Deep Transfer Learning: A Fast and Accurate Tool to Predict the Energy Levels of Donor Molecules for Organic Photovoltaics

Gareth John Moore,* Olivier Bardagot, and Natalie Banerji

Molecular engineering is driving the recent efficiency leaps in organic photovoltaics (OPVs). A presynthetic determination of frontier energy levels makes the screening of potential molecules more efficient, exhaustive, and cost-effective. Here, a convolutional neural network is developed to predict the highest occupied and lowest unoccupied molecular orbital (HOMO/LUMO) levels of donor molecules for OPV. The model takes a 2D structure image and returns a prediction of its HOMO/LUMO levels comparable to experimental values. Insufficient experimental datasets are overcome with transfer learning where the model is initially trained on the large Harvard Clean Energy Project dataset and then fine-tuned using experimental data from the Harvard Organic Photovoltaic dataset. Error margins on predicted HOMO/LUMO levels below 200 meV are achieved, without any chemical knowledge implemented. Noticeably, the model outputs have higher accuracy and precision than corresponding density functional theory (DFT) estimations. The model and its limitations are further tested on a home-built dataset of commercially available donor polymers reported in OPVs (e.g., P3HT, PTB7-Th, PM6, D18). The results demonstrate both the practical utility of this model, to foster rational molecular engineering for OPV optimization, and the potential for deep learning techniques, in general, to revolutionize the energy materials research and development sector.

1. Introduction

With the burning of fossil fuels massively contributing to the current global warming crisis, the design, optimization and implementation of renewable alternatives for energy generation is critical to curb its already devastating effects.[1] Organic photovoltaics

G. J. Moore, O. Bardagot, N. Banerji
Department of Chemistry
Biochemistry and Pharmaceutical Sciences
University of Bern
Freiestrasse 3, Bern 3012, Switzerland
E-mail: gareth.moore@unibe.ch

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adts.202100511

(OPV) offer a cost-effective, lightweight, flexible and renewable light-to-electrical energy conversion process, and with efficiencies over 19%,[2] show promise as one of the viable alternatives to fossil fuels.[3] Much of these efficiency improvements have come from molecular engineering based on empirically determined design rules and trial-and-error methods. It has been shown, however, that the energies of the frontier molecular orbitals (Highest Occupied Molecular Orbital or HOMO, and Lowest Unoccupied Molecular Orbital or LUMO) can be used as a good approximation of the expected power conversion efficiency of materials in OPV devices.[4] Consequently, a presynthetic determination of these energies makes screening of potential materials more efficient. The theoretical determination of the HOMO/LUMO levels of organic molecules is traditionally achieved using Density Functional Theory (DFT) based calculations.[5] However, the accuracy of DFT simulations is limited by the inherent trade-off between over-delocalization and under-binding.[6] Besides, DFT simulations are computationally expensive and time consuming, thereby limiting the usefulness of DFT for large scale OPV power conversion efficiency predictions and material screening. To address these limitations, deep learning methods,[7] along with the development of ever larger datasets, have emerged as a promising alternative for the development of highly predictive quantitative structure-property relationship (QSPR) models in the field of OPV.[8]

In this work a QSPR deep learning model, in the form of a deep convolutional neural network, is developed to predict the HOMO/LUMO levels of organic molecules intended for use in OPV applications. The deep learning model takes the SMILES (simplified molecular-input line-entry system) of a molecule as input, converts it to a 2D RGB image, uses the convolutional layers of the network to extract features from the image and then converts the features into energy levels using a deep dense neural network.

All of the relevant information for predicting the frontier energy levels of a molecule is contained in the SMILES string, however, this string necessarily needs to be converted into a numerical form to be used as training data for a neural network. Previous
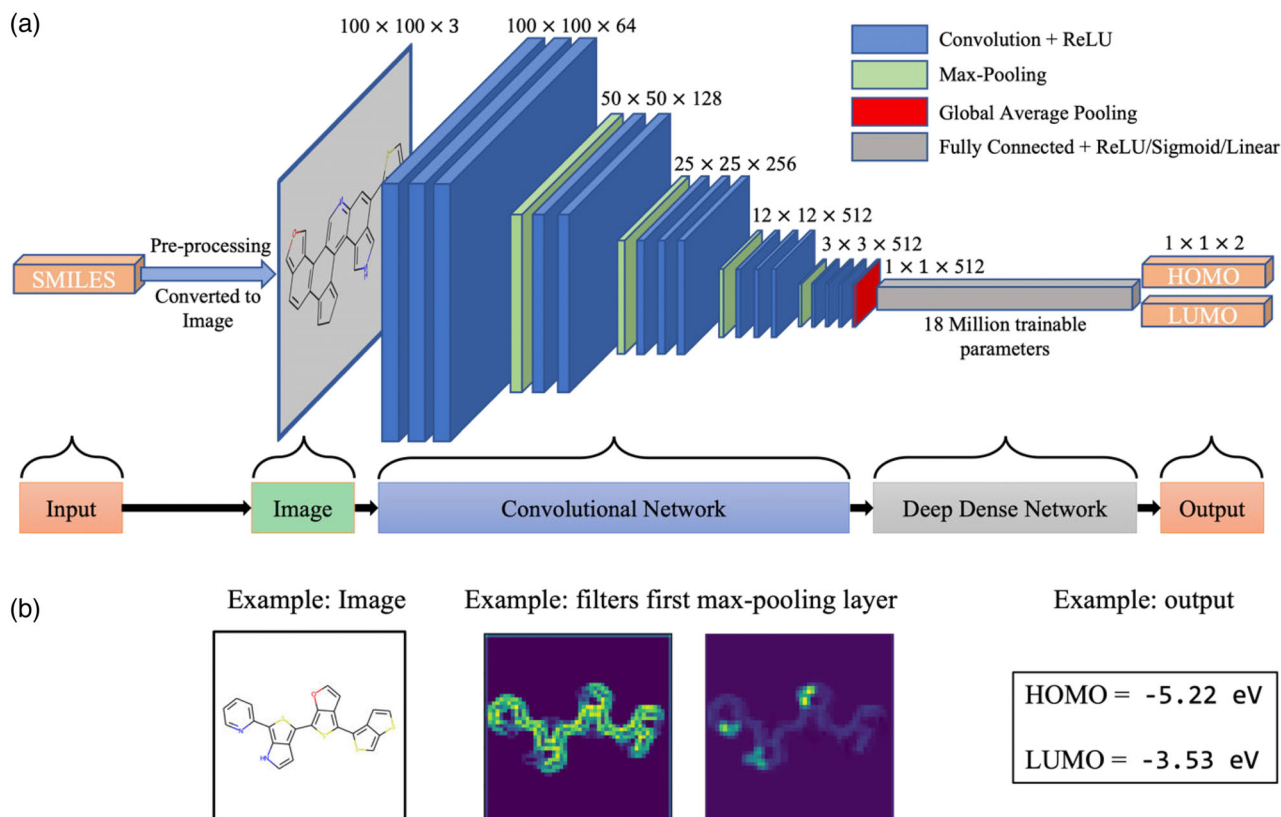
Figure 1. A) Illustration of the convolutional neural network, showing the preprocessing step, the convolutional network (with rectified linear unit (ReLU) activation function), max-pooling, global averaging, and fully connected layers (with activation functions). B) Examples of (left) a molecular image, (middle) the output of two filters after the first max-pooling layer and, (right) the output of the model.

