Computational Materials Science 202 (2022) 110967

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

A 3D orthogonal vision-based band-gap prediction using deep learning: A proof of concept

Ricardo Espinosa [a], Hiram Ponce [b,*], Josue Ortiz-Medina [a,*]

[a] Universidad Panamericana, Facultad de Ingeniería, Josemaría Escrivá de Balaguer 101, Aguascalientes, Aguascalientes, 20296, Mexico
[b] Universidad Panamericana, Facultad de Ingeniería, Augusto Rodin 498, Ciudad de México, 03920, Mexico

ARTICLE INFO

Dataset link: https://1drv.ms/u/s!AgoMuI4-UMpbu1n-3vPvyFGpZv3y?e=jotM9W

Keywords:
Band-gap prediction
Deep learning
Vision-based regression

ABSTRACT

In this work, a vision-based system for the electronic band-gap prediction of organic molecules is proposed using a multichannel 2D convolutional neural network (CNN) and a 3D CNN, applied to the recognition and classification of 2D projected images from 3D molecular structure models. The generated images are input into the CNN for an estimation of the energy gap, associated with the molecular structure. The public data set used in this research was the Organic Materials Database (OMDB-GAP1). A data transformation from the descriptive information contained in the data set to three 2D orthogonal images of molecules was done. The training set is composed of 30,000 images, whereas the testing set was composed of 7500 images, from 12,500 different molecules. The multichannel 2D CNN architecture was optimized via Bayesian optimization. Experimental results showed that the proposed CNN model obtained an acceptable mean absolute error of 0.6780 eV and root mean-squared error of 0.7673 eV, in contrast to two machine learning methods reported in the literature used for band-gap prediction based on conventional density function theory (DFT) methods. These results demonstrate the feasibility of CNN models to materials science routines using orthogonal images projections of molecules.

1. Introduction

The conventional way for theoretically simulate and analyze molecular models, in the context of materials science, has been mainly done by means of computational algorithms based on physical descriptions of atomic or molecular interactions [1]. These models describe their components in terms of quantum-physical frameworks, when used for analysis of subatomic (i.e., electronic) interactions, or in terms of parametrized factors when dealing with larger and complex systems, such as polymers or proteins [2,3]. Usually, these computational calculations are significantly expensive, given the enormous amount of mathematical operations needed to evaluate and solve such interactions between electrons or charged atoms [4]. These calculations are based on computational adaptations of the Schrödinger wavefunction equation, which can be considered as the base for the understanding of the quantum scale phenomena [5].

In recent years, with the advancements of machine learning (ML) and other computational developments related with artificial intelligence (AI), new alternatives have appeared in the materials research horizon for optimizing the study of novel materials [6–8]. These tools have already been proposed for understanding their fundamental properties, or for screening new structures with high potential in fields such as solar cells [9,10] and catalysis [11,12], graphene composites [13], as well as for optimizing characterization processes for materials where large amounts of data need to be analyzed and processed [14,15]. Among these ML methods, deep-learning networks, such as convolutional neural networks (CNNs), which is a subclass of artificial neural networks (ANNs), have been applied to the analysis of structures and their related calculated properties as a training set, with encouraging results [16,17]. For instance, Fernandez et al. [18] reported the use of different ML methods, including ANNs, linear multiple regression and support vector machines (among others) for correlating geometrical features of graphene nanoflakes with electronic properties such as electron affinity, electronic band-gap, Fermi level and ionization potential, resulting in a correlation (R^2) during tests about 0.9. Other relevant works have demonstrated that a correlation among structure features can be represented by proper descriptors, for a further use in ML algorithms. Yang et al. [11] and Sosso and Bernasconi [19], for instance, reported approaches for using simplified descriptors based on conventional but computationally expensive methods for materials research (such as density functional theory or DFT), which allow their fast and accurate study through ML.

* Corresponding authors.
E-mail addresses: respinosa@up.edu.mx (R. Espinosa), hponce@up.edu.mx (H. Ponce), jortizm@up.edu.mx (J. Ortiz-Medina).

https://doi.org/10.1016/j.commatsci.2021.110967
Received 26 August 2021; Received in revised form 7 October 2021; Accepted 10 October 2021
Available online 21 October 2021

Nowadays, one of the most developed areas of ML methods is related with the fast and efficient processing of images, specially by means of advanced deep-learning algorithms. New vision systems are being strongly based now on intelligent networks capable of identifying and tracking, in real-time, many kinds of objects and patterns [20–22]. Nevertheless, as mentioned, the tendency for using ML approaches to advanced materials research has been somewhat limited to the creation of descriptors from specific calculated properties, and training the networks on the basis of indirect structural information instead of using the molecule structure by itself. Few studies have attempted to explore this way. For instance, a recent report proposes to use a molecular representation from a concept called persistent homology, which is related with structure topology, for identification of chemical species capable of interact with CO_2 molecules [23]. In addition, few years ago, Fernandez et al. [24] and Asilar et al. [25] described approaches that, to the best of our knowledge, are among the closest attempts to utilize the power of ML methods for identifying the toxicity of specific chemical species based only in the 2D graphical representation of their molecular structures. An even more recent report makes a proposal of using the 3D coordinates of a molecular model for the determination of key chemical characteristics, by means of a specialized graph neural network architecture [26]. These reports show the potential of applying ML methods for identification and classification of molecular structures from their graphical representation perspective, which according with our survey, is still a relatively unexplored area in materials science research.

In this paper, we are proposing an approach that uses the projected images of 3D spatial models of molecules as inputs for a CNN, that would correlate the structural features (i.e., atom types and position) with the electronic band-gap as calculated by conventional methods and reported in the Organic Materials Database (OMDB-GAP1) [27]. We consider that the electronic band-gap prediction is a key feature to address, given the importance it has for the discovery and development of molecular structures susceptible to be used in high-performance electronics and energy harvesting applications. To the best of our knowledge, this is the first time in which an image-based approach is applied to band-gap prediction using ML techniques. This research would contribute to an easy and direct implementation of ML methods for new materials screening, without the intensive work that the development of complex descriptors associated with specific materials properties that is often used.

2. Methodology

This study is based on the workflow described in [28] for deep-learning. We adapted this methodology to our particular application that consists of the following steps: (i) data collection using (OMDB-GAP1) data set, (ii) data transformation changing the input data from XYZ file to three orthogonal images, (iii) deep learning applying multichannel 2D CNN and 3D CNN approaches, and (iv) band-gap prediction. A condensed scheme of our workflow is shown in Fig. 1.

