JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Electronic Descriptors for Supervised Spectroscopic Predictions

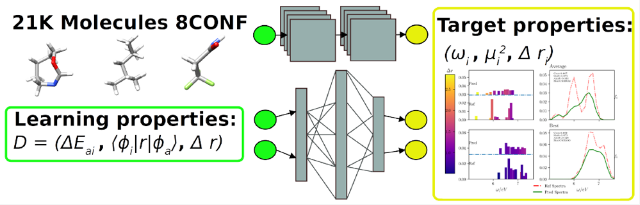
Carlos Manuel de Armas-Morejón,* Luis A. Montero-Cabrera,* Angel Rubio,* and Joaquim Jornet-Somoza*

Cite This: *J. Chem. Theory Comput.* 2023, 19, 1818−1826

Read Online

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

ABSTRACT: Spectroscopic properties of molecules hold great importance for the description of the molecular response under the effect of UV/vis electromagnetic radiation. Computationally expensive *ab initio* (e.g., MultiConfigurational SCF, Coupled Cluster) or TDDFT methods are commonly used by the quantum chemistry community to compute these properties. In this work, we propose a (supervised) Machine Learning approach to model the absorption spectra of organic molecules. Several supervised ML methods have been tested such as Kernel Ridge Regression (KRR), Multiperceptron Neural Networs (MLP), and Convolutional Neural Networks. [Ramakrishnan et al. *J. Chem. Phys.* **2015**, *143*, 084111. Ghosh et al. *Adv. Sci.* **2019**, *6*, 1801367.] The use of only geometrical-atomic number descriptors (e.g., Coulomb Matrix) proved to be insufficient for an accurate training. [Ramakrishnan et al. *J. Chem. Phys.* **2015**, *143*, 084111.] Inspired by the TDDFT theory, we propose to use a set of electronic descriptors obtained from low-cost DFT methods: orbital energy differences ($\Delta\epsilon_{ia} = \epsilon_a - \epsilon_i$), transition dipole moment between occupied and unoccupied Kohn−Sham orbitals ($\langle\phi_i|r|\phi_a\rangle$), and when relevant, charge-transfer character of monoexcitations (R_{ia}). We demonstrate that with these electronic descriptors and the use of Neural Networks we can predict not only a density of excited states but also get a very good estimation of the absorption spectrum and charge-transfer character of the electronic excited states, reaching results close to chemical accuracy (~2 kcal/mol or ~0.1 eV).



1. INTRODUCTION

The absorption spectra hold great importance for discovering photoelectric features in chemistry and materials science. The design of new photosensitive devices and materials for the energy industry as well as healthcare has become a hot topic in the last decades. A fast and accurate method that enables discrimination between hundreds or thousands of candidates becomes crucial to speed up new material discoveries with desired spectroscopic properties. The increase of the experimental and *ab initio* theoretical databases[1,3] on materials pushed forward a new way for their design, but usually they do not incorporate all the required spectroscopic information. Then, researchers rely on quantum mechanics techniques, usually Time-Dependent Density Functional Theory (TDDFT)[4,5] or multiconfigurational wave function methods,[6] for a rather confident prediction of properties and characterization. However, these types of calculations are usually complex to perform and to understand for nontrained researchers, particularly when trying to get reliable predictions of absorption spectra from an initial selection within several candidates.

Recently, Machine Learning (ML) algorithms have attracted the interest of the research community because the plausible results obtained predicted materials properties with good accuracy.[1,7] ML algorithms have been used, for example, for property classifications and group discovery,[8,9] as well as ground-state material and molecular property predictions.[7,10−12] It could also be very useful for understanding the nature of many molecular properties. The case of electronic excitations is able to be understood beyond the usual and very rough orbital descriptions, as they used to be based on rather approximate one-electron wave functions.[13] In addition, the so-called "inverse molecular design" could be aided if similarities among ML descriptors are appropriately used for such purposes.[14]

Profound research on several ML methods to be chosen, such as supervised or unsupervised models, kernel regression methods, or neural networks, etc., is required for each type of target property. Moreover, the choice of the appropriate molecular descriptors has to be made carefully in order to fulfill some desired criteria: 1) simplicity: must be easy to produce, 2) representability: must contain the required information correlated to the target property, and 3) specificity: must be unique enough to distinguish between different molecules.[15] Several descriptors have been proposed in the literature with different levels of applications.[8,16−45] Ouyang et al.[8] propose also the SISSO method for constructing these molecular or

Received: October 19, 2022
Published: March 6, 2023

https://doi.org/10.1021/acs.jctc.2c01039

material descriptors based on algebraic combinations of atomic properties.

Several attempts have been made to predict theoretical spectroscopic properties for molecules[1,12] and materials.[11,15,16,46] The seminal work done by Ramakrishnan et al.[1] proposed a kernel ridge regression model that can predict the first excited state with good results. Besides, they proposed a method called Δ*ML* for the estimation of the shift between two databases obtained using different Exchange-Correlation (XC) functionals. In that work, the authors used the so-called Coulomb Matrix[7] as a geometrical descriptor related to atomic numbers of vertex elements, which has gained notoriety because of its low computational requirements and its good performance for predicting molecular properties.[1,2] However, it proved to be insufficient for the proper prediction of the transition probability.[1] In a recent work, Westermayr et al. found a machine learning model based on the use of a complex Neural Network that using many conformers of the same molecule as a training set can be used to accurately predict its absorption spectra.[47]

In this work, we propose for the first time the use of some calculated electronic properties in order to well characterize the spectroscopic fingerprint of small molecules. By using a simple Convolutional Neural Network model trained by low-cost theoretical electronic calculations obtained from a 21k molecular database, we can predict excitation energies together with their corresponding charge-transfer character and oscillator strength. The results presented in this paper are obtained by employing electronic descriptors from ground-state DFT calculations using a simple LDA XC-functional to predict the absorption spectra at a TDDFT level using the PBE0 hybrid XC-functional. The validity of the selected model is contrasted with different Neural Network schemes, and the limitations are described on the basis of obtained results. Hence, the resulting trained Neural Network can be used to predict one or a large number of molecules with minimal computational cost.

