1. Introduction, Motivation, State of the Art

Organic semiconductors are promising materials for cheap, scalable, and sustainable electronics, light-emitting diodes, and photovoltaics. For organic photovoltaic cells, it is a challenge to find compounds with suitable properties in the vast chemical compound space. For example, the ionization energy should fit to the optical spectrum of sunlight, and the energy levels must allow efficient charge transport. Herein, a machine learning model is developed for rapidly and accurately estimating the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies of a given molecular structure. It is built upon the SchNet model [Schütt et al. (2018)] and augmented with a “Set2Set” readout module [Vinyals et al. (2016)]. The Set2Set module has more expressive power than sum and average aggregation and is more suitable for the complex quantities under consideration. Most previous models are trained and evaluated on rather small molecules. Therefore, the second contribution is extending the scope of machine learning methods by adding also larger molecules from other sources and establishing a consistent train/validation/test split. As a third contribution, a multitask ansatz is made to resolve the problem of different sources coming at different levels of theory. All three contributions in conjunction bring the accuracy of the model close to chemical accuracy.

1.1. Organic Photovoltaics

Brief overview over organic photovoltaics: The standard setup for organic photovoltaic cells is the bulk heterojunction, that is, a bicontinuous network of the donor and acceptor materials, as introduced by Yu et al. in 1995, with a semiconducting polymer as donor and C60 (or a derivative) as acceptor. After photoexcitation, an excited electron can move to the neighboring molecule if there is an unoccupied orbital available at a suitable energy. This charge separation is the principle of OPV. A certain energy difference is needed to overcome the binding energy of the excited electron and the hole it leaves behind. Thus, for high efficiency, the energy alignment of the molecular orbitals is crucial.

More elaborate OPV cells based on high-performance polymer donors and fullerene-based acceptors have reached power conversion efficiencies (PCEs) up to 11.7%. More recently, the focus has shifted to nonfullerene acceptors (NFAs) which have reached PCEs over 19% PCE already.

1.1.1. Materials Discovery for OPV

The power conversion rates of OPVs are still lower than those of their silicon counterparts, but continuous progress, in particular by discovering better organic materials, has been made in recent years. In the search for new and better materials, a series of requirements has to be met. For example, the absorption spectrum should match with the solar spectrum ($E_{\text{LUMO}} - E_{\text{HOMO}} = 2$ eV), accompanied with a high extinction coefficient. The energy levels should be suitably aligned for the intended electrodes and other interface materials (e.g., at the p–n junction). The device structure should be stable under various environments, such as humidity, heat, light, and oxygen, and in particular with respect to diffusion (i.e., the molecules should not be too small). Finally, the materials should be environmentally friendly and nontoxic, and simple fabrication processes, for example, vapor deposition (molecules not too large), should be possible. The dependence on scarce or expensive raw materials is to be avoided, as far as possible. Many of these required properties are accessible by ab initio calculations, that is, the candidate materials can be simulated before starting experiments in the lab.

1.1.2. Machine Learning in Materials Discovery for OPV

Chemical compound space is huge; therefore, it is prohibitively expensive to determine the desired properties of each compound.
with ab initio methods. This is where machine learning (ML) steps in. ML is a data-driven approach that promises to speed this up, for example, by substituting expensive density functional theory (DFT) calculations by statistical models that estimate chemical properties at a fraction of the numerical cost. Figure 1 shows a scheme where ML algorithm is used as a filter, and only those candidates that pass the ML filter are processed in the subsequent steps (ab initio calculations and, when applicable, synthesis and characterization in the laboratory).

1.2. Machine Learning for Quantum Chemistry

1.2.1. Bridging the Gap Between Molecular Properties and the Performance of OPV Device

A number of efforts has been made to relate the performance of an OPV device to the elemental properties of the materials. The Scharber model\(^9\) relates the PCE to the donor bandgap (under some assumptions on the acceptor and electrode materials). Padula et al.\(^10,11\) tried to go beyond that with ML model taking both electronic and structural features as input.

1.2.2. Goal and Context of Present Work

In the present work, we take an intermediate step and develop an ML model to reliably estimate the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies of a given molecular structure. The wider context in a materials discovery scheme is represented in Figure 1. The task of the ML filter is to provide estimates of the molecular properties (currently the HOMO and LUMO energies) in order to identify the “good” candidates from a pool of molecule. Beyond the scope of this article, the good candidates would be further investigated, first by means of ab initio (usually DFT) calculations, where again, some candidates will be discarded and others pass to the next stage, experiments in the lab, eventually leading to a better product. Along the way, the knowledge obtained from negative results should not be thrown away, but rather stored in a database in order to be used as training data for future development of the ML algorithm. Another aspect is applying a genetic algorithm and generating new candidate molecules by means of crossover and mutation of good candidates and feeding them back into the pool of molecules.