attempts include converting SMILES to a numerical vector representing the letters of the SMILES by Paul et al. who were able to predict the HOMO levels of donor molecules using a Long Short Term Memory (LSTM) type network.[9] However, even though converting the SMILES to an image significantly increases the amount of data without increasing the amount of information, there are some advantages of the image representation. First, it allows the use of powerful computer vision techniques such as Convolutional Neural Networks (CNNs) to extract the information about the molecule, it also represents potential nonbinding interactions and conformational effects in a more accessible way than a SMILES string. Images representations of molecules and CNN's was used by Sun et al. to predict the theoretical photoconversion efficiencies of donor molecules for OPV applications.[10] Lastly, it was shown by Sun et al. that expanded molecular representations of over 1000 bits (a typical SMILES contains 250 bits) result in high prediction accuracy for a number of different model architectures, with the largest representation performing the best.[10]

The model takes advantage of a machine learning technique called transfer learning,[11] whereby it is first trained on a large dataset (500 000 molecules) with HOMO/LUMO estimated by DFT simulations and then fine-tuned on a smaller dataset (194 molecules) with experimentally measured HOMO/LUMO levels. The deep learning model shows an accuracy below 200 meV, with accuracy and precision superior to DFT-estimated energies.

The validity of the QSPR model was carefully evaluated and confirmed using commercially available polymers (such as P3HT, PTB7-Th, PNTz4T, J71, PM6, D18, Figure 5D) to ensure its practical utility. As a result, the deep learning model offers an efficient way to accurately and almost instantly (≈170 ms on a personal computer) predict the frontier energy levels of molecules, without the need for molecular geometry optimization and large computing clusters, thereby allowing fast and reliable screening of donor molecules for OPV applications. This model, and models of this kind, is expected to find rapid use in both academic and industrial laboratories to realize molecular engineering at a lower cost and in a fraction of the time.

2. Experimental Section

2.1. The Deep Learning Model

Deep learning has emerged as a powerful tool for solving a variety of problems of machine learning and artificial intelligence, both in everyday[12] and scientific applications.[13] Deep learning made use of multilayer stacks of modules (in this case convolutional layers and fully connected neurons, **Figure 1**A) that mapped an input and output through nonlinear functions.[14] By having many layers and millions of trainable parameters, the system was able to model increasingly complex processes in ways that were both sensitive to minute details and invariant to noise.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

Figure 1A illustrates the architecture of the model used in this work, where it is broken up into two distinct parts, the convolutional network and the deep dense network.[7b] This model took SMILES (or InChI, International Chemical Identifier) as input and generated standardized RGB molecular images as a preprocessing step. The convolutional network was used as a way of automatically creating a nonlinear, trainable feature extracting function without the need for explicit feature engineering or hardcoded pattern recognition. Carbons were represented in black and each heteroatom in its own color (e.g., oxygen in red, sulfur in yellow, nitrogen in blue) and representative letter (e.g., oxygen "O", sulfur "S", nitrogen "N"), this allowed for heteroatoms to be easily picked out as features even with the relatively low image resolution ($100\times100\times3$) which reduced the computational requirements. The convolutional layers slid multiple initially random filters over the image, transforming it such that features like edges, large shapes (e.g., conjugated backbones, Figure 1B middle left), or color change (e.g., heteroatoms, Figure 1B middle right) were highlighted. Then, it downsampled the data using a max-pooling layer until it was in a form appropriate for input into the dense network (1D array). Finally, the result of the convolutional network (or feature map) was fed into the deep dense network, which consisted of layers of fully connected neurons each containing trainable weights and activation functions that introduced nonlinearity into the model.[14] The deep dense network had around 18 million trainable parameters starting with an input array of 512 feature elements, from the convolutional network, and outputting 2 numbers that were trained to represent the HOMO and LUMO levels of the input molecule (details in Section S1, Supporting Information). This work made use of a supervised learning technique where the model parameters (or weights and biases) were iteratively optimized based on the results of a loss function that quantified the difference between the known true output and the predicted output. [13,14] Here, the loss function was calculated as the mean squared error (MSE) between the true and predicted values (details in Section S2, Figures S1 and S2, Supporting Information).

2.2. Datasets and Training

The ultimate objective of this work was to train a deep learning model able to predict the HOMO and LUMO energy levels of a molecule with experimental (not theoretical) values taken as "true." The main constraint was the lack of a large enough dataset containing experimentally measured values. With relatively small training sets, deep learning models were unlikely to learn in a way that gave meaningful predictivity. In order to overcome this constraint, a technique called transfer learning was employed.[11,15] In transfer learning, a model is trained on a large and general dataset where basic functions, which require many iterations and large amounts of data to learn, are acquired. The model was then retrained (fine-tuned) on a smaller more specific dataset using the previously learned weights to initialize the model.[11b,15,16] The second training, or fine-tuning, was done using a significantly smaller learning rate as the weights were assumed to be already close to optimal.