1.1. Molecular Database and Descriptor Selection.

In this work, we take a subset of the *GDB-8* molecular database also used by Ramakrishnan et al.[1,3] It consists of 21k small organic molecules with relaxed geometries computed at the DFT level by using Gaussian09 with the B3LYP/6-31(2df,p) functional.[42] The selected molecules contain up to 8 carbon (C), oxygen (O), nitrogen (N), and/or fluor (F) atoms, being the number of hydrogen atoms required to make neutral the molecular charges. Hereon, we will refer to it as the 8CONF database.

Although the Sorted Coulomb Matrix and its variants have previously shown good results for the prediction of excitation energy levels and density of states,[1] the use of only geometrical molecular descriptors proved to be insufficient for the correct prediction of transition moments and oscillator strengths.

For that reason, we propose to use electronic molecular descriptors from low-cost theoretical calculations (ground-state's LDA) to predict accurate spectroscopic properties computed at the TDDFT level with a hybrid exchange-correlation functional (PBE0). We use the Octopus[48] code to compute all electronic descriptors. The current version of the code includes all required features and utilities to obtain them.[49] No further structural relaxation with the PBE0 functional for the training set was carried out in order to get predictions of spectroscopic information based on pure electronic descriptors.

The choice of the electronic descriptor has been made regarding the linear-response time-dependent DFT formulation (**LR-TDDFT**).[4,5] This approach aims to solve the time-dependent Shrödinger equation

$$\hat{H}(t)\Psi(t) = i\frac{\partial\Psi(t)}{\partial t}, \qquad \hat{H}(t) = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext}(t) \tag{1}$$

where $\hat{H}$ is the system Hamiltonian composed by a kinetic part $(\hat{T})$, an electron−electron potential component $(\hat{V}_{ee})$, and the all-other types of interaction contained in the time-dependent external potential term $\hat{V}_{ext}(t)$. Usually, the latter contains the nuclei-electron and external field interaction.

In LR-TDDFT, the time-dependent evolution of the noninteracting system under an external field is described by the noninteracting *density−density* response function $\chi_s(\boldsymbol{r}, \boldsymbol{r}', \omega)$

$$\chi_s(\boldsymbol{r}, \boldsymbol{r}', \omega) = \lim_{\eta\to 0^+} \sum_{k,j} (f_k - f_j)\delta_{\sigma_k\sigma_j} \frac{\varphi_k^{(0)*}(\boldsymbol{r})\varphi_j^{(0)}(\boldsymbol{r})\varphi_j^{(0)*}(\boldsymbol{r}')\varphi_k^{(0)}(\boldsymbol{r}')}{\omega - (\varepsilon_j - \varepsilon_k) + i\eta} \tag{2}$$

where φ_j stands for Kohn−Sham (KS) orbitals, ϵ_j and f_j stand for their corresponding energies and occupations, respectively, and $\delta_{\sigma_k\sigma_j}$ is a Kronecker delta for orbital *j* and *k* spin functions. Finally, ω is the frequency of the perturbing external field, and η is a positive infinitesimal.

This function has poles on the excitation energy of the KS system. In order to obtain the excitation energies of the full interacting system, we have to solve the Dyson-like equation. Casida et al. proposed a matrix formulation to solve this equation, and he obtained the well-known Casida's equation.[50] By solving this equation, the excitation energy levels and oscillator strengths are obtained as a combination of the biorbital function $\Phi_{ia}(\boldsymbol{r}) = \varphi_i^*(\boldsymbol{r})\varphi_a(\boldsymbol{r})$ where subindices *a* and *i* correspond to unoccupied and occupied states, respectively.

In the present work, we use the excitation energies and oscillator strengths of 15k molecules computed using Casida's equations at the PBE0 functional level in the ground state, both to train our ML model and as targets to validate it.

Let us return to the LR-TDDFT formulation in order to define the descriptors we will use. The time evolution of the polarizability function is defined in LR-TDDFT as the dipole−dipole response function, which in the space of frequencies takes the form

$$\alpha_{\mu\lambda}(\omega) = \sum_{n=1}^{\infty} \left\{ \frac{2\Omega_n \langle\Psi_0|\hat{r}_\nu|\Psi_n\rangle\langle\Psi_n|\hat{r}_\lambda|\Psi_0\rangle}{\Omega_n^2 - \omega^2} \right\} \tag{3}$$

where Ψ_0 is the actual ground-state wave function of the interacting system, Ψ_n is the *n*-th excited-state wave function, Ω_n is the excitation energy of the *n*-th excited state, μ and λ are directions in the space, and $\hat{r}_\nu$ is the dipole operator in the ν direction.

This function has also poles in the excitation energies of the system, and the corresponding oscillation strengths are proportional to the numerator. Therefore, it is also used to compute the absorption spectra of molecular systems.

Based on eq 3 and the use of biorbital functions representing monoelectronic excitations, we propose to use the following electronic descriptors:

1. Orbital energy difference: $\Delta\epsilon_{ia} = \epsilon_a - \epsilon_i$.
2. Kohn−Sham transition moment: $\mu_{ia}^2 = |\langle\varphi_i|\boldsymbol{r}|\varphi_a\rangle|^2$

Nevertheless, the only use of these two properties does not fulfill the desired criteria of specificity described above. The