1.2.3. Short Review of Algorithms for the Prediction of Molecular Properties

For this problem, a statistical model is needed that takes molecular structures, that is, collections of atoms and their relative positions, as input. The model has to handle arbitrary numbers of atoms and should be invariant under rotation, translation, and the indexing order of the atoms. Therefore applying standard ML models (like, e.g., a support vector machine or a convolutional neural network) is only possible if some kind of preprocessing of the input structures (e.g., extended-connectivity fingerprints\(^12\)) is applied. This is not ideal because molecular fingerprints are hand crafted and not invertible. Duvenaud et al.\(^13\) as well as Kearnes et al.\(^14\) integrated the fingerprint feature extraction into the deep neural network, which resulted in graph convolutional networks that operate on vertices (atoms) and edges (bonds) of a molecular graph. This was developed further by Schütz et al.\(^15\) who added continuous-filter convolutions and Chen et al. who developed MEGNet.\(^16\)

Refs. [17,18] are benchmarks of state-of-the-art models, measured on the QM9 dataset, and have shown that accuracies better than the so-called chemical accuracy (1 kcal mol\(^{-1}\) = 0.043 eV) can be reached with models like enn-s2s, SchNet, and SchNetE on a dataset of small molecules (for details on QM9, see Section 2.1).

For real-world applications of materials discovery for organic photovoltaics, one has to go beyond the scope of the QM9 dataset. The contribution of this article is developing a relevant data corpus (Section 2) and a suitable architecture for predicting the HOMO and LUMO energies of large organic molecules (Section 3) and finally the empirical validation (Section 4).

2. Data Corpus

2.1. Data Sources

Several datasets containing significant numbers of molecular structures and properties are already publicly available.\(^19–23\)
These sets consist of pairs \((x, y)\), where \(x\) is the point cloud of the atoms in space, that is, a set of quadruples representing the coordinates and the atomic charge of each atom, and \(y\) is a set of scalar properties. The geometries and properties are the outcomes of DFT calculations at a given level of theory. The properties \(y\) include, in particular, the energies of the molecular frontier orbitals HOMO and LUMO, which are proxies for the ionization energy and the electron affinity of the respective molecule.

The QM9 dataset\([19]\) contains (presumably) all molecules of up to 9 “heavy” atoms (C, N, O, F) plus hydrogens, which are 130k structures. The data was obtained at the B3LYP/6-31G(2df,p) level of theory. This dataset is commonly used in many theory papers.\([15–18]\)

In 2019, Chen et al. released a dataset of 200k molecules of up to 12 heavy atoms (C, N, O, F, S and Cl), within the scope of the “Alchemy” contest.\([20]\) The level of theory and the set of properties are the same as QM9, but the dataset is extended to slightly larger molecules and larger variety of chemical elements.

Stuke et al. released the “OE62” dataset containing 62k structures, including larger ones.\([21]\) The initial structures were extracted from organic crystals in the Cambridge Structural Database (CSD) and the molecular geometry and properties were determined via DFT. The chemical composition is much wider than those of the other sources, ranging from hydrogen to iodine (H, Li, B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Re, I).

The HOPV dataset contains 350 different molecules (5k conformers), which are large and relevant for organic photovoltaics.\([22]\) The HOPV set contains the elements H, C, N, O, F, Si, S, Se.

Kuzmich et al. published a set of 80 structures that are relevant for organic photovoltaics.\([23]\) Two molecules were discarded because of ambiguous labeling. Since this set is small, we will use it for testing only. The variety of chemical elements in the Kuzmich set is the same as HOPV plus B and Cl, which are only in OE62.

Table 1 summarizes the properties of the different data sources, and Figure 2b displays the distribution of the properties HOMO and LUMO over the molecular sizes present in the datasets. While the distribution of sizes differs strongly between the four sources, the distribution of the target values varies less. In the following, we will focus on the datasets marked in bold face in Table 1 and group them into 1) “B3LYP” exchange-correlation functional with a reasonably large basis set (QM9, Alchemy, HOPV) and 2) “PBE0” exchange-correlation functional with a reasonably large basis set (OE62, HOPV). Ideally, a sufficiently “good” level of theory will produce data close to the real values. But in practice, there are systematic shifts between data obtained from different levels of theory (Figure 2a). This problem will be addressed with a model that makes separate predictions for B3LYP and PBE0, see Section 3.1.