2.1. Data set description

This work makes use of the Organic Materials Database (OMDB-GAP1) [27], a XYZ extension file which contains information of about 12,500 organic compounds. Each organic crystal is built by x, y, z coordinates with its corresponding electronic band-gap calculated by DFT methods. The data set we used was built with 65 elements, with the heaviest element being uranium, and spans over 69 space groups and atom counts from 7 to 208 atoms, with an average of 82 atoms, per molecule.

2.2. Data collection and data transformation

One of the most challenging phases in the workflow for ML problems is the data collection task. In principle, deep-learning techniques require a large amount of data to be trained and tested correctly, including additional intensive tasks as data labeling, data augmentation, pre-processing and transformation phases [29]. By using the OMDB-GAP1 as data source, we have access to the structural information of about 12,500 compounds and their corresponding band-gap. We transformed the structural information to a visual representation that can be processed through our proposed ML algorithm, and matched directly to the calculated band-gap value for training purposes.

2.2.1. 3D molecular image generation

The transformation phase in a ML problem consists of changing the original information type to another, usually easier to process, while preserving its main features as much as possible. This procedure seeks to obtain different and relevant information for the same data set using alternative criteria [29]. In this work, we use the structural data of the chemical compounds as represented in XYZ format files as obtained from the OMDB-GAP1. A XYZ file contains a list of the atoms belonging to a specific molecule or crystal unit cell, including the chemical symbol followed by the 3D spatial coordinates, in angstroms (Å, 1×10^{-10} meters). We reconstructed each compound from its XYZ coordinates, representing the atoms and bonds as model balls-and-sticks, with a specific color and size according to the JMol default element colors scheme [30] (see Table S1 and Figure S2 in Supplementary Information). The 3D molecular models are graphically constructed using the matpotlib framework, as implemented in an own-coded script (for Python 3.7.3) [31]. From each of these 12,500 molecular models, we extract three 2D images in three different plane projections: XY, XZ and YZ. The extracted 2D images are converted to JPG format of 196×268 pixels size. All the images are generated with a standardized render stage: black background and fixed illumination. The result of this transformation phase was 37,500 RGB images of 196×268 pixels each. Fig. 2 depicts a visual summary of the process.

2.3. Feature extraction and band-gap prediction

In ML problems and specially deep-learning, CNN have revolutionized the way of computer vision problems are tackled due to the automatic discovery of structure representation in large data sets. This method has dramatically improved the state-of-the-art in image processing [28]. CNN have three factors involved in their learning process: sparse interaction, parameter sharing and quasi-variant representation. A CNN is a multi-layer neural network that consists of two different types of layers: convolution layers (c-layers) and sub-sampling layers (s-layers). C-layers and s-layers are connected alternately and form the feature extraction part of the network [28,32]. However, to find a suitable architecture of a particular CNN is a hard task [28]. In that sense, literature have reported multiple types of network architectures, depending on the solving-problem. For example, in recent years, many network structures for image recognition, classification and regression problems have been reported like: AlexNet [33], ClarifaiNet [34], GoogLeNet [35] and VGGNet [36]. All these networks have proved to be efficient in their own problem domains; but also, they can be used as pre-trained models so that users can reduce the amount of time when re-training them for another specific task.

CNN have not been widely used for regression tasks, although they are also a fundamental problem in ML [37]. Regression and classification problems are strongly correlated [38], and can be transformed from one to another in many scenarios, as in this case study. For example, the softmax classifier is related to the softmax regression [39]. Some successful applications of CNN in regression tasks are [37,40,41].

In order to test different approaches applying CNN in our image-based band-gap regression problem, we explored two different architectures using a multichannel 2D CNN and a 3D CNN. These architectures

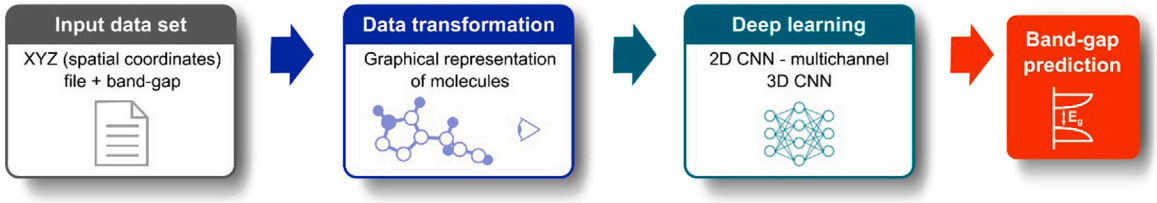

Fig. 1. Scheme of our proposed methodology for band-gap prediction using images of projected molecules in three planes (x–y, x–z, y–z). Our approach consists in four main steps: data collection, data transformation, deep-learning networks training, and electronic band-gap prediction. We take advantage of two deep-learning techniques: multichannel 2D CNN and 3D CNN, to achieve the training and inference phases resulting in the band-gap regression.

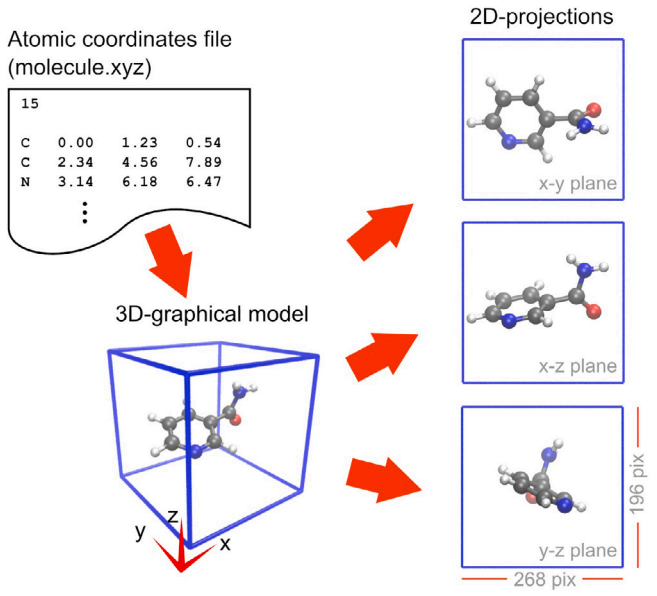

Fig. 2. Scheme of 2D-projection images generation, from the 3D molecular model constructed on the atomic coordinates (XYZ file) of each molecule. The XYZ file contains the atomic coordinates (in angstroms) and the atom type, which are rendered in a 3D molecular model with atoms represented as colored balls, and bonds (if represented) as sticks. The generated images of three 2D planes: x–y, x–z and y–z are 196 pixels wide by 268 pixels height. For illustration purposes, the black background was not rendered in this figure.

were implemented due to the excellent results achieved with classification and regression problems using 3D representations in different fields as medical resonance imaging [42] and predicting mechanical properties of materials with 3D representations [43]. Both architectures and their respective hyper-parameters were chosen through a Bayesian optimization process described next.

