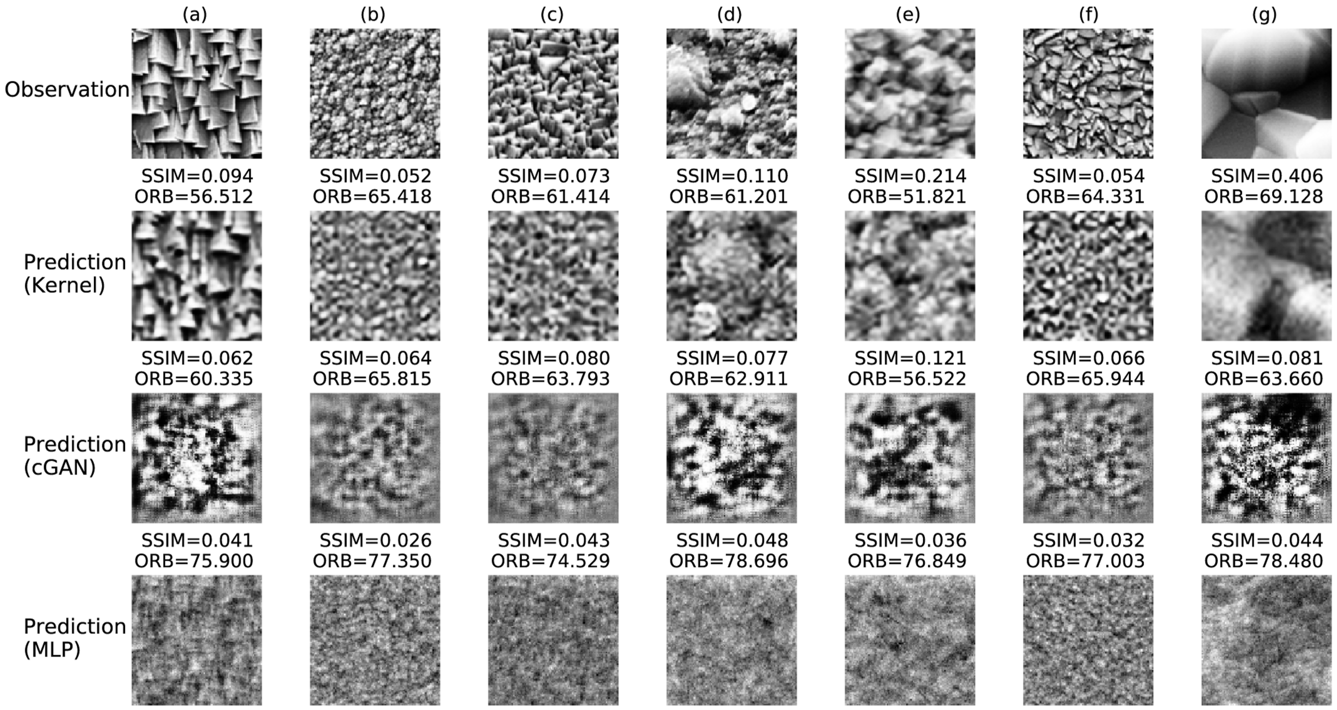
Microstructure Image Prediction. A thin-film material consisting of a chromium (Cr)-based metal plate coated with aluminum (Al) was analyzed.[34] Cr and Al metal plates were placed face-to-face, and a mixed gas of nitrogen (N) and argon (Ar) was sprayed onto the Al plate at high speed by magnetron sputtering, so that the ejected Al atoms were adsorbed onto the Cr plate. Under high-temperature conditions, the metal plates were contaminated with oxygen (O) from residual gas outgassing from the deposition equipment. The input $X \in \mathbb{R}^6$ to the model is a six-dimensional real vector representing the composition $Cr_{1-x}Al_xO_yN$ and the processing conditions.

1. Cr and Al content x
2. O content y
3. Temperature at which Al is adsorbed (T_d)
4. Pressure at which Al is adsorbed (P_d)
5. Average energy of the Ar ions when they are incident on the Al metal plate (E_i)
6. Ionization degree of the Ar gas (I_d)

Each element of the input vector X was normalized to have a mean of zero and a unit variance in the training data. The output variable Y is the SEM image of the microstructure with a resolution of 100 × 100.