2.2. Train/Validation/Test Split

In ML, the available data is commonly split into three parts, 1) the training data for fitting the free parameters of the model, 2) the validation data used to make decisions, for example, when to stop the training process, and 3) the test set to finally evaluate the accuracy of the model. For a fair evaluation, it is important that the three subsets are disjoint, in particular, the test shall represent the ability of the model to generalize data that was not “seen” in the training. Since the same structures may occur in several data sources, care has to be taken that the train/validation/test split is consistent across all sources. For example, the molecule

| Dataset          | \(N_{\text{items}}\) | \(N_{\text{InChIs}}\) | \(N_{\text{median}}\) | \(N_{\text{max}}\) | Level of theory                 | Geometry                  | properties\(^{(5)}\) |
|------------------|----------------------|----------------------|----------------------|----------------------|-------------------------------|---------------------------|----------------------|
| QM9\([19]\)      | 133 879              | 129 434              | 18                   | 29                   | B3LYP/6-31G(2df,p)            | B3LYP/6-31G(2df,p)        |
| Alchemy\([20]\)  | 202 579              | 201 542              | 22                   | 38                   | B3LYP/6-31G(2df,p)            | B3LYP/6-31G(2df,p)        |
| OE62\([21]\)     | 61 489               | 61 489               | 39                   | 174                  | PBE + vdW/Tier2\(\text{tight}\)\(^{(4)}\) | PBE/Tier2\(\text{tight}\) |
|                  | 31 k                 | 31 k                 | 38                   | 116                  | PBE + vdW/Tier2\(\text{tight}\)\(^{(4)}\) | PBE0/Tier2\(\text{tight}\) |
|                  | 5 k                  | 5 k                  | 38                   | 100                  | PBE + vdW/Tier2\(\text{tight}\)\(^{(4)}\) | PBE0/def2-TZVP            |
|                  |                      |                      |                      |                      | BP86/def2-SVP                 |                            |
| HOPV\([22]\)     | 4855                 | 350                  | 69                   | 142                  | BP86/def2-SVP                 |                            |
| Kuzmich2017\([23]\) | 80                  | 79\(^{(5)}\)         | 123                  | 270                  | 6-31G\(^{\ast}\)/B3LYP       |                            |

\(^{(4)}\) \(N_{\text{InChIs}}\) is the number of distinct InChI codes, i.e., distinct molecular graphs. \(^{(5)}\) \(N_{\text{median}}\) and \(N_{\text{max}}\) are the median respectively the max number of atoms per molecule. \(^{(6)}\) Items marked in bold are the ones considered in this work. \(^{(7)}\) PBE0 level of DFT with TS-vdW corrections. \(^{(8)}\) It denotes “Tier 2 basis” set in “tight setting”: numeric atom-centered orbitals as implemented in the FHI-aims package. \(^{(9)}\) The hybrid PBE0 exchange correlation functional of Refs. \([42,43]\). \(^{(10)}\) Multipole Expansion (MPE) implicit solvation method for water. \(^{(11)}\) Only the molecules with the identifiers FPDI-T and T2 have the same InChI. They differ only by single-bond rotations and have very similar properties.
4C6H3NO2 with the molecular graph as shown in Figure 2c occurs both in QM9 and in Alchemy, like some 18 000 other molecules.

The international chemical identifier (InChI) is a unique identifier of a molecule with a layered structure from coarse to fine information. This opens a convenient way of grouping similar structures together, namely, structures that coincide in the first InChI layers. For example, if only the first two sublayers of standard InChI codes are used, the chemical formula and the backbone structure are the same, but the locations of hydrogens, charge information and the stereochemical information, as well as isotopic information may differ. The train/validation/test split used in this work is constructed as follows. 1) Perform the split in terms of InChIs truncated after the second layer (molecules that differ only in later layers will be assigned to the same part of the split by construction). 2) Assign the test set of the Alchemy contest to test. 3) Assign the original Alchemy validation set to validation, except for those structures whose truncated InChIs are already in the test set (2% of the truncated InChIs), which go to test. 4) Assign the rest of the data such that 18% and 9% of the truncated InChIs are assigned to test and validation, respectively. The rest is assigned to train. 5) Subsequently assign the truncated InChIs of QM9, OE62, and HOPV in the same manner: Truncated InChIs that have occurred previously are assigned as before. The rest is filled up such that test and validation fraction are 18% and 9%, respectively.

Table 2 shows the sizes of the different data sources and their intersections in terms of truncated InChIs. Table 3 collects the sizes of the train, validation, and test parts resulting from the described procedure.

### 3. Model Architecture

The ML model does not question the training data; it just learns to imitate it. As mentioned above (Figure 2a), there may be systematic offsets between data obtained at the one or the other level.