Here, the data used for the initial phase of training (phase I, **Figure 2**) were 500 000 randomly sampled molecules from
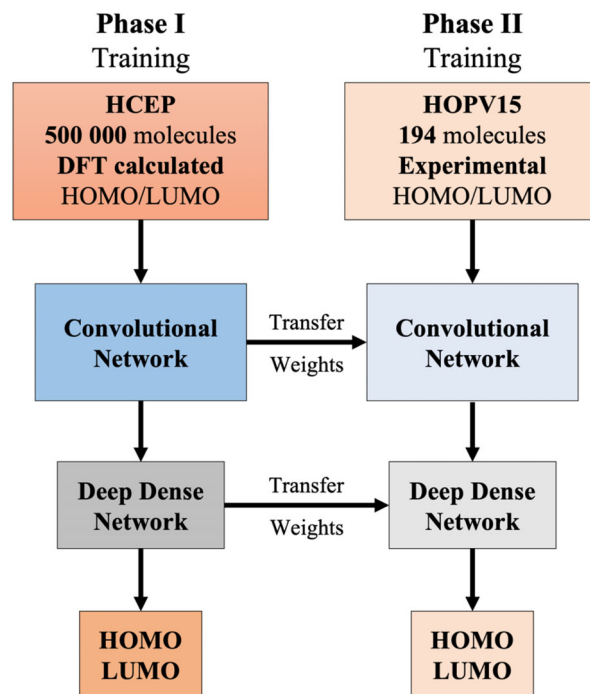

Figure 2. Schematic of phase I and phase II of training using the HCEP dataset and HOPV15 dataset respectively, with the learned weights being transferred from the first to the second phase of training.

the Harvard Clean Energy Project (HCEP) dataset.[17] The HCEP dataset consisted of around 2.3 million artificially generated potential donor molecules, for use in OPV devices. All molecules were designed combinatorially from 26 molecular building blocks. The HOMO and LUMO levels of these molecules were estimated using DFT at different levels (details in Section S3.1, Supporting Information).

The deep learning model was then fine-tuned (phase II, Figure 2) on a subset (194 molecules) of the Harvard organic photovoltaic dataset (HOPV15)[18] with the weights being carried over from the first Phase I. The HOPV15 dataset consisted of around 350 molecules whose HOMO and LUMO levels were been i) experimentally measured (extracted from literature) and ii) estimated using DFT in a range of conformations with four different functionals used for each conformation. A subset of the HOPV15 dataset was used in the training of the model (194 out of 350) consisting of only donor polymers to avoid the difficulty of unifying the SMILES of polymer and small molecule donors, as well as to represent the almost ubiquitous use of polymer donors, over small molecule donors, in OPV. The 194 molecules were converted into 5464 unique images representing all possible conformational representations of each molecule for training in Phase II. This data augmentation step ensured that the model was invariant to the specific SMILES chosen to represent a molecule, as one molecule could have multiple SMILES associated with it (details in point 2.3 and Table S2, Supporting Information). Phase II uses these experimental values i) as "true" for training and the resulting predictions of the model are later compared with the values from DFT ii). The dataset was created to represent a measurably diverse range of donor polymers used in the field of OPV (details in point 3.2 of the Supporting Information).

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

2.3. Testing and External Validation

Validation of QSPR models needed to be done rigorously in order to assess their predictivity.[19] Validation was done using three calculations: the square of the correlation coefficient (R^2), the root mean squared error (RMSE), and the standard error of prediction (SEP). The R^2 value is a measure of correlation between the true and predicted values, while the RMSE can be understood as the accuracy and the SEP as the precision of prediction.[19] The first step for validation was to split the dataset into a subset for training and a subset for testing. By doing so, the predictivity of the model could be assessed on a test set never exposed to the model and yet still representative of the entire dataset.[20] In this work, phase I was trained on 500 000 molecules and tested on 10 000 extra molecules (2%), and phase II was trained on 162 polymers and tested on 32 extra polymers (20%). In both training phases, the training and testing polymers were split with random sampling, but specifically in a way that the distribution of the HOMO and LUMO values was roughly the same. The training dataset was further split into training and validation sets and early stopping was employed to minimize overfitting (described in Section S2.1, Supporting Information). Then, to ensure the robustness of the model, a Y-scrambling test was done.[21] Here, the dependent variables (HOMO and LUMO values) were randomly scrambled and associated with the "wrong" structures (SMILES), in a way that the structure–property relationships no longer hold, and the model was retrained. If the correctly trained deep learning model shows high R^2 and the Y-scrambled model shows low R^2 values, it implies that the model outputs are neither overfitting nor chance correlations, but that there is necessarily a learnt link between the input (structure) and the output (properties).[8g,20]

2.4. The "use-case" Dataset

While it would be sufficient to test the deep learning model only on the test set from the HOPV15 dataset, from literature, an additional dataset was built with the aim of testing the model in a real use-case scenario. The so-called "use-case" dataset consisted of 26 donor polymers used in OPVs that are commercially available (i.e., including Chemical Abstracts Service or CAS numbers) and have both experimentally measured and DFT-estimated HOMO/LUMO energy levels published in peer-reviewed journals. The polymers were strictly not in the HOPV15 dataset but were composed of atoms and building-blocks represented in the Phase II training set (Figure S6, Supporting Information). Note that experimental values of the "use-case" dataset were exclusively determined by cyclic voltammetry (CV).[22] CV allowed an estimation of HOMO/LUMO energy levels with an error margin generally considered to be about ±100 mV.[23] More accurate techniques exist, such as ultraviolet photoelectron spectroscopy (UPS) or inverse photoelectron spectroscopy (IPES); however, CV was undoubtedly the most commonly used technique in the field of organic electronics due to the relative ease of measurement.[24] More details are given in Section S3.3 (Supporting Information) and a full list of the molecules with all values and predictions is given in Table S3 (Supporting Information).

3. Results

3.1. Phase I

In Phase I of training, molecules with DFT-estimated HOMO and LUMO levels from the large HCEP dataset are used to train the deep learning model with randomly initialized weights (Figure S1, Supporting Information). The goal of this phase is to leverage the large volume of data so that the model learns to extract important features from the molecular images and learns to convert those features into HOMO and LUMO energy level predictions. In order to evaluate the performance of the model at this stage, we generate predictions on the test set from the HCEP dataset. R^2 values close to 1 (**Figure 3**A,B) with SEP and RMSE values around 30 meV (Figure 3C) are found for the prediction of HOMO and LUMO levels, thereby illustrating the accuracy and precision of the prediction after phase I of training. This result is followed up by an R^2 value of 0.990 with SEP and RMSE of around 45 meV for the bandgap (Figure S3, Supporting Information), showing not only the individual predictions, but also their relative positions to be highly accurate and precise. The results from the Y-scrambling test shows R^2 values of 7.28×10^{-6} and 9.18×10^{-5} for the HOMO and LUMO levels respectively (**Table 1**). The R^2 values close to zero indicate that there is very little correlation between the predicted and "true" values, and that the Y-scrambling model is only able to predict random values within the range of the training data. The Y-scrambling test confirms that there is no overfitting and, more specifically, that there is a structure–property relationship in the dataset and that this relationship is necessarily learned by the model. Note that no chemical knowledge was implemented at any time.