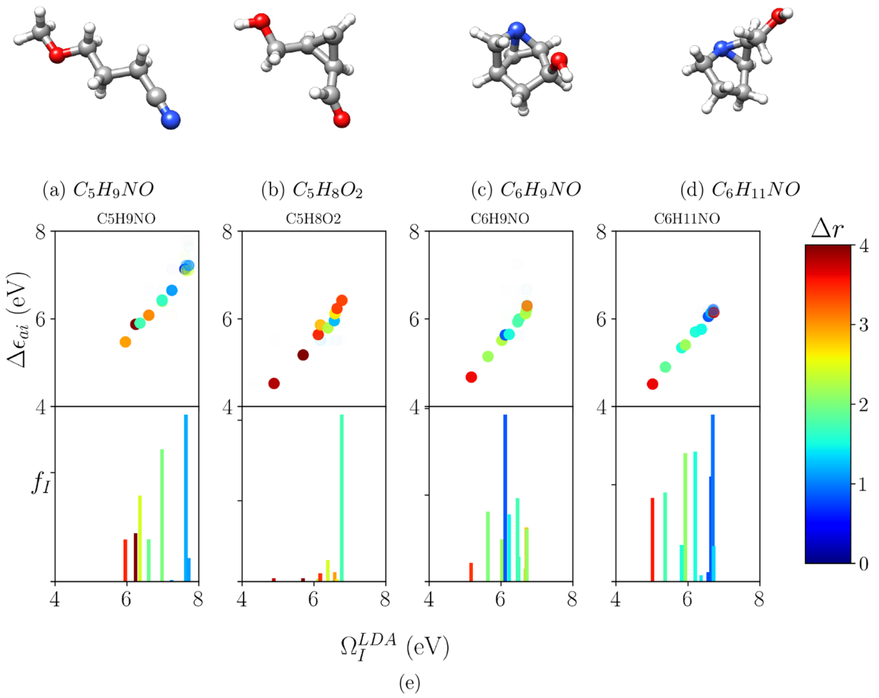


Figure 1. (a-d) Examples of four selected molecules from the 8CONF database. C, H, N, and O atoms are represented in gray, white, blue, and red, respectively. (e) Correlation between orbital energy differences and LR-TDDFT calculated excitation energies Ω_I^{LDA} (first row from top to bottom). The second row shows the calculated discrete absorption spectra for those molecules corresponding to such excitation energies. Transparency is proportional to Casida's coefficient, and color is proportional to the charge-transfer character.

calculated oscillator strengths depend on orbital overlapping between unoccupied and occupied states and are hence proportional to transition dipole moments. Besides, a work of Guido et al.[20] proposes an easy way to evaluate the charge-transfer character of an excitation by defining a new index, $\Delta\boldsymbol{r}$

$$\Delta\boldsymbol{r} = \frac{\sum_{i,a} K_{ia}^2 R_{ia}}{\sum_{i,a} K_{ia}^2} \tag{4}$$

$$R_{ia} = |\langle\varphi_a|\mathbf{r}|\varphi_a\rangle - \langle\varphi_i|\mathbf{r}|\varphi_i\rangle| \tag{5}$$

$$K_{ia} = X_{ia} + Y_{ia} \tag{6}$$

where the intervening excitation X_{ia} and de-excitation Y_{ia} coefficients of the non-Hermitean solution correspond to the TD formalism.

We would like to remark that, although R_{ia} vanishes for centrosymmetric molecules and, therefore, does not introduce information in such cases, it remains relevant for all others. A zero value of one of the descriptors for them means that they remain only defined by those evaluated, keeping the full model of descriptors for the rest of the database.

Following that index and knowing that only symmetrically similar monoexcitation contributes to the real excited state (Figure 1), we decided to also include the charge-transfer character of KS monoexcitations as a descriptor, as well as the TDDFT charge-transfer index of the excited state as a target property to predict.

Consequently, in this work, we propose the use of the combination of three electronic descriptors, namely (i) $\Delta\epsilon_{ia}$, (ii) μ_{ia}^2, and (iii) R_{ia}, computed at the ground-state LDA XC-functional level and LCAO (LDA), for the 20 lowest-lying monoelectronic transitions in order to predict three spectroscopic properties for the first ten excitations: a) excitation energy (Ω_I), b) oscillator strength (f_I), and c) charge-transfer character ($\Delta\boldsymbol{r}$) at the PBE0 accuracy level.

1.2. Supervised Machine Learning Models.

In this work, we use Neural Networks (NNs) because of their recognized versatility to find hidden correlations among several properties. We explore different NN models such as the Multi-Layer Perceptron (MLP) and the Convolution Neural Network (CNN). Each model depends on a group of internal variables, known as hyperparameters, such as the number of hidden layers, the number of neurons per layer, the number of learning iterations (also known as *epochs*), the activation functions, and many others. Some of these variables need to be optimized to fine-tune the NN.

Hyperparameter optimizations hold great importance for the correct behavior of the NN. Their values can be empirically selected, but the best combination can be only achieved by a systematic search. Then, we applied a Bayesian Optimization as implemented in *scikit-learn*[51] to find the optimal values for the *number of hidden layers* and the *number of epochs*.

Another important issue is how to feed the data to the NN. The flexibility of the NN allows many configurations for introducing the descriptors into the model. Consider $d_{m,n}$ an element of the input tensor $\boldsymbol{D}$, where n is the descriptor property ($\Delta\epsilon_{ia}$, μ_{ia}^2, or R_{ia}), and m is the considered $a \leftarrow i$ monoexcitatiton label. We used two strategies to introduce our data into the neural network: (1) the **1D model**, where each sample j is described by an array of 3m elements (one-dimensional), where all properties are introduced sequentially: $D_j = (\Delta\varepsilon_1, \mu_1^2, R_1, ..., \Delta\varepsilon_m, \mu_m^2, R_m)$, and (2) the **2D model**, where the descriptor properties for each m monoexcitation are grouped forming a two-dimensional tensor of (m,3) dimension,

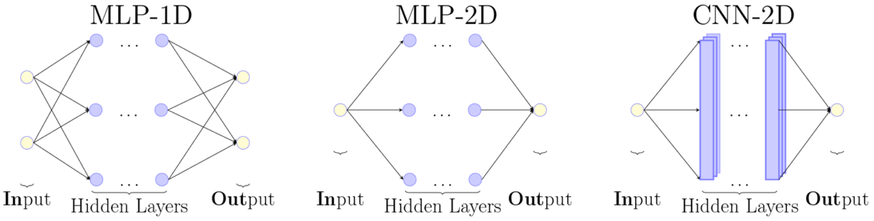


Figure 2. Simplified view of the topology built on top of *Keras*[52] and *Tensorflow*.[53] From left to right, MLP-1D, where the input data is treated independently, and MLP-2D and CNN-2D, where the input data is treated as a combination of descriptors in both cases. See text.

since all these properties are required to describe a particular excitation: $D_j = ((\Delta\varepsilon_1, \mu_1^2, R_1), ..., (\Delta\varepsilon_m, \mu_m^2, R_m))$.