2.3.1. CNN architecture optimization

We conducted a Bayesian optimization [44] to find suitable values of the hyper-parameters in both multichannel 2D CNN and 3D CNN models. This optimization approach attempted to find the global optimum in a minimum number of iterations and set the optimal architecture of both approaches. For this optimization process, we maximized the mean-absolute error (MAE) in each model over 100 iterations using the Bayesian optimization library for Python [45].

We considered the following hyper-parameters in the Bayesian optimization process for the CNN architecture: the number of convolutional layers (from 1 to 50), the number of units in each convolutional layer (from 16 to 512), the convolutional layers (from 1 to 50,000), and the number of hidden layers in the fully connected network (from 1 to 50) with ReLU units (from 128 to 50,000), and dropout (from 0.1 to 0.5) between layers. The ReLU activation function was implemented in the fully connected network, and a linear activation unit for the band-gap regression in both architectures; multichannel 2D CNN and 3D CNN. The results of the optimization architecture process are described in Fig. 3 for the multichannel 2D CNN and Fig. 4 for the 3D CNN (see Section 3 for more details). Table 1 summarizes the numerical values of the optimization process for both models.

Table 1
Optimized values of the hyper-parameters of the CNN model via Bayesian optimization.

Model	Hyper-parameter	Optimized value
Multichannel 2D CNN	Convolutional layers	6
	Filter size of c-layers	{16,32,64,128,256,256}
	Hidden layers	2
	Neurons in hidden layers	{20000,8000,1}
3D CNN	Convolutional layers	6
	Filter size of c-layers	{32,64,128,256}
	Hidden layers	2
	Neurons in hidden layers	{8196,4096,1}

2.3.2. Multichannel 2D CNN

The optimized architecture of the multichannel 2D CNN model is shown in Fig. 3. This CNN received as inputs the three images from the projected views of the molecules in the XY, XZ, and YZ planes. For each image, the CNN consisted of three independent channels of six 3×3 ReLU-based convolutional layers with varying filter sizes $(16, 32, 64, 128, 256, 256)$ and a max-pooling layer of size 2×2 per filter. These channels were then concatenated and flattened. The latter result was fed into a fully connected network (i.e., dense layer) that was comprised of two ReLU-based hidden layers with 20,000 and 8000 units each, a batch normalization between hidden layers, and the output layer with linear activation functions. Lastly, the CNN produced the band-gap prediction response accordingly to the input images. Since the concatenated channels contained the features extracted from the three orthogonal planes, our multichannel 2D CNN model effectively takes into account the 3D spatial information of the molecules.

2.3.3. 3D CNN

We also assessed the band-gap prediction by a 3D CNN model, adding an extra pre-processing step to built a 3D-vector concatenating the three orthogonal projected images (i.e. XY, XZ and YZ plane images). This $196 \times 268 \times 3$ size 3D-vector contained the concatenation of three RGB images, which was fed as input into a feature extraction step built with 4 blocks: one 3D convolutional layer with 32 filters of $3 \times 3 \times 3$ kernel size and ReLU as activation function. The second block was made of one 3D convolutional layer with 64 filter with the same configuration. The third block has two 3D convolutional layers with 128 filters, and finally, a 3D convolutional layer with 256 filters in a similar fashion. At the end of each 3D convolutional block, we included one 3D max-pooling layer with $2 \times 2 \times 2$ pool size, used to extract the relevant information from each convolutional block. The band-gap regression step was carried out by three fully connected layers: 8192 ReLU in the first one, 4096 ReLU in the second one, and finally a layer with a linear activation function. Each fully connected layer was divided by 0.30 of dropout layers. The optimized 3D CNN architecture approach is shown in Fig. 4.

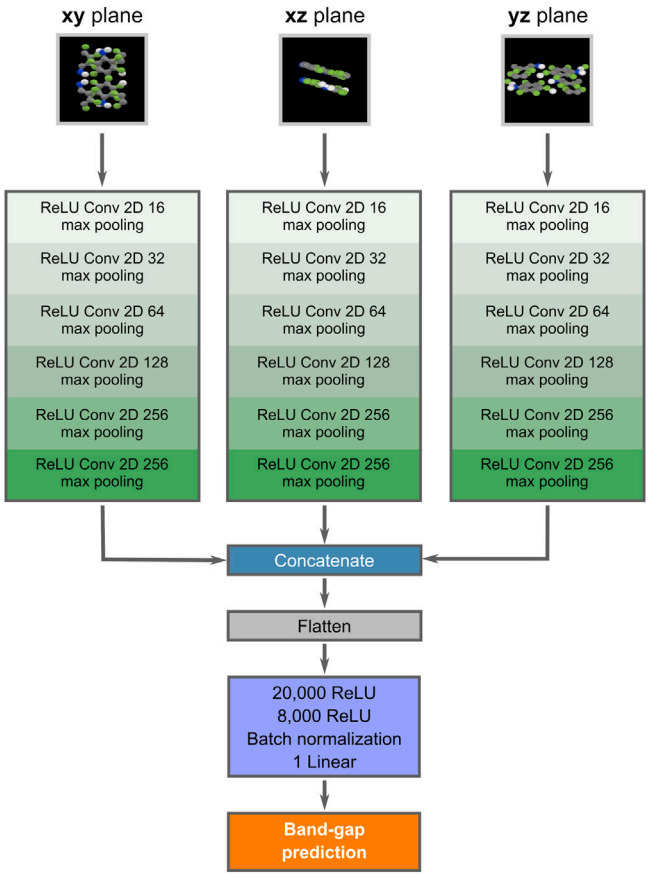

Fig. 3. Representation of the proposed multichannel 2D CNN architecture for band-gap prediction. It consists of three parallel 6-depth convolutional layers that are concatenated and flattened, and a fully connected network with two hidden layers. Three 2D-projection images of molecules are input into the network, and a band-gap (eV) estimation value is output.