### Table 2. Intersections of datasets in terms of truncated InChIs (sum formula and nonhydrogen connectivity). The numbers on the diagonal are the counts of distinct truncated InChIs in each data source.

|       | QM9      | Alchemy  | OE62     | HOPV     |
|-------|----------|----------|----------|----------|
| QM9   | 114 707  | 18 575   | 340      | 0        |
| Alchemy | 18 575  | 174 894  | 82       | 0        |
| OE62  | 340     | 82       | 61 398   | 3        |
| HOPV  | 0       | 0        | 3        | 347      |
of theory. Our approach to deal with these systematic shifts between different data sources is that the ML model makes separate predictions for the B3LYP values and the PBE0 values of the frontier orbital energies.

3.1. Multitask Approach

The idea of multitask learning\cite{25} is to train multiple target values in the same model. After a common part, the model splits into different branches, one for each task. Here, the tasks are the B3LYP estimates and the PBE0 estimates. Figure 3 depicts two example architectures with a single input and multiple outputs (more details are given in Sections 3.2 and 3.3).

A given item from the training set may or may not have target values for all tasks. For those tasks where targets are available, errors are computed in the loss function and backpropagated through the network. Notably all tasks contribute to the training of the common part of network and help to learn a meaningful representation of the input. Since the tasks are correlated, it is expected that they benefit from each other.

3.2. Molecular Graph Processing

The inputs to the model are point clouds of atoms, which are readily interpretable as graphs, where the atoms are the vertices and the pairwise distances are the edges. For the backbone of our model, we use the SchNet architecture by Schütt et al., which was designed to model atomistic systems and make use of continuous-filter convolutional layers, that is, learnt interactions that respect the symmetries of the underlying physics.\cite{15} In a nutshell, SchNet maps each atomic species to a (learnable) $n$-dimensional (here: $n = 128$) feature vector, which is then processed by several (here: 6) learnt interaction modules, each of them incorporating a continuous-filter generator.

3.2.1. The Filter Generator

For atoms $i,j$, the (continuous) distance $r = \| \mathbf{r}_i - \mathbf{r}_j \|$ is Gaussian expanded, $e_k = \exp(-\gamma (r - \mu_k)^2)$, for equidistant $\mu_k$ and $k = 0 \ldots 24$. This expansion passes through two dense layers with shifted softplus nonlinearities, yielding $W^l_{ij} \in \mathbb{R}^n$ for the $l$-th filter generator.

3.2.2. The Interaction Module

The atom features are first transformed atom-wisely (matrix multiplication plus adding a bias term). Then, the continuous-filter operation lets the atoms interact with the other atoms, weighted with the filter, $\sum_{j=0}^{n_{atm}} \chi^*_j \cdot W^l_{ij}$, where $\ast$ represents the element-wise multiplication. That is, the interaction module is factorized: the interactions are channelwise, while channels are mixed in the atom-wise operations. Then, there are three more layers inside the interaction block, atom-wise, soft-plus nonlinearity, and another atom-wise layer. Finally, there is a skip connection, that is, the sum of the outcome of the described interaction chain and its input is returned.

3.3. Aggregation and Readout

At some point, the representation on a molecular graph of arbitrary size (feature vectors per atom) has to be reduced to the set of output estimates, which are per molecule.

3.3.1. Sum and Average Aggregation

The simplest way of aggregating would be to take the sum (or the average). Sometimes this is the natural choice, for example, for the energy of a molecule, which is an extensive quantity, represented as the sum of the contributions from all atoms in their local environment. In fact, this is how the energy is constructed as the sum of local contributions in Ref. [15]: The (high-dimensional) atomic representation is reduced to one output by a small neural network. Then, these contributions from the atomic environments

![Figure 3. Multitask network architectures. The multi-layer perceptrons (MLP) are small neural networks](image-url)
are summed up (or averaged for intensive quantities), which gives the final result, see Figure 3a.

### 3.3.2. Set2Set Aggregation

The scaling of the present quantities of interest, the HOMO and LUMO energies, with the system size, is not clear a priori. Both are related to the level spacing and tend to decrease slowly with the system size, but neither sum nor average is the obvious choice.

Therefore, we resort to the Set2Set module that was developed by Vinyals et al.\[^{[26]}\] to learn a fixed-length representation for input sets of arbitrary size. It is based on the repeated application of a long short-term memory (LSTM) module: 1) The initial LSTM input and hidden (cell) state are zeros. 2) Project the atom features \(x\) (one per atom) on the LSTM output \(q\) (one per molecule) and apply softmax, weighted average \(r\) of the atom features \(x\). 3) Then, feed \((q, r)\) into the LSTM again. 4) After \(N\) repetitions (here, \(N = 3\)), return the vector \((q, r)\), one for each molecule.

Thus, the Set2Set module learns to generate a weighted average of the atom features. It has more expressive power than sum or average and, at the same time, respects the invariance with respect to the order of the input features.

The Set2Set aggregation works on the atomic embedding and doubles its dimensions. Afterward, a small neural network (here: two affine layers with a nonlinearity in-between) reduces the dimensions to the output dimension. Figure 3b represents the resulting network architecture.