3.2. Phase II

After phase I of the training, the deep learning model demonstrates an ability to predict DFT-estimated frontier energy levels of molecules represented in the HCEP dataset. The goal of this work is, however, to be able to predict experimentally equivalent HOMO/LUMO values of polymers in order to increase the potential utility of this model. To do so, phase II fine-tunes the model on polymers already reported in OPVs using experimentally determined HOMO/LUMO values as "true" values (Figure S2, Supporting Information). The model is fine-tuned on 162 polymers from the HOPV15 dataset, which are expanded to 6850 SMILES with 4595 (67%) unique images (see Section S2.3, Supporting Information, for details) with 32 polymers left for testing. By starting with the weights learned in phase I, for both the convolutional network and the deep dense network, the model can leverage the predictive structure–property relationship learnt from phase I and fine-tune itself to have better predictability for the new, experimental, dataset used in phase II. Note that transferring the weights only for the convolutional network or only for the deep dense network does not provide satisfying results (Table S1, Supporting Information). Since the HCEP and HOPV15 datasets have an incomplete overlap both in atoms and molecular building blocks present in the molecules, phase II training is expected to induce fine-tuning of the weights in the convolutional layers that extract the features of the polymers. Similarly, as the HCEP uses DFT and HOPV15 uses experimental
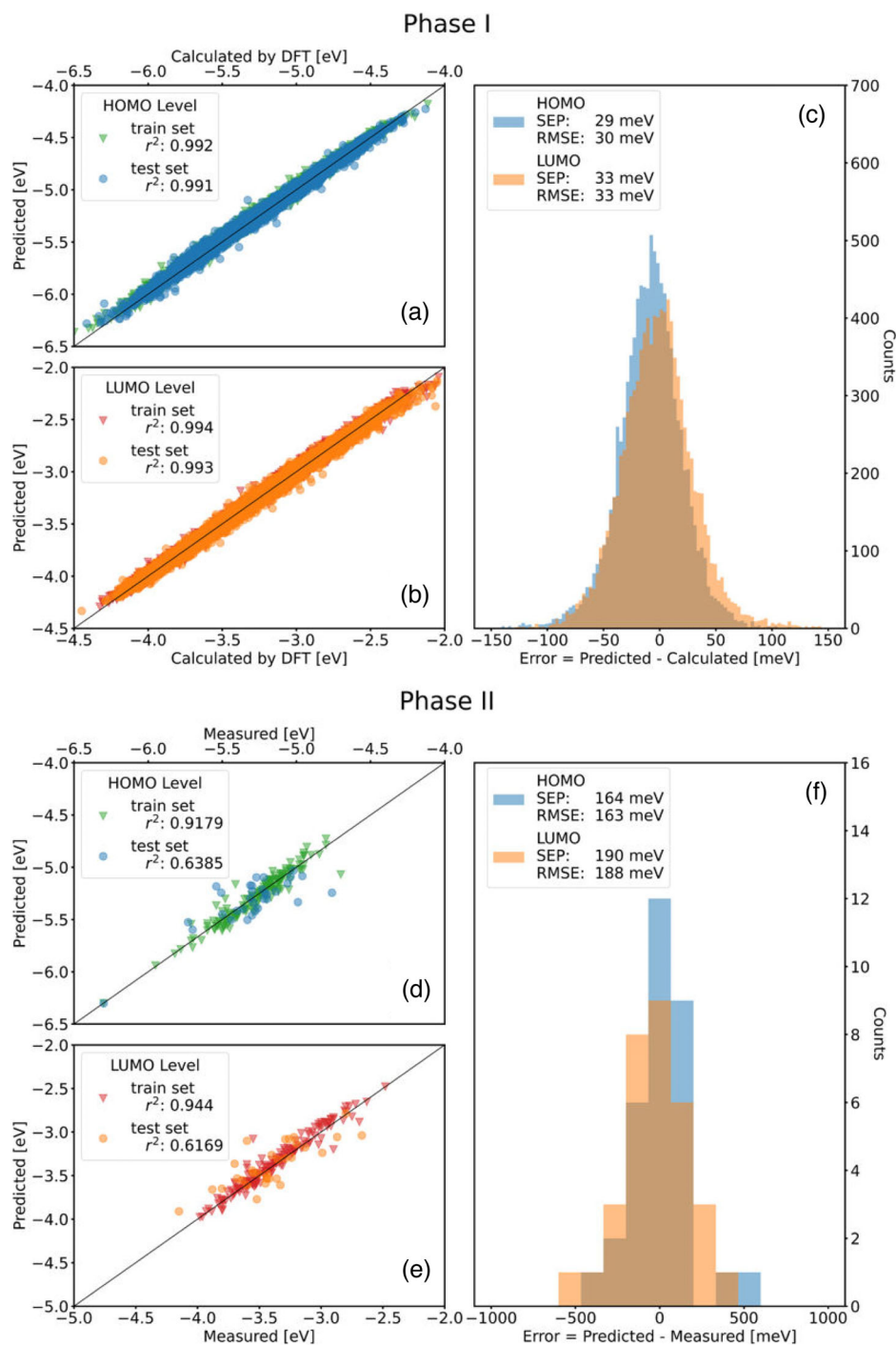
ADVANCED
SCIENCE NEWS

www.advancedsciencenews.com

ADVANCED
THEORY AND
SIMULATIONS

www.advtheorysimul.com

Figure 3. Results of both phases of deep learning model training. Phase I shows the predicted versus DFT-estimated A) HOMO and B) LUMO levels for the training set (triangles) and test set (circles). C) Distribution of the prediction errors with the HOMO level error shown in blue and the LUMO level error in orange. Phase II shows the results of the fine-tuning on HOPV15 dataset with the predicted versus experimentally measured D) HOMO and E) LUMO levels. F) Distribution of the prediction errors in phase II.

methods to define "true" values, the weights in the deep dense layers are also expected to be adjusted in phase II to accommodate the new structural features and to give predictions in accordance with experimental data. Additionally, the fine tuning of the model on HOPV15 polymers, represented by a single monomer,

with CV values as "true" forces the model to account for the decrease of the bandgaps with the average increase of the conjugation length (i.e., number of repeating units). As a result, the model learns to take into account both the shape of the monomer and the effect of polymerization on energy levels (Figure S4, Sup-

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

Table 1. Performance of the deep learning model trained on the HCEP dataset in phase I and HOPV15 in phase II including Y-scrambling for both phases and training without transfer leaning for phase II.