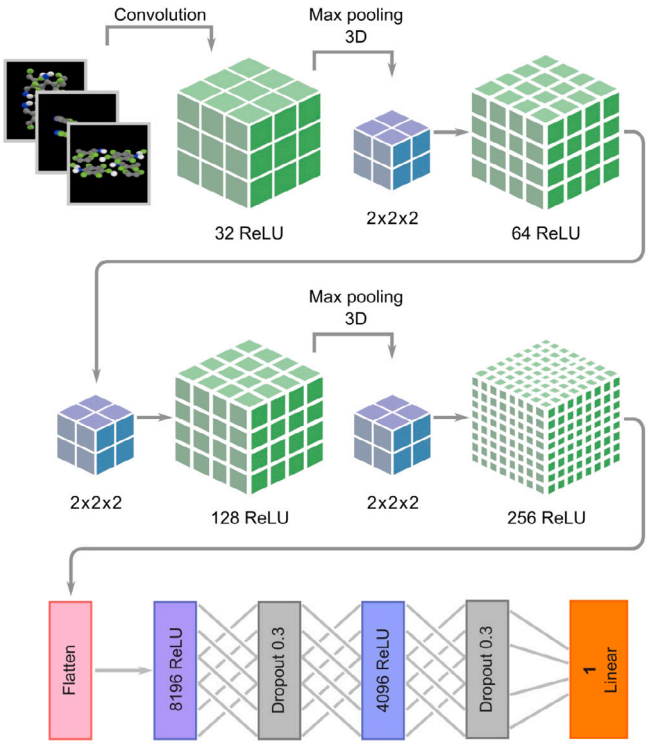

Fig. 4. Representation of the proposed 3D CNN architecture for band-gap prediction. It consists of four 3D convolutional layers that are flattened, and a fully connected network with two hidden layers. Three 2D-projection concatenated images of molecules are input into the network, and a band-gap (eV) estimation value is output.

2.3.4. Training networks

As described above, the OMDB-GAP1 data set is integrated by information from 12,500 organic compounds. In order to train both approaches in the same way, we split the data by taking 10,000 (80%) of the molecular structures for the training step, and the remaining 2500 molecules (20%) for testing. These percentages were chosen based on previous works, in order to compare our results with deep-learning approaches such as SchNet [45] and kernel ridge regression based on the Smooth Overlap of Atomic Positions (SOAP) approach [46].

The training data set was composed of 30,000 RGB images with 196 × 268 pixels size, whereas the testing data set comprised 7500 RGB images with the same size. For training purposes, we performed a Bayesian optimization process as described previously. This phase was performed over 100 iterations, using 50 epochs in each iteration, the Adam optimizer [47] and the loss function MAE.

2.4. DFT calculations

In order to further assess the applicability of our trained ML model, we performed DFT calculations for 10 additional organic crystalline structures (available as supplementary data), by means of SIESTA package [48]. These additional structures were not part of the original OMDB-GAP1 database (i.e., not part of the training nor validation set), although they were constructed based on compounds within the database. The modifications consisted of the replacement of some random atom types within the original structure, changing some of them for others with equivalent valence electrons (e.g., replacing O for S, or F for Cl), or, as it was for few cases, replacing N with C. The DFT calculations were carried out within the framework of the general gradient approximation (GGA), with the Perdew–Burke–Ernzerhof (PBE) approach. We used a double-ζ basis set plus polarization orbitals (DZP), the real-space grid used for charge and potential integration is equivalent to a plane wave cut-off energy of 100 Ry. The pseudopotentials were built from the valence electrons for each atomic species. All the calculations were performed at the Γ-point only, after a structural relaxation phase by conjugate gradient minimization until the maximum forces were lower than 0.1 eV Å^{-1}.

3. Results and discussion

This section presents the experimental results with our both CNN models. In order to analyze our proposed multichannel 2D CNN and 3D CNN architectures, the following procedures were carried out: (i) a comparison of the performance between the multichannel 2D CNN and the 3D CNN optimized models, (ii) an analysis of the band-gap prediction using our best CNN model found and a comparison against the state-of-the-art approaches, using the OMBD-GAP1 data set, and (iii) a test experiment using our best CNN model found in different unobserved data.

Regarding the metrics used to measure the performance of the CNN architectures, we calculate the training time (in minutes) and the average inference time (in milliseconds); we also computed the MAE as shown in (1) and the root mean-squared error (RMSE) as shown in (2); where, the e_i presents the difference between the target value over the predicted value, and n is the number of samples [49].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{2}$$

3.1. Performance of the CNN models for band-gap prediction

First, we provide the results obtained after optimizing and training the multichannel 2D CNN and the 3D CNN models. Table 2 shows the comparative results of both CNN models, using the testing data, in terms of MAE and RMSE. The closer to zero, the better the model is performing. As it can be seen that the multichannel 2D CNN model outperforms the 3D CNN model with lower error values (0.13 eV lower MAE and 0.31 eV lower RMSE).

Furthermore, the overall training time for 10,000 molecular models (training set) is around 30% shorter for the 3D CNN (71 min vs. 49.4 min), as shown in Table 2. Also, the average time that the multichannel 2D CNN takes to infer the band-gap is 0.8170 ms, while the 3D CNN takes 0.7194 ms (Table 2). The difference of time in executing an estimation is around 12% shorter for the 3D CNN. In this case study, the training and testing phases were performed in a computer with an Intel Core i-7 processor and 32 GB in RAM and a Graphics Processing Unit (GPU) Nvidia Titan RTX with 24 GB.

Several observations can be seen in the previous experiment. First, the multichannel 2D CNN model gets considerable advantage over the 3D CNN model, as shown in the overall performance of MAE and RMSE. Second, the 3D CNN model is less time-consuming in training and inferring than the multichannel 2D CNN, but the difference in the average inference time between the two models are around 12% which might not be very significant. Also, the multichannel 2D CNN does not require any additional processes for concatenating the projected XY, XZ and YZ images, as the 3D CNN does. Hereafter, due to the overall performance and small difference in the average inference time, the results and discussion will be related to the implementation of the multichannel 2D CNN architecture for band-gap prediction of organic compounds.

3.2. Multichannel 2D CNN band-gap prediction analysis

Using our multichannel 2D CNN model, we estimated the band-gap of the 2500 samples in the testing set. Fig. 5 shows a plot with a comparison between the actual electronic band-gap value (from DFT calculation) and the predicted band-gap value obtained from the CNN model. The results are shown in three separated categories: the predicted values which differ less than 0.5 eV from the actual band-gap (depicted as green dots), those which differ more than 0.5 but less than 1.0 eV (shown as blue dots), and those band-gap values which differ more than 1.0 eV from the real one (depicted as red dots). Roughly, the 37% of the results fall within the first category (less than 0.5 eV), which represents the best prediction performance, whereas 29% and 33% correspond to the second and third categories, respectively.

