Machine Learning Enabled Capacitance Prediction for Carbon-Based Supercapacitors

Shan Zhu\textsuperscript{a}, Jiajun Li\textsuperscript{a}, Liying Ma\textsuperscript{a}, Chunlian He\textsuperscript{a,b,c}, Enzuo Liu\textsuperscript{a,b}, Fang He\textsuperscript{a}, Chunsheng Shi\textsuperscript{a}, and Naiqin Zhao\textsuperscript{a,b,c}

\textsuperscript{a} School of Materials Science and Engineering and Tianjin Key Laboratory of Composites and Functional Materials, Tianjin University, Tianjin 300350, China

\textsuperscript{b} Collaborative Innovation Center of Chemical Science and Engineering, Tianjin 300350, China

\textsuperscript{c} Key Laboratory of Advanced Ceramics and Machining Technology, Ministry of Education, Tianjin, 300350, China

*Email: shanzhu@tju.edu.cn, nqzhao@tju.edu.cn

Abstract

Carbon is the most widely used electrode for the supercapacitors. To predict the capacitance of carbon-based supercapacitors, this work applies three machine learning (ML) methods, including linear regression, Lasso and artificial neural network. For training the ML process, we extracted data from hundreds of published papers. Moreover, five variables were selected to figure out their impact on capacitance, including specific surface area, calculated pore size, \( I_D/I_G \) ratio, N-doping level and voltage window. By evaluated with the real data, all of three methods achieve acceptable prediction results, and ANN exhibits the best performance. More importantly, this work shows the potential of ML in material science and advanced applications.

**Keywords:** supercapacitors, carbon, machine learning, artificial neural network, performance prediction
Supercapacitors enjoy many merits as a typical energy storage application, such as high power density and long-cycle life.\textsuperscript{[1]} The key component of a supercapacitor is the electrode. The most widely used material for the electrode is carbon, due to their excellent physical and chemical properties.\textsuperscript{[1,2]} To enhance the performance of supercapacitors, researchers are eager to increase the specific capacitance of carbon-based electrodes. For examples, they improve the capacitance by increasing specific surface area, controlling pore structure and introducing surface functional groups.\textsuperscript{[1,2]} However, there is currently no accepted theory to systematically explain how to increase capacity; the modification work on the capacitor electrode only relies on several empirical formulas or rough theoretical models. As the result, the reported data for capacitance are very volatile, which vary from 10 to 800 F/g and is depended on the materials and test conditions. Therefore, it is urgent to obtain a sophisticated evaluation system for predicting the capacitive properties of carbon-based supercapacitors.

To solve this problem, the feasible method is to draw insights from large amount of data. Traditionally, experimental data are often collected, followed by extrapolation predictions. However, this method does not seem to support a large amount of data for performance prediction with multi-factor control. Based on a large set of data, it is possible to draw some fundamentals, which are often neglected in small data sets. Among various data science technologies, machine learning (ML) is an effective means for creating a predictive relationship between input and output variables.\textsuperscript{[3]} Considering the complex variables in carbon-based supercapacitors, the prediction on multstructure/single-property relationships is still challenging. Thus, this is an ideal problem for ML methods, which take advantage of available datasets and predict property relationships with high efficiency and accuracy.

Considering the data sources, here comes another issue. At present, in the field of ML-assisted material design, the data come from several large scale project, such as Materials Genome Initiative. In such databases, the properties of materials including electrical, optical, mechanical properties can be obtained. Yet, in the carbon-based supercapacitors, the main materials are well-determined: carbon. As the result, the factors that really affect the performance turn to the materials' microstructures and surface functional groups. Even more, the testing details of practical application also have their own influences. However, there is no database to describe the abovementioned data. Otherwise, some researchers use the calculated data for ML. For example, Subramanian et al applied P2D model to calculate a lot of data for training their ML model;\textsuperscript{[4]} Farsi and Gobal used Lin model to creat data for their algorithm to predict supercapacitor performance.\textsuperscript{[5]} At the first glance, this method can provide enough data for ML. However, here is a paradox: researchers create the data by specific rules, then they want to find the law behind. It turns to a self justification. Therefore, we believe that the most plausible way is to directly extract data from the existing papers, and establish the correspondence between multivariate information and performance. Ghadbeigi et al. provided performance indicators for Li-ion battery electrode materials, whose data are
extracted from ~200 articles. Raccuglia et al. applied a similar human-data-retrieval technique to compile ~4,000 reaction conditions for training machine-learned syntheses of vanadium selenite crystals.