The use of different properties, units, and ranges of magnitude may affect the learning process. It is always recommended to perform data preprocessing in order to give the same weight for all properties and hence to ease the learning process. In this work, we decided to scale all the data between [0, 1] using the tool *MinMaxScaler* provided by the *preprocessing* package of *scikit-learn*.[51] Since the range of transition dipole moments is always positive and presents a high density distribution for values between 0 and 1, we transformed this property to a logarithmic scale.

Alongside with the preprocessing methods several NN models have been tested to find which will best perform for predicting properties. Figure 2 represents the different ML models tested in this work.

We construct our models using Keras and TensorFlow[52,53] with the hyperparameters shown in Table 1 resulting from a Bayesian Optimization against the accuracy values obtained over the test set. Since the number of hyperparameters is large, only two have been optimized: 1) the number of layers and 2) the number of epochs. The activation function, which transforms the values between neurons across layers, has been selected empirically resulting from the use of the eLU[54] and ReLU[55] functions. Both activation functions were alternated starting with eLU. The full description of the models used in this work can be found in the Supporting Information.

Table 1. Relevant Hyperparameters Used to Build Our Models[a]

model	epochs	n. hidden layers	activation function
MLP-1D	1756	17	eLU/ReLU(negative slope = 0.01)[54,55]
MLP-2D	1419	4	eLU/ReLU(negative slope = 0.01)[54,55]
CNN-2D	1500	2	eLU/ReLU

[a] The second and third columns show the number of *Epochs* and *Hidden Layers*, respectively, as optimized for the model shown in the first column. The fourth column specifies the type of activation function and properties. For a complete list, see the Supporting Information.

By using this configuration, the training process with 10k molecules as a learning set and the prediction of 1k molecules from the test set takes ~1528 s for the slowest NN (CNN-2D model) on a commercial laptop with an Intel i7 and 12 Gb of RAM.

Table 2 shows the required times for the training and predicting processes. It is important here to remark that once the model is trained, the prediction of the spectroscopic properties is almost instantaneously obtained. Comparing this performance with the arduous task of solving the complex TDDFT equations shows the potential of using such an ML tool.

Table 2. Approximate *Training* and *Predicting* Times in Seconds (s) for Each Model Using 10k Molecules and 1k Molecules, Respectively[a]

model name	use *log*	training process (s)	predicting process (s)	total time (s)
MLP-1D	yes	919	<1	919
MLP-2D	yes	319	<1	319
CNN-2D	yes	1521	<1	1522

[a] Even if CNN-2D takes 1521 s to train, which is the slowest one, the response when predicting the properties is almost instantaneous.

The learning process is obviously biased by the learning data set. In order to validate the input data distribution, the optimized model, and its uncertainty, Musi et al. proposed to systematically perform a stability test.[56] The learning data set is validated by repeating the neural network construction process (training and predicting process) over 10 experiments, by randomly selecting 10k learning molecules out of 15k PBE0 available calculations. We used 1k of the remaining 3k molecules as a control, which will remain unchanged across all experiments.

Figure 3 shows the mean average errors (MAEs) for the first 10 excited states of 1k 8CONF small molecules used in the control set. For each state, the mean value of the error over the 10 experiments is represented, and its standard deviation is depicted as a bar line.

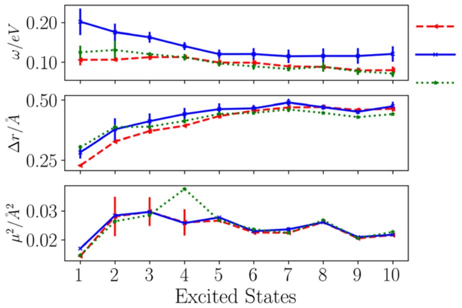


Figure 3. Mean absolute error (MAE) on the prediction of the excited-state energies Ω (top), charge-transfer character index Δr (middle), and transition moments μ^2 (down) for 10 low-lying states averaged over the 10 repetitions of the same ML model. The MAE standard deviation of those experiments is also represented by the bar amplitude.

We can see that CNN-2D and MPL-1D models reach errors of the excitation energy predictions that are close to chemical accuracy (~0.1 eV). Besides, they can also correctly predict the charge-transfer character of the low-lying excited state. Notice that Figure 1 shows Δr property as ranging from 0 to 4 Å. It