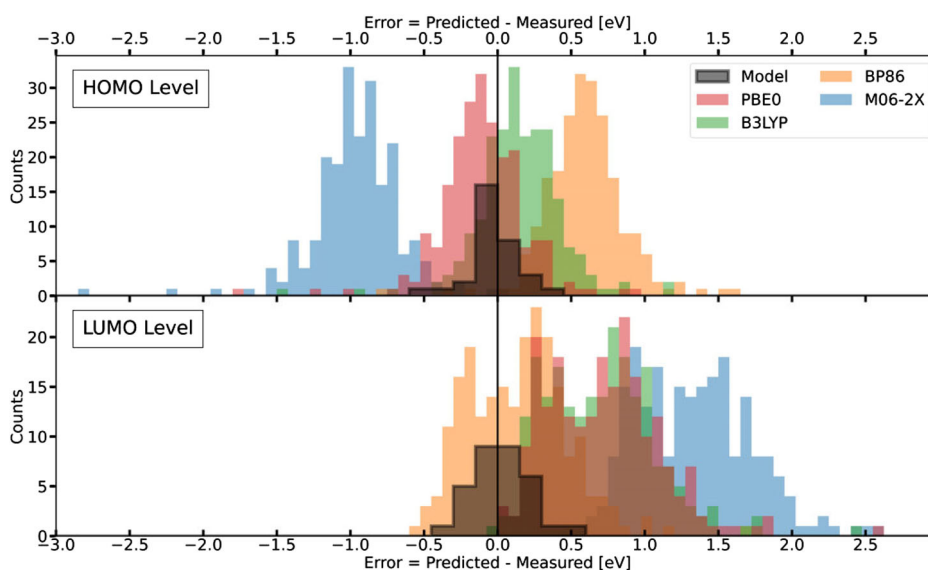|	Phase I—HCEP		Phase II—HOPV15		
	Learning	Y-scrambling	Transfer learning	Nontransfer learning	Y-scrambling
R^2 (HOMO)	0.991	7.28×10^{-6}	0.639	0.0223	7.28×10^{-4}
R^2 (LUMO)	0.993	9.18×10^{-5}	0.617	0.0366	5.86×10^{-4}
R^2 (Gap)	0.987	1.32×10^{-4}	0.477	0.0401	3.39×10^{-5}

Figure 4. Distribution of the prediction errors of the HOMO (top) and LUMO (bottom) levels for the deep learning model (black) as well as DFT-estimated values for all conformers using various functionals from the HOPV15 dataset (colors). The difference in counts results from the multiple DFT values (one for each conformer) reported for a given molecule with a given functional in the HOPV15 dataset.

porting Information), which is in sharp contrast with deterministic DFT simulations.[25] For both HOMO and LUMO predictions after phase II training (Figure 3D,E respectively), satisfying R^2 values are obtained: greater than 0.9 for the training set, and greater than 0.6 for the test set, which indicates a good correlation between the prediction and true values. More importantly, SEP and RMSE values below 170 and 190 meV are achieved for the HOMO and LUMO levels respectively (Figure 3F). In other words, with this model, the frontier energy levels of any newly designed donor polymer composed of atoms and building blocks seen in the HOPV15 training set, can be predicted with, on average, an error of less than 200 meV compared to the experimental value. Such finding offers a fast and accurate tool to guide molecular engineering for OPV optimization.

The decrease in the predictability of the model after phase II, compared to phase I, comes from the effects of the smaller dataset (despite transfer learning), polymerization effects and the measurement error in experimental CV data. Machine learned models inherently carry through the error from training data and it is impossible, therefore, not improve on its accuracy. As in phase I, the validity of the phase II results and the importance of using the transfer learning technique are confirmed by attempting to train the model without transferring the weights from the phase I model (nontransfer learning) and Y-scrambling tests. In-

deed, low R^2 values are obtained when the model is not initialized with the previously learned weights (Non-Transfer Learning, Table 1). This illustrates the necessity of the transfer learning and confirms that the HOPV15 dataset is not large enough to train this kind of deep learning model alone, even with the data augmentation techniques described above. The data augmentation step did, however, result in a model showing good molecular conformation and orientation invariance (Table S2). Lastly, the poor R^2 values from the Y-scrambling test again demonstrates the learned relationship between the new chemical structures and the experimental values of the frontier energy levels (Table 1).

3.3. Comparison with DFT from HOPV15 Dataset

After exploring the results of the phase II training against experimental data, the model is compared to the common method for energy level estimation, DFT simulations. The error distributions, i.e., the distributions of the difference between predicted (deep learning model or DFT) and experimental "true" values, are shown in **Figure 4**. The distributions of the absolute values are shown in Figure S5 (Supporting Information). For the DFT-estimated results, we see rather poor accuracy of the prediction leading to large RMSE values (**Table 2**), which are illustrated by the shift in the error distributions from zero (Figure 4). The

Table 2. Comparison of the performance of the deep learning model and the various DFT estimations for the HOMO level, LUMO level and the bandgap (= LUMO − HOMO), showing the standard error of prediction (SEP) and the root mean square error (RMSE).