Compared with other recent and relevant works [8,11,13,18], which use specific descriptors such as stoichiometry, crystalline space group, band-gap category and geometry (among others), our ML proposed model performs similarly to some of them. Nevertheless, it is worth to consider that our approach is solely based on visual representations of molecular models, instead of feature-based ML models using large sets of alphanumerical descriptors associated to specific characteristics of materials. Literature reports a limited number of works for predicting the electronic band-gap using ML models with the OMDB-GAP1 data set. In fact, some of the most interesting works proposing ML methods for processing graphical representations of molecular structures, predict only the toxicity of such molecules, with an outcome classification limited to “toxic” or “non-toxic” [16,24,25], which, could be quite useful in practical terms, does not deal with the difficulty of predicting a key feature as the electronic band-gap within reasonable value ranges. In this regard, we compare our proposal with two feature-based ML approaches found in the state-of-the-art, SchNet [45] and SOAP [46], as reported in Table 3. Even though our proposal gets a MAE difference of 0.290 eV and 0.263 eV against the SchNet and SOAP, respectively, and gets an RMSE difference of 0.2483 eV and 0.2133 eV respectively,

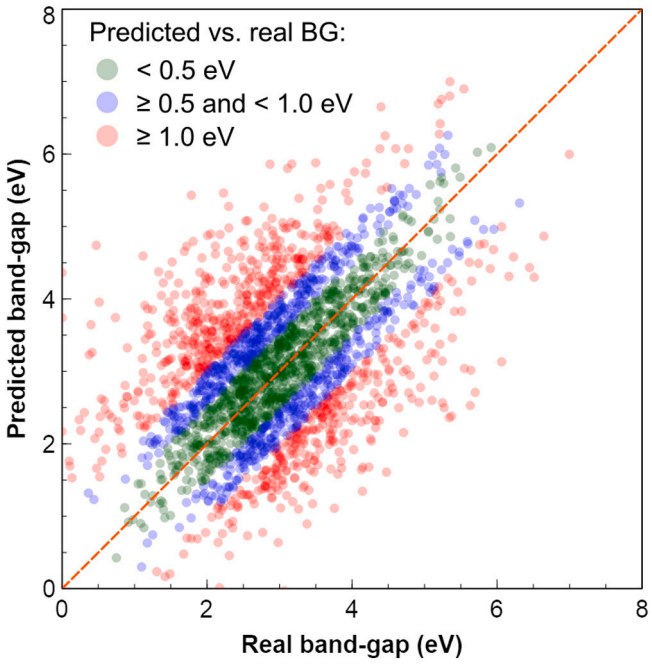

Fig. 5. Image-based band-gap prediction for the 2500 test molecules using our proposed multichannel 2D CNN model. The results are represented in three categories: real vs. predicted band-gap (BG) difference < 0.5 eV (green dots), BG difference ≥ 0.5 and < 1.0 eV (blue dots), and BG difference ≥ 1.0 eV (red dots). The dashed line represents the ideal (no difference) case.

it is worth noting to say that the SchNet and SOAP methods are based on the prior calculation of descriptors (features) and ours not. The multichannel 2D CNN model only requires the representation of the molecular structure as a set of 2D projection images that is less consuming in both time and computational efforts.

Further analysis of the results demonstrates interesting features of the multichannel 2D CNN model. Fig. 6a shows the normalized count of C, O and *N* atoms (with respect to H atoms), whereas Fig. 6b shows the count of Cl, S, F, and Br (also normalized with respect to H). It can be seen that the electronic band-gap prediction is in general not influenced by the relative amount of elements such as carbon, oxygen and nitrogen, which remains basically constant for all the analyzed organic compounds. Nevertheless, it can be observed that the band-gap prediction accuracy worsens as the relative amount of chlorine, sulfur, fluorine or bromine increases in 37%, 79%, 43%, and 22% (respectively) for the mid-range values, and 73%, 47%, 44%, and 27% (respectively) for the results with larger differences between predicted and real band-gap values. It is interesting that three of these four elements are halogens, which in practice can exert an enormous difference in the reactivity of the compounds that contain them, given their intrinsic electronegativity. A recent work addresses this particular issue [50], also within the context of ML methodologies for molecular properties prediction. The authors describe the extension of a deep learning molecular potential (the ANI-1x neural network potential), which is a deep learning network adapted specifically to accurately predict the energy of structures containing a defined set of atomic species, originally trained for dealing with compounds containing H, C, O, and N [51]. The described extension includes S, F and Cl, which result specially complex to parametrize in terms of molecular torsion profiles. Therefore, our results are relevant and promising, since our approach was sensitive to the presence of these precise chemical species by simply using an image-based architecture, which rendered a consistent behavior for prediction of the electronic band-gap of organic crystalline structures.

Table 2
Test performance of the multichannel 2D CNN and the 3D CNN models.

Model	MAE (eV)	RMSE (eV)	Training time (min)	Average inference time (ms)
Multichannel 2D CNN optimized	0.6780	0.7673	71.0	0.8170
3D CNN optimized	0.8080	1.0780	49.4	0.7194

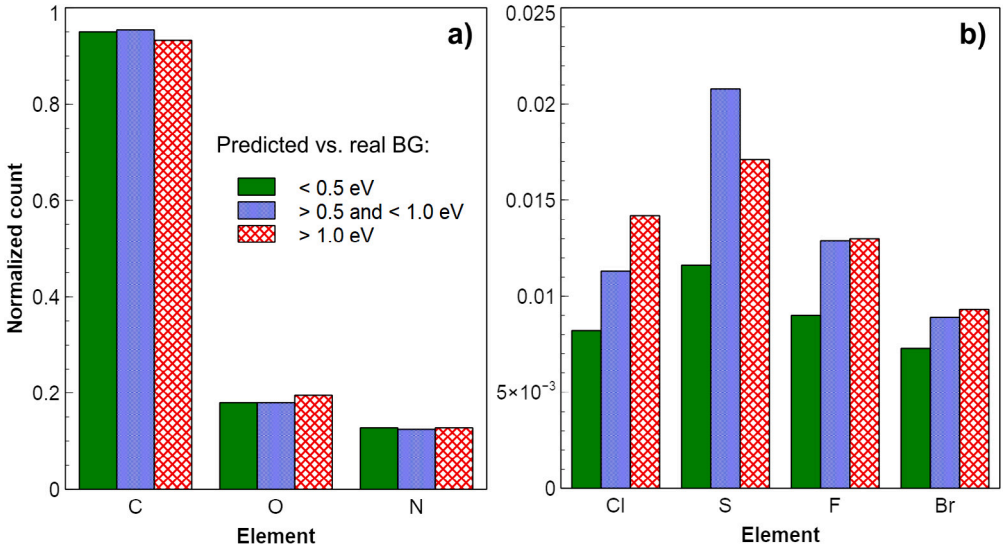

Fig. 6. Normalized count species (with respect to hydrogen) for: (a) C, O and N, and (b) Cl, S, F, and Br. The green bars depict the count of the compounds which real vs. predicted electronic band-gap difference is less than 0.5 eV, the blue bars correspond to differences in between 0.5 and 1.0 eV, and the red crossed hatch corresponds to difference values larger than 1.0 eV.