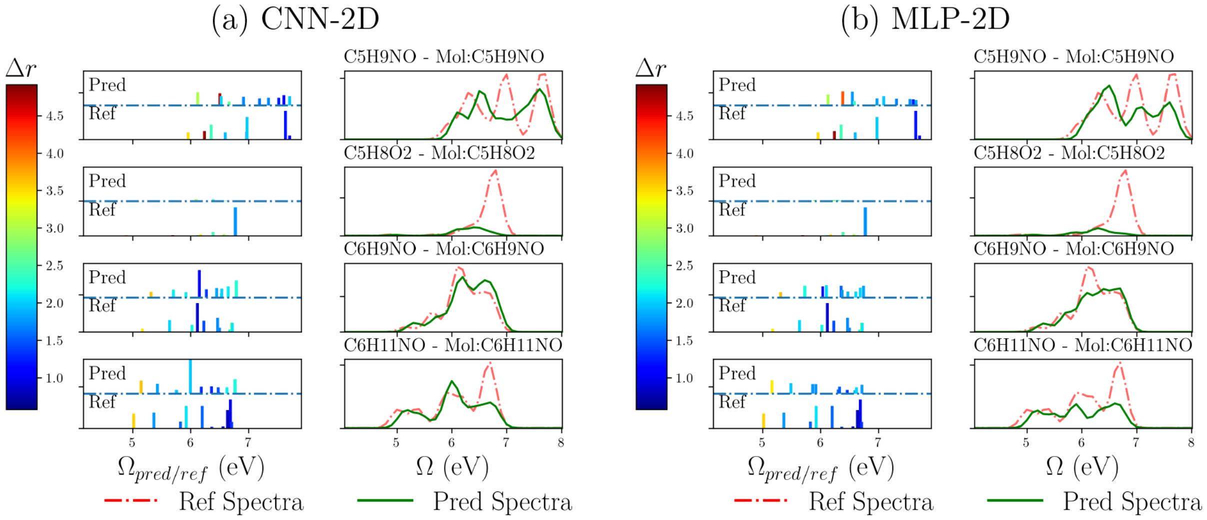


Figure 4. Discrete and broadened excitation spectra obtained with CNN-2D (a) and MLP-2D (b) for some example molecules contained in the control group. On the *X* axis are the calculated excited-state energies Ω, and on the *Y* axis appear their oscillator strengths, f_I. Green curves represent reconstructed spectra from NN predictions, while the red ones represent those from PBE0-CASIDA calculations.

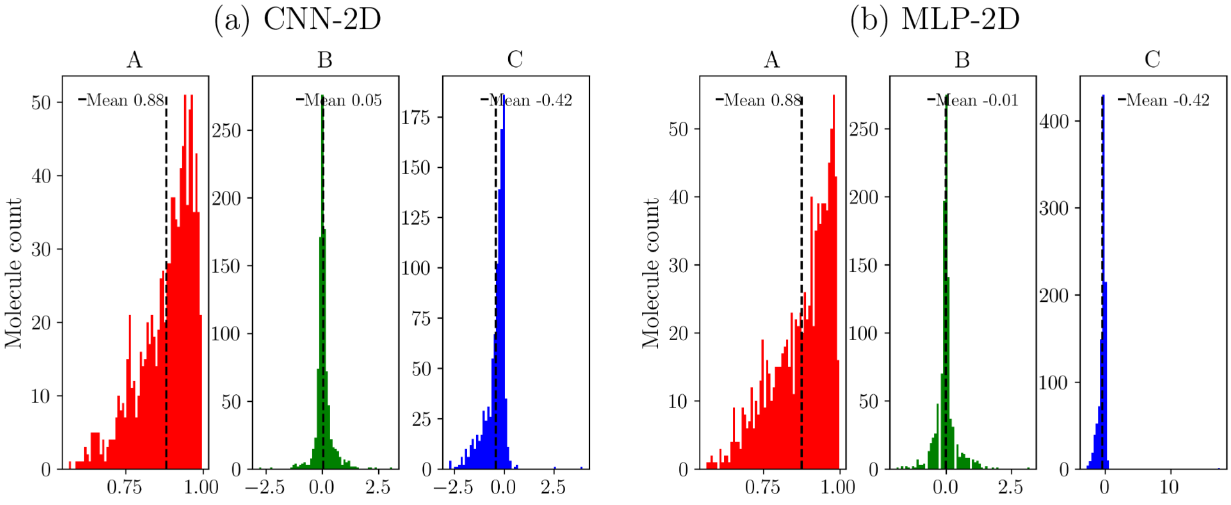


Figure 5. Distribution of the different metrics used to evaluate the prediction of the broadened spectra: (A, red) cross-correlation, (B, green) area under the curve, and (C, blue) curve shift for (a) CNN-2D and (b) MLP-2D.

means that a resulting error being smaller than 0.50 Å clearly distinguishes between short- and long-range charge-transfer characters. Regarding the transition moment prediction, we see that the models can just fairly predict the first excitation probability.

In the following section, we discuss the spectroscopic properties obtained using the more promising models, and we will remark on their strength and drawbacks.

2. RESULTS AND DISCUSSION

As already described above, the main goal of this work is to find an adequate molecular-electronic descriptor that enables us to obtain accurate spectroscopic properties using machine learning techniques. From the analysis of accuracy and stability (Figure 3), we see that CNN-2D and MLP-2D produce the lowest error for predictions. Therefore, we selected these models to perform a deeper analysis.

In order to easily visualize the agreement of the models for predicting the optical response of a molecule, we must also look at the absorption spectra. Figure 4 shows some examples of reconstruction for discrete and broadened absorption spectra (More examples can be found in the Supporting Information.).

The discrete spectra is represented as impulses positioned at the specific excitation energy of a given state. Their heights are proportional to the calculated oscillator strength. The shown color of each impulse represents the index for charge-transfer character according to the color scales of **Δ*r*** on the left side. It is well-known that LDA functionals tend to underestimate excitation energies between Kohn−Sham's orbitals when they involve a charge-transfer process. This is avoided in TDDFT calculations because of the consideration of a fraction of the exact Hartree exchange potential for hybrid functionals, such as PBE0. If we look at the spectra of C_5H_9NO, the first excited-state energy and nature in Figure 1, that was obtained with simpler LDA calculation results used for NN learning, they can be compared with the predicted results of Figure 4. A switch on the charge-transfer character of the first excited state appears together with a blue shift of the energy in the prediction. It