Method	HOMO [meV] [SEP / RMSE]	LUMO [meV] [SEP/RMSE]	Gap [meV][SEP/RMSE]
This model	164/163	190/188	246/245
DFT: PBE0	292/309	381/814	530/978
DFT: B3LYP	288/324	370/791	515/753
DFT: BP86	269/652	339/367	449/637
DFT: M06-2X	307/1003	407/1407	589/2376

low RMSE values and the error distributions centered around zero for the deep learning model clearly indicate its better accuracy compared to this level of DFT in predicting experimental HOMO/LUMO levels. Even if the shifts in energy values obtained by DFT can conceivably be improved by an empirical correction factor, the precision remains inferior to the model as highlighted by the SEP values (Table 2). Indeed, all the DFT methods show comparable SEP values (between 269 and 407 meV) for the estimation of the HOMO and LUMO levels (Table 2), while the SEP values for the model predictions are around half of that (164 meV and 190 meV for the HOMO and LUMO levels, respectively). In particular, the better accuracy and precision of the deep learning model for both the HOMO and LUMO levels is emphasized by
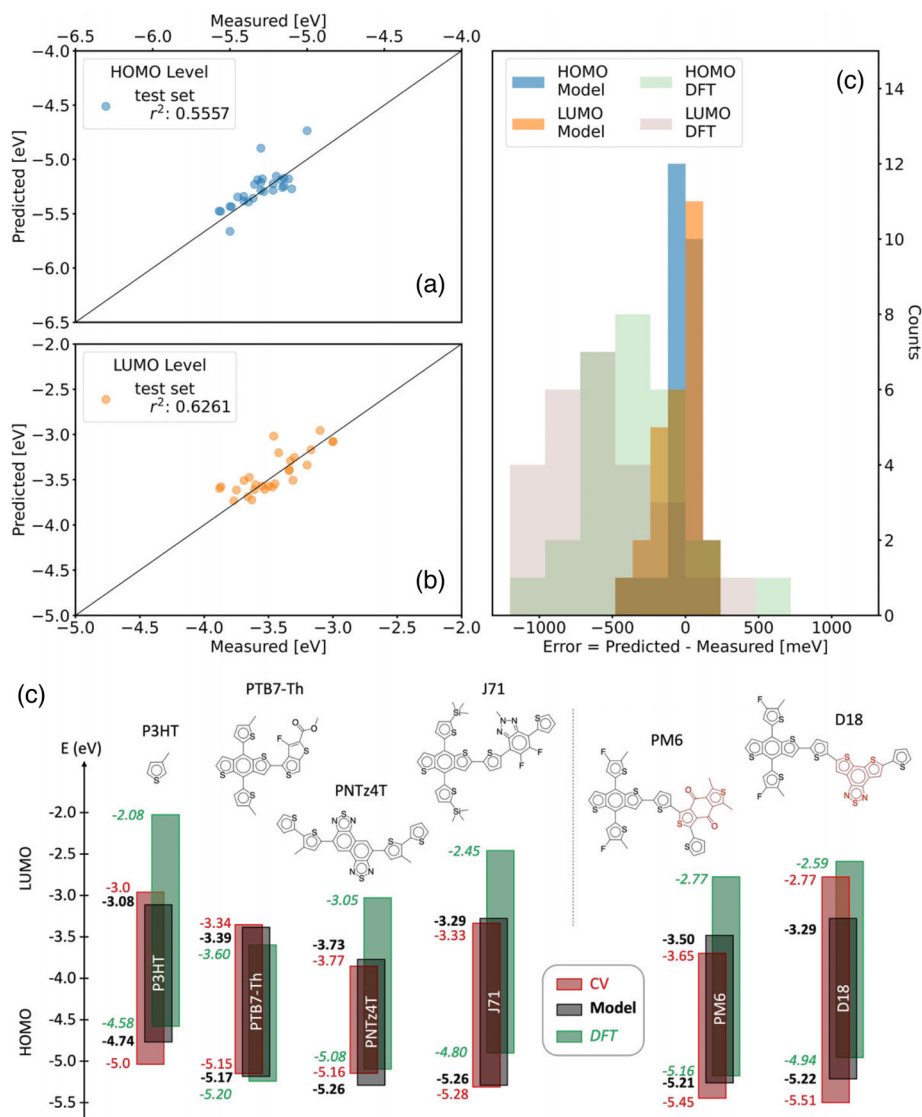

Figure 5. Results of testing the deep learning model on the "use-case" dataset. Deviation of the A) HOMO and B) LUMO level predicted compared to the experimental values. C) Distributions of the prediction errors for the values outputted by the deep learning model (Model) and the values estimated by DFT (*DFT*) extracted from literature. "Measured" correspond to energy levels determined experimentally (by cyclic voltammetry (CV)) and extracted from peer-reviewed journals. D) Schematic comparison of the model and DFT predictions compared to experimental values of commonly used donor polymers in OPVs. PM6 and D18 are not part of the "use-case" dataset as they are partly composed of building blocks, highlighted in red in their chemical structures, not represented in the HOPV15 training set.

**ADVANCED
SCIENCE NEWS**

www.advancedsciencenews.com

**ADVANCED
THEORY AND
SIMULATIONS**

www.advtheorysimul.com

Table 3. Comparison of the results of the deep learning model and published DFT of molecules from the "use-case" testing dataset. The table shows SEP and RMSE of the HOMO, LUMO, and bandgap (= LUMO − HOMO).

Method	HOMO [meV] [SEP/RMSE]	LUMO [meV] [SEP/RMSE]	Gap [meV] [SEP/RMSE]
This model	121/125	154/156	186/182
DFT	329/500	390/727	506/549

the low SEP (precision, 246 vs 449 meV) and low RMSE (accuracy, 245 vs 637 meV) obtained for the predicted bandgap compared to BP86, the best performing DFT functional for predicting the bandgap (Table 2). We can therefore conclude that the deep transfer learning model is not only more accurate but also more precise than any of the DFT methods used in the HOPV15 dataset.