Table 3
Comparison between the multichannel 2D CNN model and the-state-of-the-art methods for band-gap prediction.

Model	MAE (eV)	RMSE (eV)
SchNet [45]	**0.388**	**0.5190**
SOAP [46]	0.4150	0.5540
Multichannel 2D CNN (ours)	0.6780	0.7673

3.3. Band-gap prediction over new crystalline organic structures

As mentioned before, in order to further test the validity and applicability of our proposed multichannel 2D CNN model, we performed a DFT-based analysis of 10 new crystalline organic structures, by the random modification of structures already existent in the OMDB-GAP1 database. The methodology for the DFT calculation was described before, and the results of the band-gap determination by DFT and our proposed multichannel 2D CNN can be compared in Fig. 7. As it can be observed, the performance of our multichannel 2D CNN compared with conventional DFT methods, in terms of band-gap prediction performance, can be considered as remarkable since it get a MAE of 0.4371 eV and an RMSE of 0.6611 eV. Also, the R^2 of 0.946 value reveals a notable correlation between the real band-gap and the estimated band-gap. Thus, our proposal, given the improved processing speed of the trained network, represents a huge advantage for a quick screening for new materials structures.

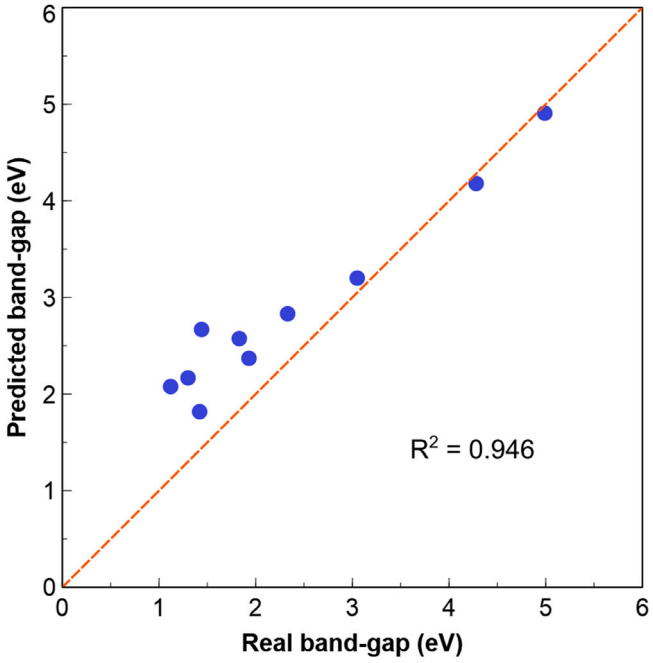

Fig. 7. Results for the band-gap prediction using the multichannel 2D CNN in new molecular structures outside the OMDB-GAP1 data set. The new molecules were relaxed and their electronic ground state calculated by a conventional DFT methodology for band-gap determination. Results report a MAE of 0.4371 eV and an RMSE of 0.6611 eV.

3.4. Discussion

At this point, it is important to highlight again the differences and novelty of our proposed ML model: in terms of performance, the image-based multichannel 2D CNN model for band-gap prediction offers a comparable solution to the state-of-the-art methods (such as SOAP and SchNet). The observed results support the evidence about the promising performance of ML methodologies in general, and of our proposed band-gap prediction model based on visual representation of molecular structures in particular. It is remarkable to say that SOAP and SchNet,

as most of the related research, are based on descriptive information of the molecules as inputs, while our proposal is based on images that are easily generated from XYZ coordinates files, and provides the potential advantage of exploring geometrical variations in molecules as factors for electronic band-gap. The advantages of our proposal can be summarized as follows: (i) the image-based band-gap prediction approach offers a flexible graphical solution for materials characterization with acceptable results, (ii) it paves the way to a new synergistic approach by combining the potential of ML methods in vision solutions, with their also promising performances of materials properties prediction, which undoubtedly will produce further interest and research efforts, and (iii) it adds an interesting alternative to material scientists to quickly assess and screen potential useful molecular structures.

Nevertheless, it is important to consider some limitations of our proposal, which would be related to the fact that a vision-based approach is subjected to the quality of images. Thus, the reconstruction of the analyzed images is a critical step, which in terms of crystalline compounds, could be problematic since the generation of suitable projections of the 2D images is a delicate task. The latter due to the predominant viewpoints each molecule could have, and that it could impact to a considerable extent the ability of the vision-based system to provide an accurate prediction. This problem is more related to the finding of a better way to graphically represent the molecular models, in terms of variations in sphere's size for each atomic species, different color schemes, etc., which could make the CNN better to recognize and predict the desired properties. This topic could actually be of interest for future research, in order to optimize the proposal presented here.

Lastly, the obtained results demonstrated that our proposal is acceptable compared to the state-of-the-art for band-gap prediction methods, new in terms of the visual information used in the band-gap predictions, and with the inherent advantages of ML methods by avoiding the high computational costs of the traditional models.

4. Conclusion

In this paper, we presented a 3D orthogonal image-based band-gap prediction approach, aiming to take advantage of the robustness that ML methods have shown in vision applications for image-based developments. Using the OMDB-GAP1 data set as a case study, we studied first the optimization of two candidate CNN architectures –a multichannel 2D CNN and a 3D CNN–, through Bayesian optimization, for their training and use in electronic band-gap prediction, which is a key characteristic within materials science research. Then, we compared our multichannel 2D CNN proposal (which resulted suitable for this study) with two state-of-the-art ML-based models, to say, SOAP and SchNet. Our multichannel 2D CNN model reported a MAE of 0.6780 eV and a RMSE of 0.7673 eV. These results are acceptable for a first attempt to predict band-gap using only the projection images of a 3D molecular model, compared with recent feature-based ML methods from the literature.

We also believe that our proposal contributes with an interesting option for quick exploration of new crystalline materials, due to the prediction made only with projected structural images and not with a full description of the x, y, z positioning of atoms in molecules, as it is with traditional and costly approaches such as DFT. It is important to highlight that this particular advantage of ML approaches over conventional ones, such as *ab-initio*, lies precisely on the possibility of quick prediction of certain features (e.g., electronic band-gap) that, in situations where speed could outweigh precision, would allow materials scientists to quickly screen relevant structures for deeper study, saving important amounts of time and computational resources.