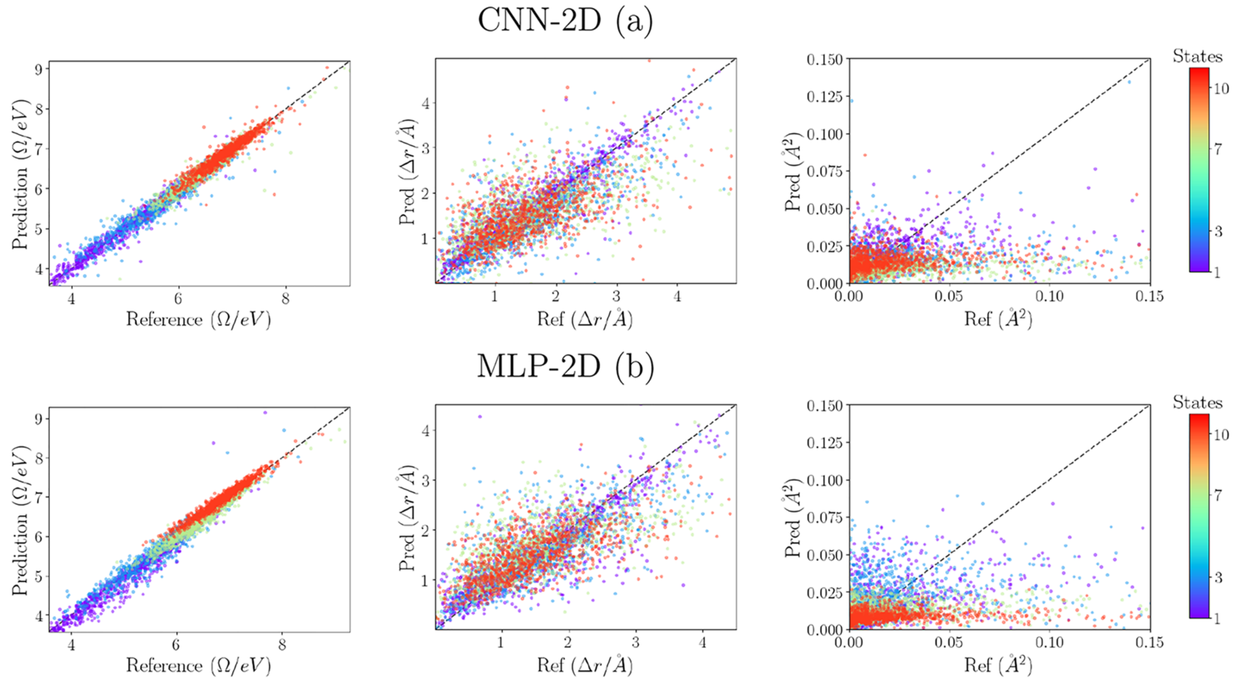


Figure 6. Correlation graphics between predicted and calculated values for best models (CNN-2D (a) and MLP-2D (b)). From left to right, the following are represented: excitation energies (Ω), charge-transfer coefficients (Δr), and transition moments (μ^2). The color of the points corresponds to the state number as ordered from the lowest energy according to the scale on the right-hand side.

means that even though our models were fed with rough LDA calculated properties, the learning process appears to add the effect of the exchange-correlation, being this is one of the more time demanding parts when computing spectroscopic properties by TDDFT routines. It could be very significant to save computational resources and time for excited-state predictions of large molecules.

Besides, the broadened spectra shown in Figure 4 have been reconstructed as a sum of Gaussians centered at the excitation energy which area is proportional to the oscillator strength. The broadening factor has been chosen to be 0.15 eV at the half width at the half-maximum (HWHM). We used the cross-correlation between the normalized spectra, the curve shift, and the curve area difference for comparison metrics between reference and predicted broadened spectra, because the relevant property in spectroscopy is usually the relative intensity instead of its absolute value. Our models sometimes have difficulties in distinguishing between closely lying excitations and can produce a switch between states. However, the broadening procedure enables us to mitigate this error by producing a good estimation of the continuum absorption spectra.

Figure 5 shows the distribution of the parameters used to evaluate the obtained broadened spectra. Both models show a very good distribution of the cross-correlation values having more than 93% situated above a value of 0.90. In addition, almost all tests produce an area under the curve close to the 0 values, which indicates that the number of electrons is conserved in the prediction. Regarding the shift of the predicted broadened spectra, we observe that in spite of the fact the great majority presents just a slight shift, both models tend to produce a small red-shift of the absorption spectra when comparing with the PBE0-CASIDA results.

Figure 6 shows the correlation graphics between the predicted values and the reference for each of the spectroscopic properties analyzed in this work. We can see that our models produce a very good correlation of the excitation energies for all excited states. A better correlation is observed for higher excitation energies, which can be attributed to a higher density of states found in the database at such frequencies. Regarding the charge-transfer index, we see that the CNN-2D model produces better correlation for the low-lying states, while it loses this correlation for higher states. It seems that both models hardly reproduce the proper transition dipole moments, being the CNN-2D model is slightly better for the low-lying states. A possible source of error can be attributed to the diverse distribution of the values that increases the complexity of the learning process and/or to intrinsic inconsistencies of the theoretical calculations of transition dipole moments when Kohn−Sham's virtual orbitals are involved. Other tests performed by increasing the size of our learning set including up to 15k molecules suggest that enlarging the database can improve the correlation of the **Δ*r*** for higher excitations but just produces a slight improvement in the transition dipole moments.

3. CONCLUSIONS

Accurate knowledge of the spectroscopic properties of molecules has been of great interest since long ago for academical as well as industrial sectors. The high computational cost of the quantum chemistry/physics techniques, mostly for large molecules, and the lack of an extended experimental database, as well as the increase on the reliability of the machine learning and artificial intelligence methods, are inviting researchers to apply those techniques for predicting such physical properties. Nevertheless, the major difficulty usually relies on finding the proper descriptors being able to correlate with properties of interest.

As mentioned above, the main objective of this work is to find adequate molecular-electronic descriptors to be used with proper NN models to predict the theoretical absorption spectra for a group of small organic molecules. We prove that the combination of certain selected electronic properties ($\Delta\epsilon_{ia}$, R_{ia}, μ_{ia}^2) resulting from low cost *ground state LDA* calculations appears as good descriptors for the prediction of such spectroscopic properties at a higher level of theory (e.g., TDDFT with the PBE0 functional without a geometry relaxation). Besides, we demonstrate that a simple optimized Convolutional Neural Network (CNN-2D) as well as a Multi Layer Perceptron (MLP-2D) network can learn to supply the exchange correlation correction required for predictions going from LDA to PBE0 levels of theory.