3.4. External Validation: Comparison with DFT from Literature ("use-case" Dataset)

In order to validate this claim, the deep learning model is tested on the "use-case" dataset made up of 26 commercially available donor polymers whose HOMO/LUMO levels were both experimentally determined by CV measurements and calculated using DFT based on routinely used hybrid functionals (B3LYP or PBE0) and advanced basis sets (mainly 6-31G), published in peer-reviewed journals (details in Tables S3 and S4, Supporting Information). The model undergoes no further training. The predictions of both the deep learning model and DFT simulations are compared to the experimental CV data. The R^2 values of circa 0.56 for the HOMO and 0.63 for the LUMO levels (**Figure 5**A,B respectively) confirm the practical predictability of this model.[20] Again, the distribution of the prediction errors is better centered around zero than for DFT simulations (Figure 5C). More importantly, the SEP and RMSE values under 160 meV for both HOMO and LUMO model predictions compared to 390 meV for DFT (**Table 3**) demonstrate the ability of the deep learning model to outperform routinely used DFT simulations in determining the frontier energy levels of this set of polymers. Compared to machine learning models that predict DFT-B3LYP HOMO/LUMO levels with RMSE of around 200 meV, the model presented here offers state-of-the-art accuracy for predicting experimental HOMO/LUMO levels with RMSE of 125 meV and 156 meV respectively (Table 3).[8c] The list of predicted values for each molecule (including PCE11, PCDTBT, F8T2, MEH-PPV, TQ1, etc.) is given in Table S3 (Supporting Information). The performance of the deep learning model compared to DFT is illustrated graphically in Figure 5D for a few polymers of broad interest: P3HT, PTB7-Th, PNTz4T, J71, PM6, and D18 (PCE18).

The limitations of this deep transfer learning model, however, must be considered. It is mainly limited by the broadness of the training data, in that its ability to give consistently good predictions decreases when tested on polymers containing atoms or building blocks that are not represented in the training data (overlap of building blocks are shown in Figure S6, Supporting Information). As an example, D18, which has a dithienobenzothiadiazole unit not represented in the HOPV15 training set, shows a

better HOMO level prediction but an overall worse model prediction than the B3LYP-based DFT estimation. Indeed, the LUMO level is significantly under-estimated most likely because wide-bandgap polymers such as D18 are not present in the training set.[26] Note that PM6 contains a benzo[1,2-*c*:4,5-*c'*]dithiophene-4,8-dione unit also not represented in the HOPV15 training set and yet the deep learning model remains more accurate than B3LYP-based DFT.[27] The HOMO/LUMO predictions of 16 other polymers used in OPVs that are commercially available, but not fully represented in the training set, is given in Table S4 (including PCE12, PCE13, J61, DRCN5T, PDBT-T1, etc.) with statistical analysis in Figure S7 and Table S5 (Supporting Information). Finally, it is crucial to remember that, unlike the model presented here, the capabilities of DFT simulations go far beyond the simple prediction of HOMO/LUMO energy levels. DFT allows the estimation of electronic distribution of each orbital, dipole moment, electronic coupling, molecular electrostatic potential, optical absorption, and many more complex properties.[28]

4. Conclusion and Outlook

In this work, a QSPR deep transfer learning model is successfully created which takes the SMILES of a molecule as input, converts it to an RGB image, and predict its HOMO/LUMO levels with an accuracy (RMSE) of below 200 meV. This model makes use of a convolutional neural network architecture and transfer learning techniques in order to train on experimental data despite the relatively small dataset. The practical use of this model is successfully validated on real-use donor polymers used in OPVs from both the HOPV15 dataset (test set) and an external "use-case" dataset made up of commercially available polymers with frontier energy levels reported in literature. The model predictions are also compared to the results of DFT simulations, using four different functionals, and DFT results reported in literature, whereby the model is found to be substantially more accurate and precise. This suggests that this deep learning model performs better at predicting the frontier energies of this class of molecules than the computationally expensive and time consuming DFT simulations. As a result, this model offers a reliable and quick way of screening potential donor polymers for optimizing OPVs, thereby saving costly and time-consuming synthesis and experimental testing.

One downside of models created using deep learning techniques is our inability to extract how the model converts the image into the energy values. Nevertheless, our findings suggest that with enough data, parameters and a well optimized model, there is enough information in the 2D diagram of a molecule to predict its HOMO/LUMO energies. Currently the major limitation hindering the broad adoption of this model is the limited training set, as it does not process atoms and molecular building blocks that are not represented in the training data consistently. We believe that if one could gather the data already available in peer-reviewed journals and train this model on it, the model will grow in accuracy for a wide variety of newly designed materials. Considering that no chemical knowledge is implemented, we do not see any limitations in extending the use of this model to acceptor molecules.[29] In particular, the prediction of the LUMO level of accepting small molecules, such as nonfullerene acceptors, would be of high interest for the design of efficient donor-

acceptor systems for bulk heterojunction solar cells. In general, the expansion of datasets in size and diversity would greatly improve the prediction power and practical utility of these kinds of QSPR models.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that supports the findings of this study are available in the supplementary material of this article

Keywords

[1] M.-D. V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, B. Zhou, *Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge **2021**.

[2] J. Wang, Z. Zheng, Y. Zu, Y. Wang, X. Liu, S. Zhang, M. Zhang, J. Hou, *Adv. Mater.* **2021**, *33*, 2102787.

[3] a) J. Hou, O. Inganas, R. H. Friend, F. Gao, *Nat. Mater.* **2018**, *17*, 119; b) Y. Cui, H. Yao, L. Hong, T. Zhang, Y. Xu, K. Xian, B. Gao, J. Qin, J. Zhang, Z. Wei, J. Hou, *Adv. Mater.* **2019**, *31*, 1808356; c) B. Fan, D. Zhang, M. Li, W. Zhong, Z. Zeng, L. Ying, F. Huang, Y. Cao, *Sci. China Chem.* **2019**, *62*, 746; d) Y. Cui, H. Yao, J. Zhang, T. Zhang, Y. Wang, L. Hong, K. Xian, B. Xu, S. Zhang, J. Peng, Z. Wei, F. Gao, J. Hou, *Nat. Commun.* **2019**, *10*, 2515; e) Q. Liu, Y. Jiang, K. Jin, J. Qin, J. Xu, W. Li, J. Xiong, J. Liu, Z. Xiao, K. Sun, S. Yang, X. Zhang, L. Ding, *Sci. Bull.* **2020**, *64*, 272.

[4] a) M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, *Adv. Mater.* **2006**, *18*, 789; b) F. H. Alharbi, S. N. Rashkeev, F. El-Mellouhi, H. P. Lüthi, N. Tabet, S. Kais, *npj Comput. Mater.* **2015**, *1*, 15003; c) M. C. Scharber, N. S. Sariciftci, *Prog. Polym. Sci.* **2013**, *38*, 1929.