In the functional output kernel regression, the centers of 32 × 32 RBF kernels ($d = 32 \times 32$) were placed at equally spaced positions in the two-dimensional image coordinate space. The modeling and training conditions were almost the same as those used to predict the spectral functions. The hyperparameter set to be explored was given by the network structure and the variance and length scale of the RBF kernel. The input variable was transferred to the embedding latent space via a fully connected layer. A repeating unit consisting of a transpose convolution layer and a leaky ReLU activation function was applied several times to this embedding vector. Here, the basic structure of the model was the same as that in the spectral function prediction, but a Gaussian noise was augmented to the input system, as in cGAN, to increase the diversity of representable image patterns as detailed in the Supporting Information Note. The cGAN modeling was also similar to that in the spectral function prediction. The input variable was transformed into the output variable via a series of fully connected layers to obtain a feature embedding, a repeated application of the transpose convolution layer and leaky ReLU to the embedding feature, and a fully connected layer to generate a 100 × 100 image Y (Supporting Information Note).

The data contained 123 SEM images with their compositional and processing conditions, of which 90, 5, and 5% were randomly assigned to the training, validation, and test sets,



Parameters	(a)	(b)	(c)	(d)	(e)	(f)	(g)
T_d [° C]	20	200	300	500	500	600	800
P_d [P_a]	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Al [at. %]	0	27	0	57	0	33	0
O [at. %]	0	10	3	0	0	10	3
E_I [eV]	1.43	7.23	6.40	41.4	202	6.44	2.80
I_d [a. u.]	0.14	0.50	0.22	1.0	0.20	0.50	0.12

Figure 7. Results of the microstructure prediction using the functional output kernel regression and cGAN and pixel-by-pixel prediction using a MLP with respect to seven different SEM images.

respectively. The data partitioning was independently repeated thrice. The generalization performance of the model was investigated by varying the number of layers from 1 to 4. The number of neurons in each layer was designed as described in the Supporting Information Note. The numbers of stacked hidden layers for the NNs in the kernel regression, generator, and discriminator of cGAN were selected as 4, 1, and 2, respectively. The values of σ^2 and l were selected as 0.0001 and 5, respectively.

Figure 7 shows the experimental and predicted SEM images for the functional output kernel regression and cGAN for seven randomly selected test conditions. To obtain a comprehensive view of the prediction performance, the Supporting Information Note provides the prediction results for 14 randomly selected test conditions and training results for 60 randomly selected conditions. Comparing with the experimental SEM images, the predicted images of the functional output kernel regression properly captured the difference in morphological features of microstructures, such as grain sizes, even though the amount of training data was as small. The cGAN model was unable to predict observed features of microstructures, possibly due to the limited amount of training data. We calculated the negative-transformed oriented FAST and rotated BRIEF (ORB)[55] and structural similarity (SSIM)[56] as measures of image similarity invariant to shifts in position, scale, and rotation (Supporting Information Note). Figure 8 summarizes the comparison of these two measures for the functional output kernel regression and cGAN with respect to the 21 test instances from the three independent trials. Clearly, the functional output kernel regression outperformed cGAN in both similarity measures.

As with the spectral prediction, a high learning potential of the function output kernel regression on small data was confirmed. However, the experimental results also indicated