Many others advantages arise as (i) the trained NN model can be reused for further predictions on previously unknown molecules; (ii) geometry optimizations are not required although could be recommended; and (iii) the trained NN model expects as input the ground state LDA electronic data, and it can be obtained from any (TD)DFT code that managed to produce this output.

Previous works were focused on the prediction of the first excited state[1] or on the density of states near to the LUMO[2] by using only geometrical or spacial descriptors. In this work, we demonstrate the need of an electronic descriptor not only to extend the prediction of the excitation energies at chemical accuracy but also to give information about their charge-transfer character. Oscillator strength values proved to be the most challenging property for our models. Although we demonstrate an enhancement on the prediction of the low-lying excitation probabilities when the training set is augmented, the transition dipole moments for high energy excitation remained poorly correlated. Different sources of error that can be addressed for this problem are discussed.

We can conclude that by constructing a large database which would include all types of molecules (small, medium, and large sized molecules), the presented method would be able to precisely obtain spectroscopic properties for any unknown molecule by just computing a low-cost DFT ground state using the LDA functional. This goal can be achieved by using, for example, the Novel Materials Discovery Laboratory (NOMAD) open database.[57]

The natural next step is to provide even more fundamental properties as descriptors to the Neural Network. However, in progress work reveals that the use of the same properties coming from unoptimized Linear Combination of Atomic Orbitals (LCAO) calculations, typically used to begin a ground state calculation, requires a more complex network optimization to overcome the big gap between an LCAO level of theory and that of PBE0 hybrid functional calculations. Besides, we are also working on the use of the simplest ML techniques to predict optimized geometries, that could avoid the tedious task of geometry relaxations.[58]

All data from the calculations done in this work have been stored in the database of the Novel Material Discovery Laboratory (NOMAD) project[57] and can be downloaded: LDA from 10.17172/NOMAD/2021.10.18-2 and PBE0 from 10.17172/NOMAD/2021.10.18-3.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c01039.

Additional data and NN descriptions (PDF)

AUTHOR INFORMATION

Corresponding Authors

Carlos Manuel de Armas-Morejón − *Nano-Bio Spectroscopy Group, Departamento de Polímeros y Materiales Avanzados: Fisica, Química y Tecnología, Universidad del País Vasco UPV/EHU, 20018 San Sebastián, Spain; Laboratorio de Química Computacional y Teórica, Facultad de Química, Universidad de La Habana, 10400 La Habana, Cuba*; Email: carlosdearmasm@gmail.com

Luis A. Montero-Cabrera − *Laboratorio de Química Computacional y Teórica, Facultad de Química, Universidad de La Habana, 10400 La Habana, Cuba; Donostia International Physics Center, 20018 Donostia, Spain*; orcid.org/0000-0002-4128-1203; Email: lmc@fq.uh.cu

Angel Rubio − *Theory Department, Max Planck Institute for the Structure and Dynamics of Matter and Center for Free-Electron Laser Science, 22761 Hamburg, Germany; Nano-Bio Spectroscopy Group, Departamento de Polímeros y Materiales Avanzados: Fisica, Química y Tecnología, Universidad del País Vasco UPV/EHU, 20018 San Sebastián, Spain*; orcid.org/0000-0003-2060-3151; Email: angel.rubio@mpsd.mpg.de

Joaquim Jornet-Somoza − *Nano-Bio Spectroscopy Group, Departamento de Polímeros y Materiales Avanzados: Fisica, Química y Tecnología, Universidad del País Vasco UPV/EHU, 20018 San Sebastián, Spain; Theory Department, Max Planck Institute for the Structure and Dynamics of Matter and Center for Free-Electron Laser Science, 22761 Hamburg, Germany*; orcid.org/0000-0002-6721-1393; Email: j.jornet.somoza@gmail.com

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c01039

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC-2015-AdG694097), the Cluster of Excellence 'CUI: Advanced Imaging of Matter' of the Deutsche Forschungsgemeinschaft (DFG) - EXC 2056 - project ID 390715994, Grupos Consolidados (IT1249-19), and the SFB925 "Light induced dynamics and control of correlated quantum systems". We kindly recognize the partial support of the project ID PN223LH010-002 "Inteligencia Artificial Aplicada, Espectroscopía y Bioactividad" of the Cuban Ministry of Science, Technology and Environment as well as the overall support given to L.A.M.C. by the Universidad de La Habana and the Donostia International Physics Center.

REFERENCES

(1) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. *J. Chem. Phys.* **2015**, *143*, 084111.

(2) Ghosh, K.; Stuke, A.; Todorović, M.; Jörgensen, P.; Schmidt, M.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Advanced Science* **2019**, *6*, 1801367.

(3) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(4) Marques, M. A. L.; Maitra, N. T.; Nogueira, F. M. S.; Gross, E. K. U.; Rubio, A. *Fundamentals of Time-Dependent Density Functional*

Theory; Physics Ed.ial Department I: 2012; DOI: 10.1007/978-3-642-23518-4.

(5) Ullrich, C. *Time-Dependent Density-Functional Theory: Concepts and Applications*; Oxford University Press: 2012; DOI: 10.1093/acprof:oso/9780199563029.001.0001.

(6) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry Introduction to Advanced Electronic Structure Theory*; Dover Publications, Inc.: 1996.

(7) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(8) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. SISSO: a compressed-sensing method for systematically identifying efficient physical models of materials properties. *Phys. Rev. Materials* **2018**, *2*, 083802.

(9) Goldsmith, B. R.; Boley, M.; Vreeken, J.; Ghiringhelli, L. M.; Scheffler, M. Uncovering structure-property relationships of materials by subgroup discovery. *New J. Phys.* **2017**, *19*, 013031.

(10) Pilania, G.; Gubernatis, J.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.

(11) Schütt, K.; Tkatchenko, A.; Gastegger, M.; Müller, K.-R. Unifying machine learning and quantum chemistrywith a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.

(12) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Vinyals, O.; Dahl, G. E.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.

(13) Kimber, P.; Plasser, F. Toward an understanding of electronic excitation energies beyond the molecular orbital picture. *Phys. Chem. Chem. Phys.* **2020**, *22*, 6058−6080.