Herein, we introduce ML to predict the capacitance performance of carbon materials. Although there are countless different algorithms associated with ML, this paper is primarily concerned with linear regression, Lasso and artificial neural network (ANN). Figure 1 illustrates the process for capacitance prediction based on ML in this work.

We have collected more than 1000 published papers about carbon-based supercapacitors (Figure 2). Then, over 10,000 data points were extracted, including the physical and chemical features of carbon material (i.e. specific surface area, pore volume, micropore volume, the ratio of I_D/I_G, doping elements) and the test system (i.e. electrolyte, test voltage window, and corresponding specific capacity). Moreover, the data were divided into two parts (one for training and one for verification). When the data are used for model training, the cross-validation is the best way to evaluate the predictive models.

Three kinds of ML models were applied in this work, which are linear regression, Lasso and ANN. In statistics, linear regression is a regression analysis of the relationship between one or more independent variables and dependent variables using a least-squares function called a linear regression equation. By comparison, Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical models. Lasso regression is characterized by variable selection and regularization while fitting a generalized linear model. Therefore, regardless of whether the dependent/response variable is continuous or binary or multivariate, it can be modeled and then predicted using Lasso regression. ANN is an effective estimation technique to approximate the parameters for complex and their non-linear behaviors. Since ANNs can discover relationships between inputs and outputs of a system without a detailed understanding of the mechanisms involved, they can be effectively applied to estimate models for systems where such relationships are not clearly understood.

Since there are many kinds of variables in our database, the next step is to choose the suitable ones. This part in ANN is called feature selection, and the number of features is equal to the number of nodes in the input layer of the network. The independent variables or the feature
should be physically meaningful and have a bearing on the property, meaning that they can explain the dependent variable well. There are multiple ways to select the right set of variables for the model. First among them would be the scientific understanding and domain knowledge.

Figure 2. The relation between capacitance and (a) specific surface area, (b) calculated pore size, (c) pore volume, (d) N-doping amount, (e) the ratio of \( \frac{I_D}{I_G} \) and (f) voltage window.

As the electrodes of supercapacitors, the most important feature of carbon materials is their specific surface area. The influence of surface area was initially introduced by Helmholtz and later refined by Stern and Geary, which can be described according to \(^{1,2}\)
\[ C = \frac{\varepsilon \varepsilon_0 A}{d} \]

where \( C \) is the capacitance, \( \varepsilon \) is the electrolyte dielectric constant, \( \varepsilon_0 \) is the vacuum permittivity, \( A \) the surface area accessible to ions, and \( d \) the distance between the center of the ion and the carbon surface. To increase the charge stored, it is necessary to increase the carbon surface area. Unfortunately, there is no simple linear relationship between the surface area and the capacitance (Figure 2a).

The accessibility of electrolyte ions is also affected by the pore size distribution of carbon. For example, there is a greatly enhanced capacity in non-aqueous electrolytes when pore sizes are below 1 nm.\[^{10}\] However, there are many kinds of methods for analyzing the pore size, such as different detection molecules and different mathematical models for fitting the data. All these differences will undoubtedly bring incompatibility among various researches. Therefore, we would like to use mathematical calculations to obtain a unified pore by pore volume (Figure 2b and Figure 2c). To this regard, the pore size was calculated by following:\[^{11}\]

\[
Pore\ size(\text{calculated}) = \frac{4 \times \text{Pore volume}}{\text{Specific surface area}}
\]