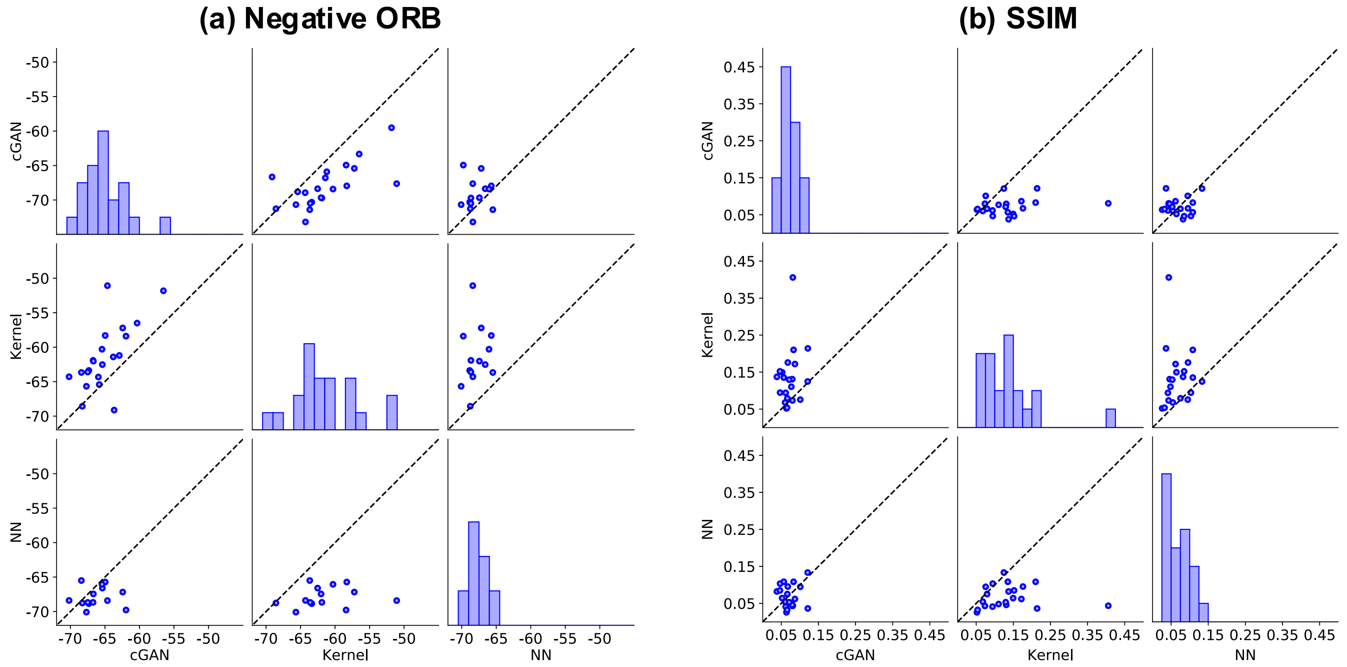


Figure 8. Scatter plot matrices showing the distribution of two similarity measures (the ORB with negative transformation and SSIM) of the predicted SEM images with respect to test data for cGAN, functional output kernel regression, and pixel-by-pixel independent prediction using conventional NNs.

the weakness of the trained model as an image generator. Although the predicted images successfully captured the morphological features of the microstructures, they were less clear than those of the experimental images. The reduction in sharpness was caused by the blurring effect of using kernel functions. For functional output regression, the parameter savings using a basis set of kernel functions leads to high tolerance for small data, but at the same time, it inevitably leads to a reduction in the quality of the generated images. However, unlike image generation in computer vision, the generation of high-quality images is not as important for prediction tasks in materials science. If sharpness or contrast needs to be increased, various techniques of image synthesis can be employed. Besides the blurring effect caused by using the kernel functions, the use of the squared loss of per-pixel image intensity would induce a reduction in the sharpness of generated images. To enhance sharpness and contrast, the gradient image similarity between the generated and real images could be added to the loss function, that is, sharpness loss regularization.[57,58] Furthermore, replacing the squared loss with *l*1-norm loss, as in pix2pix-GAN,[59] a well-known generative model for image-to-image translation, would reduce blurring.

Tolerance to Data Size in Functional Output Kernel Regression. We investigated the underlying mechanisms behind the learning success despite the exceedingly small amount of data, that is, 54 training samples for spectral prediction and 100 or more for microstructure prediction. The functional output regression has a learning mechanism that is common to multi-task learning. The tasks of predicting multiple function values are not independent but are related to each other. Joint learning of multiple related tasks is expected to be advantageous for the model to acquire common representations across tasks. In addition, data accumulation by the simultaneous use of data from multiple tasks is generally advantageous in suppressing overlearning of task-specific noises. These mechanisms may be responsible for the success in building the model with a small amount of data. To confirm this hypothesis, we conducted three numerical experiments using the optical absorption spectra in Dataset I.