### 4. Numerical Experiments and Results

#### 4.1. Impact of Graph Aggregation

To start with, we train SchNet models with different aggregation and output modules on the training part of the QM9 source, that is, only on rather small molecules. The accuracy of the trained algorithm is then evaluated on the test parts of the data sources, that is, on data that is disjoint from the training data, see Figure 4.

**4.1.1. Sum**

For each of the targets (HOMO, LUMO, gap), atomic contributions are obtained with a two-layer neural network and summed up. (For training details and parameters see Section 6). Results, in particular the measured mean absolute errors (MAEs), are shown in column (a) of Figure 4. The model achieves good accuracy on test data from the same distribution as the training (QM9), but is completely off for out-of-distribution data, like the large molecules from HOPV and Kuzmich (2017).

**4.1.2. Average**

Results for average aggregation (with otherwise the same parameters) are given in panel (b) of Figure 4. Average aggregation is not as far off as sum aggregation.

![Figure 4](https://www.advancedsciencenews.com/)

**Figure 4.** Regression accuracy of SchNet models with different output modules [a) Sum, b) Average, and c) Set2Set], trained only on the QM9 part of the training data. Shown are the target-estimate pairs of 25 molecules randomly sampled from each test set. The MAE measured on the entire test set is given in the legend.
4.1.3. Set2Set

The set2set output module aggregates in a flexible and learnt manner, the dimension is reduced after aggregation. This yields even better results (panel (c) of Figure 4) than average aggregation. Of course systematic deviations of the order of 1 eV are still not acceptable, but Set2Set is the best starting point for the next stage, extending the training set.

4.2. Impact of the Training Data

We successively extend the training set by adding sets of larger molecules (QM9 → Alchemy → OE62 → HOPV), Figure 5a–d. This significantly reduces the regression error on the test sets of large molecules (HOPV and Kuzmich2017).

From panels (e) to (d), we go in the opposite direction. Whereas (e) is trained without the small molecules, the errors on the QM9 and Alchemy test sets are large. Adding the datasets of small molecules (d) improves this without harming the accuracy on the large molecules (only HOPV LUMO is a bit worse in (d); HOPV HOMO, OE62, and Kuzmich LUMO are better).

The accuracy of the main algorithm (Set2Set output module for both tasks, full training set is represented in see Figure 5d, and also Table 4). We note that for all quantities and all tests the RMSE value is larger than \(\sqrt{2/\pi} \times \text{MAE}\), indicating that the error distributions are less bounded than the Gaussian distribution. We have verified that the error distributions are still approximately exponentially bounded and not dominated by outliers.

Altogether, test errors are of the magnitude 0.05 eV for HOMO and LUMO for small molecules (QM9, Alchemy) and 0.1 eV for larger molecules with more chemical variety (OE62, HOPV, Kuzmich2017). The increase of the errors with the molecule size is plausible since the subspace of small molecules is rather well covered in the training data by QM9 and Alchemy, while the space.
of larger molecules (say from \( C_{25} \) to \( C_{100} \)) is larger and is covered only by the OE62 and HOPV datasets, which contain less molecules than QM9 and Alchemy. Molecules of up to 270 atoms as they occur in the Kuzmich2017 test are not in the training data at all. We stress that the errors of our main algorithm are already close to the so-called “chemical accuracy” (1 kcal mol\(^{-1} = 0.043\) eV) and should be good for practical applications.

### 4.3. Ablation Study

Three efforts have lead to the accuracy achieved in the full model (Figure 5d): the concept of multitask training, the choice of the output module, and the extended training set. Previously, in Section 4.2, we have analyzed the impact of the training set, given the design decisions “multitask training” and “set2set aggregation”. In this section, we verify that these design decisions were actually important to achieve the accuracy reported earlier.

#### 4.3.1. Multitask Learning

In Table 4, the column “Main Algorithm” is the multitask model, that is, B3LYP and PBE0 tasks trained a set2set output module, on top of the representation learnt by a common SchNet graph-processing network, as represented in Figure 3b (with \( N = 2 \times 3 \) predictions because 3 outputs (HOMO, LUMO and gap) are trained for two tasks, B3LYP and PBE0). In the other columns, B3LYP task and PBE0 task are trained separately (Figure 3b with \( N = 3 \) predictions).