Besides the microstrcutures, the hybridization pattern and crystallization degree of carbon are inheret characteristics (Figure 2e), which can be investigated by Raman spectroscopy. In Raman results, the D-bands at around 1360 cm\(^{-1}\) represent the existence of the defects and the amorphous carbon; the G-bands reflecting the sp\(^2\)-hybridized carbon locate at around 1570 cm\(^{-1}\); and the intensity ratio between the D-band and G-band (\( I_D/I_G \)) can characterize the defect condition and the crystallization degree of carbon materials.\[^{12-14}\] In general, the integral intensity ratio of D peak to G peak (\( I_D/I_G \)) is used to evaluate the crystallization degree of carbon materials. According to recent studies, the crystallinity of materials has an effect on the electrical conductivity, and the main cause of the crystallinity is the heating temperature, the lattice arrangement of the impression carbon.

Besides of the optimized structures, the suitable components also play a significant role in carbonaceous materials for electrochemical energy storage applications. Among various elements, nitrogen is the most common heteratoms that introduced in carbon matrix.\[^{13,14}\] the nitrogen-content functional groups greatly increase the capacitive properties of carbon-based materials by introducing additional pseudocapacitance and quantum capacitance.

In addition to the abovementioned variables, the material testing system is also important, especially the electrolyte.\[^{1,2}\] The main consequence introduced by various electrolytes is the voltage window. There is an equation:

\[ E = CV^2 \]
where $E$ is the energy density of the supercapacitor, $C$ is the capacitance and $V$ is the voltage window. To this regard, voltage window is an important indicator for the overall performance of the device.

After selected the variables, these ML methods were conducted to analyze the database. It is worth pointing out that the ANN is comprised by three-layer: an input layer with a neuron per input, a hidden layer, and a single neuron in the output layer (Figure 3). In this experiment, the network was trained with the training data, and the trained network was then evaluated using the test data. Each experiment was comprised of a large number of runs, for various network configurations.

![Figure 3. Illustration of ANN.](image)

Since these three methods were trained by the data, it is reasonable to evaluate their performance. For this purpose, we selected data from several materials other than the data set and substituted it into the corresponding model for testing. The feature parameters of these samples are very different from each other (Table S1). For example, the specific area varies from 20 to 2856 m$^2$ g$^{-1}$. Also, the test systems, the heteroatom-doping contents and the pore size distributions vary in a wide range. To this regard, the feasibility of these ML method can be tested in a comprehensive way. The prediction results of three methods are in the same order of magnitude with the real capacitance (Figure 4). Among them, the numerical values of linear regression and Lasso's predictions are close to each other. This phenomenon is caused by that the regulation effect of Lasso is not obvious with small number of variables. In the aspect of accuracy, ANN show the best performance, which is obviously closer to the real value. Especially in the sample of No.5, the error of ANN is only 3.3 %, while the errors of linear regression and Lasso are almost 90 %. The superiority of ANN derived from its ability to recognize patterns in a series of input and output data without any prior assumptions about their nature and interrelations. As the result, ANN can minimize the error between the calculated output and the experimental known targets during and achieve a better prediction
performance. The deviation of the experiment was mainly due to the small number of data points. As the number of samples increases, the accuracy of predictions will increase.

**Figure 4.** The comparison between real capacitances and the predictions of three models.

In conclusion, this work utilized three ML-based models (linear regression, Lasso and ANN) to predict the capacitance performance of carbon materials in supercapacitors. The data were extracted more than 1000 published papers. Moreover, five variables, including specific surface area, calculated pore size, I_D/I_G ratio, N-doping level and voltage window, serve as the features in the ML processes. By tested with the real data, all of three methods achieve acceptable prediction results, and ANN exhibits the better accuracy than others. More importantly, ML shows its great potential for assisting researchers in material science and application design.

**Acknowledgement**

This work was supported by the National Natural Science Foundation of China (Grant Nos. 51472177, 51772206 and 11474216), and the Science and Technology Support Program of Tianjin (No. 16ZXCLGX00110, and 16ZXCLGX00070). Shan Zhu was supported by China Scholarship Council.

**Reference**