Finally, we would like to note that the results presented here could fuel more research in the application of ML methods commonly used in other areas (e.g. automated vision systems) for materials research. We are particularly interested in further analysis of image-based determination of electronic properties for molecules including specific types of atomic species, such as halogens, given that we could observe that our multichannel 2D CNN was able to indirectly differentiate structures with this kind of elements. It could be expected then that the results presented here could work as a reference point for further refinement of ML architectures and training methods, and for expanding the applicability of computational tools to novel materials discovery and research.

CRediT authorship contribution statement

Ricardo Espinosa: Idealized the concept, Designed and tested the Machine Learning algorithms, Contributed in the manuscript writing and revision. **Hiram Ponce:** Designed and tested the Machine Learning algorithms, Contributed in the manuscript writing and revision. **Josue Ortiz-Medina:** Idealized the concept, Performed the DFT calculations, Contributed in the manuscript writing and revision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The XYZ files for the additional structures (apart from structures within the OMDB-GAP1 database) used for our ML model validation are available to download from https://1drv.ms/u/s!AgoMuI4-UMpbu1n-3vPvyFGpZv3y?e=jotM9W.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.commatsci.2021.110967.

References

[1] S.V. Shishkina, Using of quantum-chemical calculations to molecular crystals studying, Struct. Chem. 30 (2019) 1565–1577, http://dx.doi.org/10.1007/s11224-019-01397-8.

[2] A. Jothi, Principles, challenges and advances in ab initio protein structure prediction, Protein Pept. Lett. 19 (11) (2012) 1194–1204, http://dx.doi.org/10.2174/092986612803217015.

[3] J. Pokluda, M. Černý, M. Sob, Y. Umeno, Ab initio calculations of mechanical properties: Methods and applications, Prog. Mater. Sci. 73 (2015) 127–158, http://dx.doi.org/10.1016/j.pmatsci.2015.04.001.

[4] R.-Q. Zhang, W.-J. Fan, Economical basis sets and their uses in ab initio calculations, Int. J. Quantum Chem. 115 (9) (2015) 570–577, http://dx.doi.org/10.1002/qua.24830, URL http://doi.wiley.com/10.1002/qua.24830.

[5] S. Luber, Recent progress in computational exploration and design of functional materials, Comput. Mater. Sci. 161 (2019) 127–134, http://dx.doi.org/10.1016/j.commatsci.2019.01.040.

[6] A. Bianco, Y. Chen, E. Frackowiak, M. Holzinger, N. Koratkar, V. Meunier, S. Mikhailovsky, M. Strano, J.M. Tascon, M. Terrones, Carbon science perspective in 2020: Current research and future challenges, Carbon 161 (2020) 373–391, http://dx.doi.org/10.1016/j.carbon.2020.01.055, URL https://linkinghub.elsevier.com/retrieve/pii/S0008622320300622.

[7] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, Machine learning driven new material discovery, Nanoscale Adv. (2020) http://dx.doi.org/10.1039/d0na00388c, URL http://xlink.rsc.org/?DOI=D0NA00388C.

[8] C.E. Belle, V. Aksakalli, S.P. Russo, A machine learning platform for the discovery of materials, J. Cheminformatics 13 (1) (2021) 42, http://dx.doi.org/10.1186/s13321-021-00518-y, URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00518-y.

[9] Ç. Odabaşı Özer, R. Yıldırım, Performance analysis of perovskite solar cells in 2013–2018 using machine-learning tools, Nano Energy 56 (2019) 770–791, http://dx.doi.org/10.1016/j.nanoen.2018.11.069.

[10] M.H. Lee, Insights from machine learning techniques for predicting the efficiency of fullerene derivatives-based ternary organic solar cells at ternary blend design, Adv. Energy Mater. 9 (26) (2019) http://dx.doi.org/10.1002/aenm.201900891.

[11] W. Yang, T.T. Fidelis, W.H. Sun, Machine learning in catalysis, from proposal to practicing, ACS Omega 5 (2020) 83–88, http://dx.doi.org/10.1021/acsomega.9b03673.

[12] Q. Meng, Optical, electrical, and catalytic properties of metal nanoclusters investigated by ab initio molecular dynamics simulation: A mini review, in: Photoinduced Processes At Surfaces and in Nanomaterials, American Chemical Society, 2015, pp. 215–234, http://dx.doi.org/10.1021/bk-2015-1196.ch011, URL https://pubs.acs.org/doi/10.1021/bk-2015-1196.ch011.

[13] F. Shayeganfar, R. Shahsavari, Deep learning method to accelerate discovery of hybrid polymer-graphene composites, Sci. Rep. 11 (1) (2021) 15111, http://dx.doi.org/10.1038/s41598-021-94085-9, URL http://www.nature.com/articles/s41598-021-94085-9.

[14] N. Borodinov, S. Neumayer, S.V. Kalinin, O.S. Ovchinnikova, R.K. Vasudevan, S. Jesse, Deep neural networks for understanding noisy data applied to physical property extraction in scanning probe microscopy, Npj Comput. Mater. 5 (1) (2019) http://dx.doi.org/10.1038/s41524-019-0148-5.

[15] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N.T.P. Hartono, S. Ramasamy, B.L. DeCost, S.I. Tian, G. Romano, A. Gilad Kusne, T. Buonassisi, Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks, Npj Comput. Mater. 5 (1) (2019) http://dx.doi.org/10.1038/s41524-019-0196-x, arXiv:1811.08425.

[16] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, A. Agrawal, ElemNet: Deep learning the chemistry of materials from only elemental composition, Sci. Rep. 8 (1) (2018) 17593, http://dx.doi.org/10.1038/s41598-018-35934-y, URL http://www.nature.com/articles/s41598-018-35934-y.

[17] V. Venkatraman, The utility of composition-based machine learning models for band gap prediction, Comput. Mater. Sci. 197 (2021) 110637, http://dx.doi.org/10.1016/j.commatsci.2021.110637, URL https://linkinghub.elsevier.com/retrieve/pii/S0927025621003645.

[18] M. Fernandez, H. Shi, A.S. Barnard, Geometrical features can predict electronic properties of graphene nanoflakes, Carbon 103 (2016) 142–150, http://dx.doi.org/10.1016/j.carbon.2016.03.005.

[19] G. Sosso, M. Bernasconi, Harnessing machine learning potentials to understand the functional properties of phase-change materials, MRS Bull. 44 (9) (2019) 705–709, http://dx.doi.org/10.1557/mrs.2019.202.