For most quantities, the multitask model achieves better accuracy. The only exception are the HOPV-PBE0 errors, which are

| Quantity | Test          | Main Algorithm | SchNet(6) + Average | SchNet(6) + Sum  |
|----------|---------------|----------------|---------------------|-----------------|
| HOMO (B3LYP) | MQ9           | 0.006 0.043 0.063 | 0.004 0.043 0.064 | 0.008 0.049 0.079 |
|           | Alchemy       | 0.001 0.047 0.077 | 0.000 0.048 0.076 | 0.004 0.055 0.085 |
|           | HOPV          | 0.004 0.057 0.090 | 0.006 0.071 0.108 | 0.070 0.222 0.368 |
| LUMO (B3LYP) | MQ9           | 0.003 0.042 0.062 | 0.001 0.041 0.059 | 0.010 0.046 0.068 |
|           | Alchemy       | 0.003 0.048 0.086 | -0.001 0.048 0.087 | 0.010 0.052 0.088 |
|           | HOPV          | -0.047 0.066 0.088 | -0.069 0.145 0.262 | -0.087 0.167 0.325 |
| Kuzmich2017 | MQ9           | 0.010 0.179 0.162 | 0.107 0.199 0.240 | -1.572 1.697 2.363 |
| HOMO (PBE0) | OE62          | 0.001 0.092 0.135 | 0.006 0.109 0.135 | -0.011 0.152 0.253 |
|           | HOPV          | -0.001 0.056 0.081 | 0.006 0.051 0.073 | -0.063 0.185 0.274 |
| LUMO (PBE0) | OE62          | 0.009 0.094 0.142 | 0.009 0.109 0.158 | 0.005 0.117 0.169 |
|           | HOPV          | -0.047 0.069 0.095 | -0.050 0.143 0.258 | -0.102 0.154 0.318 |
lower in the model trained with PBE0. But on the other hand, those quantities that are hard in the single-task setting (error measure larger than 0.1 eV, marked in red in the columns “B3LYP only” and “PBE0 only”), the errors reduce significantly when the tasks are trained together (column “Main Algorithm”). We conclude that the concept of multitask-training is indeed important for the accuracy and reliability of the model.

4.3.2. Aggregation Modules

We have shown above (Section 4.1) that the Set2Set output module is best if the training represents only a small part of the test distribution (QM9 only). Here we repeat this investigation with the full training set. Table 5 collects the results. The columns labeled “Main Algorithm” is the model SchNet(6) + Set2Set output module.

Average Aggregation: The model with the average output modules is competitive on small molecules (QM9 and Alchemy), which are represented in the training set. It also performs slightly better on the HOMO(PBE0) task on the HOPV test. However, it performs much worse on those four tests and quantities that are already difficult for the baseline model Set2Set (error measure above 0.1 eV, marked in red in the table). Altogether, the average output module has 12 error measures above 0.1 eV.

Sum Aggregation: When moving on from average to sum, the picture is similar: Sum performs worse than average and set2set on those tests and quantities that are already difficult for average and set2set. In total, the sum output module yields 16 error measures (out of 33 that are above 0.1 eV. The estimates of the LUMO values for the Kuzmich (2017) data are systematically shifted: The mean of the ground-truth LUMO values is $-3.30 \, \text{eV}$, but the mean estimated LUMO is $-4.86 \, \text{eV}$, manifesting unphysical scaling of the estimate with the system size.

In conclusion, out of the three output modules considered, the Set2Set output module remains the most reliable one, especially on challenging data.

5. Conclusion

5.1. Achievements

We have achieved a model to predict the energies of frontier orbitals HOMO and LUMO. It is reliable for large molecules, which are relevant for organic electronics and photovoltaics.

In practice, it is not enough for a model to be good on a particular subspace of chemical compound space, like, for example, the small molecules of the QM9 dataset. Rather, the reliability of a model is determined by the difficult test data (here: OEl2, HOPV and Kuzmich). There is no point in reducing the test errors on QM9 below chemical accuracy on QM9-like data if the errors on large molecules are much larger.

We have achieved such a model with reasonable errors on the large molecules 1) by incorporating the Set2Set output module for improved generalization, 2) by extending the training set with data from four different sources, and 3) by introducing a multitask ansatz to handle the domain shifts due to different levels of theory in different sources.

5.2. Outlook and Future Work

The method is not restricted organic photovoltaics and is in general transferable to any problem in materials discovery where material properties are available as training data, for example, of organic light-emitting diodes, organic circuits, materials for batteries, or supercapacitors.

Our model can be used in future explorations of the chemical compound space in the search for better materials for organic photovoltaics. As indicated in Figure 1, we have interdisciplinary approaches in mind that incorporate machine learning for fast prediction, but also algorithms that generate new candidates, as well as ab initio methods and experiments for further verification and technical exploitation.