(a) The maximum absorption wavelength $\lambda_{\max}$ and its intensity value were calculated from the spectral function predicted by the functional output kernel regression, and the accuracy of predicting $\lambda_{\max}$ and the maximum intensity were verified against their observed values.

(b) Using $\lambda_{\max}$ and its intensities as a data set extracted from the observed spectra in advance, we built a multi-layer perceptron (MLP) NN that directly predicts the extracted scalar values of $\lambda_{\max}$ or the maximum intensity from the fingerprinted chemical structure (the same as the one used in the cGAN and kernel regression).

(c) We trained *m* MLPs independently and separately to predict the spectral values of each of the *m* equally spaced grid points. The estimated $\lambda_{\max}$ and the maximum intensity values were calculated from the pointwise prediction of the spectral function.

In the peak detection from a spectral profile, $\lambda_{\max}$ was selected as the wavelength of the maximum intensity among the maxima that appeared 232 nm from the left end. In these three experiments, two different ratios of the training, validation, and test sets were set to 649:151:150 and 100:425:425, respectively. The data partitioning was repeated 10 times independently, and training and testing were performed. For the evaluation of the accuracy of the test set, cases in which the predicted peak position was within 10 nm of the observed peak position were determined as correct and cases in which the predicted peak intensity was within 0.1 of the observed intensity were determined as correct. As shown in Table 2, for both models obtained from the smaller and larger data sets, the prediction accuracy of the functional output kernel regression with the entire function trained simulta-

Table 2. Tolerance to Small Data in Predicting the Maximum Absorption Wavelength $\lambda_{\max}$ and Peak Intensity[a]

method	data split	position $\lambda_{\max}$ (%)	intensity of $\lambda_{\max}$ (%)
(a) kernel	649:151:150	46.3	43.9
(b) MLP for $\lambda_{\max}$		23.5	11.7
(c) MLP for each wavelength		45.9	45.5
(a) kernel	100:425:425	36.1	32.6
(b) MLP for $\lambda_{\max}$		22.4	11.1
(c) MLP for each wavelength		32.9	25.5

[a] Three cases were tested: (a) feature values obtained from the predicted spectral function of the functional output kernel regression, (b) the direct prediction of the two feature quantities using MLPs, and (c) feature values obtained from a set of MLPs with pointwise learning and prediction. The ratios of the training, validation, and test sets were set to 649:151:150 and 100:425:425, respectively.

neously (model (a)) overwhelmingly exceeded the performance of models (b) and (c). In particular, compared to model (c), which learned and predicted spectral values independently, model (a) showed little degradation in prediction performance when the amount of training data was reduced to 100.

We performed a similar test with the microstructure image prediction. The performance of the functional output kernel regression was compared to that of the pixelwise prediction by MLPs that learned the intensity of each pixel in the image independently. As illustrated in Figure 7, the pixelwise MLPs clearly failed to predict any test instances. Figure 8 shows that the similarity values of the predicted images from the pixelwise learning to the 21 test images were significantly lower than those of the functional output kernel regression in all cases.

This observation has implications for the methodological construction of machine learning-based material-property prediction. For example, in the prediction of a temperature-dependent property, it is often the case that the temperature range is limited to room temperature to define a target property to be predicted. However, such an approach makes the problem more difficult and reduces prediction accuracy. In fact, the pixel-by-pixel machine learning failed to predict the image intensity at all, but successfully predicted the entire images with a fairly high degree of accuracy. When functional data are available, direct prediction of the functional output variable can significantly improve prediction accuracy.

■ CONCLUSIONS

In the past few years, several machine learning techniques have been introduced in material science. In this context, machine learning techniques for predicting the physicochemical values of scalar quantities from input materials have matured. On the other hand, in this study, we focused on the regression problem where the output variable is in the form of a function. We described the potential problem setting in material science and presented two methodologies based on deep generative models and statistical functional data analysis. Because machine learning research based on this perspective is still in its infancy, there must be other promising methods besides the one proposed here. As a starting point, we demonstrated the potential of functional output regression in two cases: the prediction of the light absorption spectral function of a molecular system and the prediction of the microstructure from experimental condition parameters.