[20] R. Espinosa, H. Ponce, S. Gutiérrez, L. Martínez-Villaseñor, J. Brieva, E. Moya-Albor, A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-fall detection dataset, Comput. Biol. Med. 115 (2019) http://dx.doi.org/10.1016/j.compbiomed.2019.103520, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074271336&doi=10.1016%2Fj.compbiomed.2019.103520&partnerID=40&md5=ef8677c4cdef81a0237ce459adca348c.

[21] T.D. Linh, M. Arai, Two-stage deep neural network for general object detection, J. Inf. Process. 27 (2019) 268–277, http://dx.doi.org/10.2197/ipsjjip.27.268.

[22] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, International Conference on Learning Representations, ICLR, pp. 1–14.

[23] J. Townsend, C.P. Micucci, J.H. Hymel, V. Maroulas, K.D. Vogiatzis, Representation of molecular structures with persistent homology for machine learning applications in chemistry, Nature Commun. 11 (1) (2020) 3230, http://dx.doi.org/10.1038/s41467-020-17035-5, http://www.nature.com/articles/s41467-020-17035-5.

[24] M. Fernandez, F. Ban, G. Woo, M. Hsing, T. Yamazaki, E. LeBlanc, P.S. Rennie, W.J. Welch, A. Cherkasov, Toxic colors: The use of deep learning for predicting toxicity of compounds merely from their graphic images, J. Chem. Inf. Model. 58 (8) (2018) 1533–1543, http://dx.doi.org/10.1021/acs.jcim.8b00338, https://pubs.acs.org/doi/10.1021/acs.jcim.8b00338.

[25] E. Asilar, J. Hemmerich, G.F. Ecker, Image based liver toxicity prediction, J. Chem. Inf. Model. 60 (3) (2020) 1111–1121, http://dx.doi.org/10.1021/acs.jcim.9b00713, URL https://pubs.acs.org/doi/10.1021/acs.jcim.9b00713.

[26] C. Qian, Y. Xiong, X. Chen, Directed graph attention neural network utilizing 3D coordinates for molecular property prediction, Comput. Mater. Sci. 200 (2021) 110761, http://dx.doi.org/10.1016/j.commatsci.2021.110761, URL https://www.sciencedirect.com/science/article/pii/S0927025621004882.

[27] B. Olsthoorn, R.M. Geilhufe, S.S. Borysov, A.V. Balatsky, Band gap prediction for large organic crystal structures with machine learning, Adv. Quantum Technol. 2 (7–8) (2019) 1900023, http://dx.doi.org/10.1002/qute.201900023.

[28] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.

[29] Y. Roh, G. Heo, S.E. Whang, A survey on data collection for machine learning: A Big Data - AI integration perspective, IEEE Trans. Knowl. Data Eng. (2019) 1, http://dx.doi.org/10.1109/tkde.2019.2946162, arXiv:1811.03402.

[30] Jmol element colors scheme, URL http://jmol.sourceforge.net/jscolors/.

[31] J. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, http://dx.doi.org/10.1109/MCSE.2007.55.

[32] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.

[33] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), in: Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105, URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[34] M. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision, ECCV 2014 - 13th European Conference, Proceedings, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), PART 1, Springer Verlag, 2014, pp. 818–833, http://dx.doi.org/10.1007/978-3-319-10590-1_53, 13th European Conference on Computer Vision, ECCV 2014; Conference date, 06-09-2014 Through 12-09-2014.

[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, CoRR abs/1409.1556. URL http://arxiv.org/abs/1409.1556.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1–9, http://dx.doi.org/10.1109/CVPR.2015.7298594, URL https://ieeexplore.ieee.org/document/7298594.

[37] J. Yuan, B. Ni, A.A. Kassim, Half-CNN: A general framework for whole-image regression, 2014, CoRR abs/1412.6885. arXiv:1412.6885.

[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, in: Springer Series in Statistics, Springer, 2009, URL https://books.google.com.mx/books?id=eBSgoAEACAAJ.

[39] C. Bishop, Pattern Recognition and Machine Learning, in: Information Science and Statistics, Springer New York, 2016, URL https://books.google.com.mx/books?id=kOXDtAEACAAJ.

[40] N. Ilyas, A. Shahzad, K. Kim, Convolutional-neural network-based image crowd counting: Review, categorization, analysis, and performance evaluation, Sensors 20 (1) (2020) http://dx.doi.org/10.3390/s20010043.

[41] Y. Jiang, C. Li, Convolutional neural networks for image-based high-throughput plant phenotyping: A review, Plant Phenomics 2020 (2020) 1–22, http://dx.doi.org/10.34133/2020/4152816.

[42] J. Hu, Y. Kuang, B. Liao, L. Cao, S. Dong, P. Li, A multichannel 2D convolutional neural network model for task-evoked fMRI data classification, Comput. Intell. Neurosci. 2019 (i) (2019) http://dx.doi.org/10.1155/2019/5065214.

[43] C. Raoa, Y. Liua, Three-dimensional convolutional neural network (3D-CNN) for heterogeneous material homogenization, 2020, pp. 1–21, arXiv arXiv:2002.07600.

[44] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 4 (2012) 2951–2959, arXiv:1206.2944.

[45] F. Nogueira, BayesIan optimization: Open source constrained global optimization tool for python, 2014, URL https://github.com/fmfn/BayesianOptimization.

[46] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (18) (2013) 184115, http://dx.doi.org/10.1103/PhysRevB.87.184115, arXiv:1209.3140.

[47] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, pp. 1–15, arXiv:1412.6980.

[48] J.M. Soler, E. Artacho, J.D. Gale, A. García, J. Junquera, P. Ordejón, D. Sánchez-Portal, The SIESTA method for ab initio order- N materials simulation, J. Phys.: Condens. Matter 14 (11) (2002) 2745, URL http://stacks.iop.org/0953-8984/14/i=11/a=302.

[49] W. Wang, Y. Lu, Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in assessing rounding model, IOP Conf. Ser. 324 (1) (2018) http://dx.doi.org/10.1088/1757-899X/324/1/012049.

[50] C. Devereux, J.S. Smith, K.K. Davis, K. Barros, R. Zubatyuk, O. Isayev, A.E. Roitberg, Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens, J. Chem. Theory Comput. 16 (7) (2020) 4192–4202, http://dx.doi.org/10.1021/acs.jctc.0c00121, URL https://pubs.acs.org/doi/10.1021/acs.jctc.0c00121.

[51] J. Smith, O. Isayev, A. Roitberg, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, Chem. Sci. 8 (4) (2017) 3192–3203, http://dx.doi.org/10.1039/C6SC05720A, URL http://xlink.rsc.org/?DOI=C6SC05720A.