Approaches for molecular geometry generation range from genetic algorithms\cite{27,28} over fragment-based approaches\cite{29} to autoencoder-based methods\cite{30-32} and generative neural networks\cite{33,34}. One challenge for using the current model for such applications is that it is trained with the exact molecular geometries as obtained from DFT geometry optimizations. If the model is to be applied to structures that come from generating algorithm, the geometries will not be exact and may result in unexpected predictions by the model. Thus, the sensitivity of the model accuracy with respect to noise in the input coordinates needs be investigated. The problem can probably be alleviated by training the model with random noise added to the input geometries. Note, however, that this would change the regression task from “What are the properties of this molecular geometry?” to the question “Imagine this geometry relaxes to its near-by ground state, what would be its properties?”

6. Experimental Section

This section contains technical details that should facilitate the reproduction of the numerical experiments presented in this article.

Training Details: The common SchNet network was configured with six interaction units, learning a 128-dimensional representation of the atomic environments. The MLPs before/after the aggregation had 2 layers with the size of the hidden layer being 64 for the sum and average architectures (Figure 3a) and 32 for the Set2Set architecture (Figure 3b).

The Adam optimizer is with learning rate decay from $10^{-4}$ to $5 \times 10^{-7}$ (ReduceLROnPlateauHook with decay factor 0.5 and patience parameter $= 25$, i.e., the learning rate is reduced if the training error has not reduced in the last 25 epochs). After each pass through the training data (epoch), the regression error was computed on the validation set. The model with the lowest validation error was kept (early-stopping criterion based on the validation error). Typically, the best validation accuracy was achieved close to the first or second learning rate decay. Training took $\approx 350$ epochs (about 5 days on a nvidia Quadro GP100 GPU).

Code and Data: The code to reproduce this work is available at github.com/chgaul/maltose, DOI: 10.5281/zenodo.7328587, including the trained models and scripts to download the primary data from their respective sources. The maltose code contains 1) the maltose library, which provides the extensions to the SchNet code described in this article, and 2) a collection of scripts to execute and analyze the numerical experiments of this article.

The library is installed as a python module and provides the set2set output module and multitask functionality. It builds upon the schnetpack package\cite{33} (at least v1.0.1), which again makes use of the “Atomic Simulation Environment (ASE)”\cite{34} and pytorch-1.8.0. The implementation of the Set2Set output module depends on the packages pytorch-geometric and pytorch_scatter.
The scripts for data download and preparing the datasets (i.e., to convert to use format) are located at scripts/primary_data/ in the maltose repository, as well as the script that created the unified train/test/validation split. These scripts depend on the xyz2mol code from https://github.com/jensengroup/xyz2mol, which implements the algorithm from Ref. [37], and on the chemical toolkit RDKit.[38] The configurations of the algorithms are provided in scripts/configs/. There are scripts to run the training, to evaluate the errors, and to create the plots included in this article.

Acknowledgements

This work was supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no 883256.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo/GitHub at http://doi.org/10.5281/zenodo.7328587, Zenodo reference number 7328587.

Keywords

machine learning, molecular orbital energies, organic semiconductors

Received: November 23, 2022
Revised: February 2, 2023
Published online: March 2, 2023