Of particular interest is the mechanism of the high tolerance of the functional output regression to the limited amount of training data. In this study, the amount of data was much smaller than in general situations: 54 samples in the prediction of absorption spectra and 109 samples for microstructure image prediction. Despite the extremely small sample sizes, we successfully obtained models with sufficiently high prediction accuracies. In general, machine learning prediction of physical or chemical properties is often performed by transforming the functional data into a few features and then predicting the scalar output variables, instead of directly predicting the functional data. However, the experimental results suggest that the direct prediction of functional output variables may involve a learning mechanism that favors the acquisition of higher generalization performance than the prediction of scalar variables. Intuitively, the higher dimensionality of the output variable is likely to lead to overlearning and a decrease in the generalization performance of the trained model. However, because the tasks of simultaneously predicting multiple function values are not independent but strongly related to each other, the simultaneous use of data from multiple tasks can suppress task-specific noise due to data expansion. In addition, in multi-task modeling, the complexity of the model does not increase significantly as usually only one- or two-dimensional input variables t are added. In the trade-off between increasing model complexity and data expansion, the advantages of the latter tend to outweigh those of the former.

There are many other potential applications in material research where the prediction of functional outputs is required. Most material properties are given as functions of temperature or frequency. The dielectric properties are defined as functions of temperature and frequency. Polymeric properties such as the specific volume, coefficient of linear expansion, bulk modulus, specific heat, thermal conductivity, and χ parameters are also determined in a temperature-dependent manner. From these temperature-dependent curves, important properties such as the glass transition temperature, crystal melting point, and crystallization temperature can be calculated. The dependence of properties on processing conditional parameters is another typical application of functional output regression. Furthermore, in the 2D and 3D imaging of material structures, the output variable is given as a multidimensional function in the image coordinate space. Due to the lack of data availability, we have presented only a few limited applications, but there are many problem settings for functional output regression that remain unexplored in material research. In principle, the proposed methods are designed to handle arbitrary high-dimensional functional data. We hope that the distributed Python code can be utilized to discover more problem settings.

■ DATA AND SOFTWARE AVAILABILITY

The Python codes for the functional output kernel regression and pretrained models are available on GitHub.[41] Optical absorption spectra are available from Urbina et al.[31] for Dataset I and Dataset II and Kokaly et al.[54] for the USGS library. SEM images of the microstructures with compositional and processing data are available from GitHub https://github.com/lbanko/generative-structure-zone-diagrams provided by Banko et al.[34]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00626.

> Modeling details for cGAN and the functional output kernel regression, procedure of hyperparameter search, results of predicting the optical absorption spectra in the UV–vis region and in the USGS spectral library, and the microstructure image prediction (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ryo Yoshida – *Department of Statistical Science, The Graduate University for Advanced Studies, Tachikawa 190-8562, Japan; Research Organization of Information and Systems, The Institute of Statistical Mathematics, Tachikawa 190-8562, Japan; Research and Service Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba 305-0047, Japan;* orcid.org/0000-0001-8092-0162; Email: yoshidar@ism.ac.jp

Authors

Megumi Iwayama – *Department of Statistical Science, The Graduate University for Advanced Studies, Tachikawa 190-8562, Japan; Production Management Headquarters, Process Technology Division, Daicel Corporation, Himeji 671-1283, Japan;* orcid.org/0000-0003-3032-9600

Stephen Wu – *Department of Statistical Science, The Graduate University for Advanced Studies, Tachikawa 190-8562, Japan; Research Organization of Information and Systems, The Institute of Statistical Mathematics, Tachikawa 190-8562, Japan;* orcid.org/0000-0002-7847-8106

Chang Liu – *Research Organization of Information and Systems, The Institute of Statistical Mathematics, Tachikawa 190-8562, Japan;* orcid.org/0000-0002-9511-4283

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c00626

Author Contributions

R.Y. and M.I. devised the project, main conceptual ideas, and proof outline. M.I. implemented the machine-learning algorithms and conducted the experiment with the support of C.L. M.I. and R.Y. wrote the manuscript with the support of S.W. R.Y. supervised this project.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by JST CREST grant number JPMJCR19I3, MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" hp210264, MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas 19H05820, Grant-in-Aid for Scientific Research (A) 19H01132, and Grant-in-Aid for Young Scientists 21K14675 from the Japan Society for the Promotion of Science.