[5] a) P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864; b) W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133; c) M. Mesta, J. H. Chang, S. Shil, K. S. Thygesen, J. M. G. Lastra, *J. Phys. Chem. A* **2019**, *123*, 4980.

[6] a) B. G. Janesko, *Chem. Soc. Rev.* **2021**, *50*, 8470; b) P. Verma, D. G. Truhlar, *Trends Chem.* **2020**, *2*, 302.

[7] a) K. He, X. Zhang, S. Ren, J. Sun, Springer, Cham **2016**, 9908; b) K. Simonyan, A. Zisserman, *arXiv preprint arXiv* **2014**, *1409*.

[8] a) V. Yangali-Quintanilla, A. Verliefde, T. U. Kim, A. Sadmani, M. Kennedy, G. Amy, *J. Membr. Sci.* **2009**, *342*, 251; b) R. Olivares-Amaya, C. Amador-Bedolla, J. Hachmann, S. Atahan-Evrenk, R. S. Sánchez-Carrera, L. Vogt, A. Aspuru-Guzik, *Energy Environ. Sci.* **2011**, *4*, 4849; c) F. Pereira, K. Xiao, D. A. R. S. Latino, C. Wu, Q. Zhang, J. Aires-De-Sousa, *J. Chem. Inf. Model.* **2017**, *57*, 11; d) H. Sahu, W. Rao, A. Troisi, H. Ma, *Adv. Energy Mater.* **2018**, *8*, 1801032; e) D. Padula, J. D. Simpson, A. Troisi, *Mater. Horiz.* **2019**, *6*, 343; f) W. Sun, M. Li, Y. Li, Z. Wu, Y. Sun, S. Lu, Z. Xiao, B. Zhao, K. Sun, *Adv. Theory Simul.* **2019**, *2*, 1800116; g) N. Meftahi, M. Klymenko, A. J. Christofferson, U. Bach, D. A. Winkler, S. P. Russo, *npj Comput. Mater.* **2020**, *6*, 166. h) A. Mahmood, J.-L. Wang, *Energy Environ. Sci.* **2021**, *14*, 90.

[9] A. Paul, D. Jha, R. Al-Bahrani, W.-K. Liao, A. Choudhary, A. Agrawal, 2019 International Joint Conference on Neural Networks (IJCNN), **2019**, p. 1.

[10] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, K. Sun, *Sci. Adv.* **2019**, *5*, eaay4275.

[11] a) C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, Springer, Cham, **2018**, p. 270; b) M. Hussain, J. J. Bird, D. R. Faria, Springer, Cham, **2018**, p. 191.

[12] a) C. Farabet, C. Couprie, L. Najman, Y. LeCun, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915; b) C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, *arXiv pre-print server* **2014**.

[13] a) J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, *J. Chem. Inf. Model.* **2015**, *55*, 263; b) T. Ciodaro, D. Deva, J. M. de Seixas, D. Damazio, *J. Phys.: Conf. Ser.* **2012**, *368*, 012030; c) M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk, *Nature* **2013**, *500*, 168; d) M. K. K. Leung, H. Y. Xiong, L. J. Lee, B. J. Frey, *Bioinformatics* **2014**, *30*, i121; e) H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, T. R. Hughes, Q. Morris, Y. Barash, A. R. Krainer, N. Jojic, S. W. Scherer, B. J. Blencowe, B. J. Frey, *Science* **2015**, *347*, 1254806.

[14] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.

[15] G. Vrbančič, V. Podgorelec, *IEEE Access* **2020**, *8*, 196197.

[16] J. Y. Ching, A. K. C. Wong, K. C. C. Chan, *IEEE Trans. Pattern Anal. Mach. Intell.* **1995**, *17*, 641.

[17] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241.

[18] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Hachmann, A. Aspuru-Guzik, *Sci. Data* **2016**, *3*, 160086.

[19] D. L. J. Alexander, A. Tropsha, D. A. Winkler, *J. Chem. Inf. Model.* **2015**, *55*, 1316.

[20] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69.

[21] C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **2007**, *47*, 2345.

[22] N. Elgrishi, K. J. Rountree, B. D. McCarthy, E. S. Rountree, T. T. Eisenhart, J. L. Dempsey, *J. Chem. Educ.* **2018**, *95*, 197.

[23] C. M. Cardona, W. Li, A. E. Kaifer, D. Stockdale, G. C. Bazan, *Adv. Mater.* **2011**, *23*, 2367.

[24] J. Sworakowski, *Synth. Met.* **2018**, *235*, 125.

[25] S. S. Zade, M. Bendikov, *Org. Lett.* **2006**, *8*, 5243.

[26] a) Q. Liu, Y. Jiang, K. Jin, J. Qin, J. Xu, W. Li, J. Xiong, J. Liu, Z. Xiao, K. Sun, S. Yang, X. Zhang, L. Ding, *Sci. Bull.* **2020**, *65*, 272; b) X. Li, J. Xu, Z. Xiao, X. Wang, B. Zhang, L. Ding, *J. Semicond.* **2021**, *42*, 060501.

[27] a) H. Zhang, H. Yao, J. Hou, J. Zhu, J. Zhang, W. Li, R. Yu, B. Gao, S. Zhang, J. Hou, *Adv. Mater.* **2018**, *30*, 1800613; b) W. Liu, Q. Liu, C. Xiang, H. Zhou, L. Jiang, Y. Zou, *Surf. Interfaces* **2021**, *26*, 101385.

[28] a) A. Aboulouard, S. Mtougui, N. Demir, A. Moubarik, M. E. Idrissi, M. Can, *Synth. Met.* **2021**, *279*, 116846; b) U. Yaqoob, A. Raza Ayub, S. Rafiq, M. Khalid, J. Iqbal, *J. Mol. Liq.* **2021**, *341*, 117428; c) A. Guillén-López, C. Delesma, C. Amador-Bedolla, M. Robles, J. Muñiz, *Theor. Chem. Acc.* **2018**, *137*, 85; d) D. Rana, A. Materny, *Spectrochim. Acta, Part A* **2021**, *253*, 119565; e) B. G. Sumpter, V. Meunier, *J. Polym. Sci., Part B: Polym. Phys.* **2012**, *50*, 1071.

[29] S. A. Lopez, B. Sanchez-Lengeling, J. De Goes Soares, A. Aspuru-Guzik, *Joule* **2017**, *1*, 857.

2100511 (10 of 10)