(14) Green, J. D.; Fuemmeler, E. G.; Hele, T. J. H. Inverse molecular design from first principles: Tailoring organic chromophore spectra for optoelectronic applications. *J. Chem. Phys.* **2022**, *156*, 180901.

(15) Ghiringhelli, L.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science - Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(16) De, S.; Bartók, A. P.; Csányic, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phisycal Chemistry Chemical Physics* **2016**, *18*, 13754−13769.

(17) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(18) Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.

(19) Marín, R. M.; Aguirre, N. F.; Daza, E. E. Graph Theoretical Similarity Approach To Compare Molecular Electrostatic Potentials. *J. Chem. Inf. Model.* **2008**, *48*, 109−118.

(20) Guido, C. A.; Cortona, P.; Mennucci, B.; Adamo, C. On the Metric of Charge Transfer Molecular Excitations: A Simple Chemical Descriptor. *J. Chem. Theory Comput.* **2013**, *9*, 3118−3126.

(21) Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for measuring distances in configuration spaces. *J. Chem. Phys.* **2013**, *139*, 184118.

(22) Dong, J.; Cao, D. S.; Miao, H. Y.; Liu, S.; Deng, B. C.; Yun, Y. H.; Wang, N. N.; Lu, A. P.; Zeng, W. B.; Chen, A. F. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform* **2015**, *7*, 60.

(23) Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.

(24) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(25) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495−4501.

(26) Himanen, L.; Jäger, M. O. J.; Morooka, Eiaki V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(27) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(28) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

(29) Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309−3313.

(30) Huang, B.; Symonds, N. O.; von Lilienfeld, O. A.Quantum Machine Learning in Chemistry and Materials. In *Handbook of Materials Modeling*; Andreoni, W., Yip, S., Eds.; Springer, Cham, 2018.

(31) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Computational Chemistry* **2018**, *57*, 4164−4169.

(32) Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **2011**, *83*, 153101.

(33) Artrith, N.; Urban, A.; Ceder, G. Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species. *Phys. Rev. B* **2017**, *96*, 014112.

(34) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Scientific Advance* **2017**, *3*, e1701816.

(35) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(36) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(37) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121* (1), 511.

(38) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **2019**, *5*, 22.

(39) Hughes, Z. E.; Thacker, J. C. R.; Wilson, A. L.; Popelier, P. L. A. Description of Potential Energy Surfaces of Molecules Using FFLUX Machine Learning Models. *J. Chem. Theory Comput.* **2019**, *15*, 116−126.

(40) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

(41) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. The Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett.* **2018**, *9*, 2725.

(42) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.

(43) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045017.

(44) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(45) Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27−35.

(46) Kennedy, M. C.; Hagan, A. H. Predicting the output from a complex computer code when fast approximations are avalible. *Biometrika* **2000**, *87*, 1.

(47) Westermayr, J.; Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *J. Chem. Phys.* **2020**, *153*, 154112.

(48) Andrade, X.; Alberdi-Rodriguez, J.; Strubbe, D. A.; Oliveira, M. J. T.; Nogueira, F.; Castro, A.; Muguerza, J.; Arruabarrena, A.; Louie, S.

G.; Aspuru-Guzik, A.; Rubio, A.; Marques, M. A. L. Time-dependent density-functional theory in massively parallel computer architectures: the octopus project. *J. Phys.: Condens. Matter* **2012**, *24*, 233202.

(49) Tancogne-Dejean, N.; Oliveira, M. J. T.; Andrade, X.; Appel, H.; Borca, C. H.; Le Breton, G.; Buchholz, F.; Castro, A.; Corni, S.; Correa, A. A.; De Giovannini, U.; Delgado, A.; Eich, F. G.; Flick, J.; Gil, G.; Gomez, A.; Helbig, N.; Hübener, H.; Jestädt, R.; Jornet-Somoza, J.; Larsen, A. H.; Lebedeva, I. V.; Lüders, M.; Marques, M. A. L.; Ohlmann, S. T.; Pipolo, S.; Rampp, M.; Rozzi, C. A.; Strubbe, D. A.; Sato, S. A.; Schäfer, C.; Theophilou, I.; Welden, A.; Rubio, A. Octopus, a computational framework for exploring light-driven phenomena and quantum dynamics in extended and finite systems. *J. Chem. Phys.* **2020**, *152*, 124119.

(50) Casida, M. E. Time-Dependent Density Functional Response Theory for Molecules. *Recent Advances in Density Functional Methods*; 1995; p 155, DOI: 10.1142/9789812830586_0005.

(51) Szymański, P.; Kajdanowicz, T. scikit-multilearn: A Python library for Multi-Label Classification. *Journal of Machine Learning Research* **2019**, *20*, 1−22.

(52) Chollet, F. *Keras*; 2015; https://keras.io, {https://github.com/fchollet/keras} (accessed 2023-03-01).

(53) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015, arXiv:1603.04467. *arXiv*. https://arxiv.org/abs/1603.04467 (accessed 2023-03-01). Software available from https://www.tensorflow.org (accessed 2023-03-01).

(54) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). 2016, arXiv:1511.07289v5. *arXiv*. https://arxiv.org/abs/1511.07289 (accessed 2023-03-01).

(55) Nair, V.; Hinton, G. E. Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27 International Conference on Machine Learning*; 2010.

(56) Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906−915.

(57) Draxl, C.; Scheffler, M. NOMAD: The FAIR Concept for Big-Data-Driven Materials Science. 2018, arXiv:1805.05039v1. *arXiv*. https://arxiv.org/abs/1805.05039 (accessed 2023-03-01).

(58) de Armas-Morejón, C. M.; Larsen, A. H.; Montero-Cabrera, L. A.; Rubio, A.; Jornet-Somoza, J. A basic electro-topological descriptor for the prediction of organic molecule geometries by simple machine learning. 2022, arXiv:2210.10700. *arXiv*. https://arxiv.org/abs/2210.10700 (accessed 2023-03-01).