[1] B. Lüssem, M. Riede, K. Leo, Phys. Status Solidi A 2013, 210, 9.
[2] Y. Firdaus, V. M. Le Corre, J. I. Khan, Z. Kan, F. Laquai, P. M. Beaupjuge, T. D. Anthopoulos, Adv. Sci. 2019, 6, 1802028.
[3] G. Han, Y. Yi, Acc. Chem. Res. 2022, 55, 869.
[4] G. Yu, J. Gao, J. C. Hummelen, F. Wudl, A. J. Heeger, Science 1995, 270, 1789.
[5] J. Zhao, Y. Li, G. Yang, K. Jiang, H. Lin, H. A. W. Ma, H. Yan, Nat. Energy 2016, 1, 15027.
[6] A. Wadsworth, M. Moser, A. Marks, M. S. Little, N. Gasparini, C. J. Brabec, D. Baran, I. McCulloch, Chem. Soc. Rev. 2019, 48, 1596.
[7] K. Chong, X. Xu, H. Meng, J. Xue, L. Yu, W. Ma, Q. Peng, Adv. Mater. 2022, 34, 2109516.
[8] W. Gao, F. Qi, Z. Peng, F. R. Lin, K. Jiang, C. Zhong, W. Kaminsky, Z. Guan, C.-S. Lee, T. J. Marks, H. Ade, A. K.-Y. Jen, Adv. Mater. 2022, 34, 2200897.
[9] M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldau, A. J. Heeger, C. J. Brabec, Adv. Mater. 2006, 18, 789.
[10] D. Padula, J. D. Simpson, A. Troisi, Mater. Horiz. 2019, 6, 343.
[11] D. Padula, A. Troisi, Adv. Mater. 2019, 9, 1902463.
[12] D. Rogers, M. Hahn, J. Chem. Inf. Comput. Sci. 2010, 50, 1549.
[13] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, in Advances in Neural Information Processing Systems, vol. 28 (Eds.: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett), Curran Associates, Inc., New York 2015, https://papers.nips.cc/paper/2015/hash/9be311e65d81a9ad8150a60844bb94c-Abstract.html.
[14] S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, J. Comput. Aided Mol. Des. 2016, 30, 1573.
[15] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, J. Chem. Phys. 2018, 148, 241722.
[16] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Chem. Mater. 2019, 31, 3564.
[17] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, J. Chem. Theory Comput. 2017, 13, 5255.
[18] J. Gilmer, S. Schoenholz, P. Riley, O. Vinyals, in Machine Learning Meets Quantum Physics (Eds.: K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K.-S. Tsuda, K.-R. Müller), Springer, Berlin 2020, pp. 199–214.
[19] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Sci. Data 2014, 1, 140022.
[20] C. Chen, P. Chen, C.-Y. Hsieh, C.-K. Lee, B. Liao, R. Liao, W. Liu, J. Oiu, Q. Sun, J. Tang, R. S. Zemel, S. Zhang, arXiv:1906.09427, 2019.
[21] A. Stuke, C. Kunkel, D. Golze, M. Todorovic, J. T. Margraf, K. Reuter, P. Rinke, H. Oberhofer, Sci. Data 2020, 7, 58.
[22] S. A. Lopez, E. O. Pyzer-Knapp, G. N. Simm, T. Lutzow, K. Li, L. R. Seress, J. Bachmann, A. Aspuru-Guzik, Sci. Data 2016, 3, 160086.
[23] A. Kuzmich, D. Padula, H. Ma, A. Troisi, Energy Environ. Sci. 2017, 10, 395.
[24] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, J. Cheminf. 2015, 7, 23.
[25] R. Caruana, Mach. Learn. 1997, 28, 41.
[26] O. Vinyals, S. Bengio, M. Kudlur, arXiv:1511.06391, 2015.
[27] S. Arapan, P. Nieves, S. Cuesta-López, J. Appl. Phys. 2018, 123, 083904.
[28] E. S. Henault, M. H. Rasmussen, J. H. Jensen, PeerJ. Phys. Chem. 2020, 2, e11.
[29] D. Flam-Shepherd, A. Zhigalin, A. Aspuru-Guzik, arXiv:2202.00658, 2022.
[30] M. J. Kusner, B. Paige, J. M. Hernández-Lobato, in Proc. 34th Int. Conf. on Machine Learning (Eds.: D. Precup, Y. W. Teh), Proceedings of Machine Learning Research, vol. 70, Sydney 2017, pp. 1945–1954, http://proceedings.mlr.press/v70/kusner17a.html.
[31] M. Simonovsky, N. Komodakis, in Artificial Neural Networks and Machine Learning – ICANN 2018 (Eds.: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis), Springer International Publishing, Cham 2018, pp. 412–422, https://arxiv.org/abs/1802.03480.
[32] D. Flam-Shepherd, T. C. Wu, A. Aspuru-Guzik, Mach. Learn. Sci. Technol. 2021, 2, 045009.
[33] N. Gebauer, M. Gastegger, K. Schütt, in Advances in Neural Information Processing Systems, vol. 32 (Eds.: H. Wallach, L. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, R. Garnett), Curran Associates, Inc., New York 2019, https://proceedings.neurips.cc/paper/2019/hash/ a4d8e2a7e0d0c102339f97716d2fdfb6-Abstract.html.
[34] N. W. A. Gebauer, M. Gastegger, S. S. P. Hessmann, K.-R. Müller, K. T. Schütt, Nat. Commun. 2021, 13, 973.
[35] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, K.-R. Müller, Nat. Commun. 2019, 15, 448.
[36] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dukåk, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, J. Phys.: Condens. Matter 2017, 29, 273002.
[37] Y. Kim, W. Y. Kim, Bull. Korean Chem. Soc. 2015, 36, 1769.
[38] G. Landrum, et al., Rdkit: Open-source cheminformatics. 2021_09_2 release, 2021, https://www.rdkit.org.
[39] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 1996, 77, 3865.
[40] A. Tkatchenko, M. Scheffler, Phys Rev Lett. 2009, 102, 073005.
[41] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, M. Scheffler, Comput. Phys. Commun. 2009, 180, 2175.
[42] C. Adamo, V. Barone, J. Chem. Phys. 1999, 110, 6158.
[43] J. P. Perdew, M. Ernzerhof, K. Burke, J. Chem. Phys. 1996, 105, 9982.
[44] M. Sinstein, C. Scheurer, S. Matena, V. Blum, K. Reuter, H. Oberhofer, J. Chem. Theory Comput. 2017, 13, 5582.