■ REFERENCES

(1) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M. S.; Kim, W.; Huang, S. I.; Hong, M.; Baldo, R. P.; Adams, A.; Aspuru-Guzik, A. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120–1127.

(2) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **2019**, *5*, 66.

(3) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

(4) Carrete, J.; Li, W.; Mingo, N.; Wang, S.; Curtarolo, S. Finding Unprecedentedly Low-Thermal-Conductivity Half-Heusler Semiconductors via High-Throughput Materials Modeling. *Phys. Rev. X* **2014**, *4*, 011019.

(5) Liu, C.; Fujita, E.; Katsura, Y.; Inada, Y.; Ishikawa, A.; Tamura, R.; Kimura, K.; Yoshida, R. Machine learning to predict quasicrystals from chemical compositions. *Adv. Mater.* **2021**, *33*, 2102507.

(6) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **2017**, *95*, 144110.

(7) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(8) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679.

(9) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.

(10) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(11) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217–241.

(12) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(13) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(14) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Model.* **1987**, *27*, 82–85.

(15) Choudhary, K.; DeCost, B.; Tavazza, F. Machine learning with force-field inspired descriptors for materials: fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2018**, *2*, 083801.

(16) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **2019**, *5*, 1717–1730.

(17) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.

(18) Geary, R. C. The contiguity ratio and statistical mapping. *Inc. Statistician* **1954**, *5*, 115–146.

(19) Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **1950**, *37*, 17–23.

(20) Moreau, G.; Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *New J. Chem.* **1980**, *4*, 359–360.

(21) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **2009**, *130*, 104504.

(22) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(23) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P.Convolutional networks on

graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems 28*; NIPS: Montreal, 2015; pp 2224−2232.

(24) Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; MüllerSchnet, K.-R.A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems 30*; NIPS: California, 2017; pp 992−1002.

(25) Li, X.; Zhang, Y.; Zhao, H.; Burkhart, C.; Brinson, L. C.; Chen, W. A transfer learning approach for microstructure reconstruction and structure-property predictions. *Sci. Rep.* **2018**, *8*, 13461.

(26) Cang, R.; Xu, Y.; Chen, S.; Liu, Y.; Jiao, Y.; Yi Ren, M. Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design. *J. Mech. Des.* **2017**, *139*, 071404.

(27) Hiraoka, Y.; Nakamura, T.; Hirata, A.; Escolar, E. G.; Matsue, K.; Nishiura, Y. Hierarchical structures of amorphous solids characterized by persistent homology. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 7035−7040.

(28) Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Tu, Q.; Ceder, G. Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra. *Chem. Mater.* **2021**, *33*, 4204−4215.

(29) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.

(30) Guo, Z.; Wu, S.; Ohno, M.; Yoshida, R. Bayesian Algorithm for Retrosynthesis. *J. Chem. Inf. Model.* **2020**, *60*, 4474−4486.

(31) Urbina, F.; Batra, K.; Luebke, K. J.; White, J. D.; Matsiev, D.; Olson, L. L.; Malerich, J. P.; Hupcey, M. A.; Madrid, P. B.; Ekins, S. UV-adVISor: Attention-Based Recurrent Neural Networks to Predict UV−Vis Spectra. *Anal. Chem.* **2021**, *93*, 16076−16085.

(32) Mirza, M.; Osindero, S.Conditional generative adversarial nets. **2014**, arXiv preprint arXiv:1411.1784.

(33) Sutskever, I.; Vinyals, O.; Le, Q. V.Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems 27*; NIPS: Montreal, 2014; pp 3104−3112.

(34) Banko, L.; Lysogorskiy, Y.; Grochla, D.; Naujoks, D.; Drautz, R.; Ludwig, A. Predicting structure zone diagrams for thin film synthesis by generative machine learning. *Commun. Mater.* **2020**, *1*, 15.

(35) Yang, Z.; Li, X.; Catherine Brinson, L.; Choudhary, A. N.; Chen, W.; Agrawal, A. Microstructural materials design via deep adversarial learning methodology. *J. Mech. Des.* **2018**, *140*, 111416.

(36) Li, X.; Yang, Z.; Brinson, L. C.; Choudhary, A.; Agrawal, A.; Chen, W.A deep adversarial learning methodology for designing microstructural material systems. *44th Design Automation Conference*, Quebec, 2018; Vol. *2B*.

(37) Ramsay, J. O.; Silverman, B. W.*Applied Functional Data Analysis: Methods and Case Studies*; Springer: New York, 2002; Vol. 77.

(38) Mamede, R.; Pereira, F.; Aires-de-Sousa, J. Machine learning prediction of UV-Vis spectra features of organic compounds related to photoreactive potential. *Sci. Rep.* **2021**, *11*, 23720−11.

(39) Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41−75.

(40) Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30−43.

(41) Iwayama, M.; Chang, L.; Stephen, W.; Yoshida, R.XenonPy platform. https://github.com/yoshida-lab/XenonPy/blob/master/samples/kernel_neural_network.ipynb (accessed May 07, 2022).

(42) Baldo, M. A.; O'Brien, D. F.; You, Y.; Shoustikov, A.; Sibley, S.; Thompson, M. E.; Forrest, S. R. Highly efficient phosphorescent emission from organic electroluminescent devices. *Nature* **1998**, *395*, 151−154.

(43) Mazzio, K. A.; Luscombe, C. K. The future of organic photovoltaics. *Chem. Soc. Rev.* **2015**, *44*, 78−90.

(44) Shaath, N. A. Ultraviolet filters. *Photochem. Photobiol. Sci.* **2010**, *9*, 464−469.

(45) Marques, M.; Rubio, A.; Gross, E. K.; Burke, K.; Nogueira, F.; Ullrich, C. A.*Time-Dependent Density Functional Theory*; Springer Science & Business Media: Berlin, 2006; Vol. *706*.

(46) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(47) Chen, L.; Kim, C.; Batra, R.; Lightstone, J. P.; Wu, C.; Li, Z.; Deshmukh, A. A.; Wang, Y.; Tran, H. D.; Vashishta, P.; Sotzing, G. A.; Cao, Y.; Ramprasad, R. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput. Mater.* **2020**, *6*, 61.

(48) Koster, A.; Ziese, U.; Verkleij, A.; Janssen, A.; de Jong, K. Three-dimensional transmission electron microscopy: a novel imaging and characterization technique with nanometer scale resolution for materials science. *J. Phys. Chem. B* **2000**, *104*, 9368−9370.

(49) Garcea, S.; Wang, Y.; Withers, P. X-ray computed tomography of polymer composites. *Compos. Sci. Technol.* **2018**, *156*, 305−319.

(50) Takeuchi, A.; Suzuki, Y. Recent progress in synchrotron radiation 3D-4D nano-imaging based on X-ray full-field microscopy. *Microscopy* **2020**, *69*, 259−279.

(51) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436−444.

(52) Ioffe, S.; Szegedy, C.Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. **2015**, arXiv preprint arXiv:1502.03167.

(53) Maas, A. L.; Hannun, A. Y.; Ng, A. Y.Rectifier Nonlinearities Improve Neural Network Acoustic Models. *ICML*, 2013; p 30.

(54) Kokaly, R.; Clark, R.; Swayze, G.; Livo, K.; Hoefen, T.; Pearson, N.; Wise, R.; Benzel, W.; Lowers, H.; Driscoll, R.; Klein, A.*USGS Spectral Library Version 7*; U.S. Geological Survey data release, 2017.

(55) Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G.ORB: an efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision*, Barcelona, 2011; pp 2564−2571.

(56) Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600−612.

(57) Yu, B.; Zhou, L.; Wang, L.; Shi, Y.; Fripp, J.; Bourgeat, P. Ea-GANs: Edge-Aware Generative Adversarial Networks for Cross-Modality MR Image Synthesis. *IEEE Trans. Med. Imag.* **2019**, *38*, 1750−1762.

(58) Butte, S.; Wang, H.; Xian, M.; Vakanski, A.Sharp-GAN: Sharpness Loss Regularized GAN for Histopathology Image Synthesis. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022; pp 1−5.

(59) Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A. A.Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017; pp 1125−1134..