THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

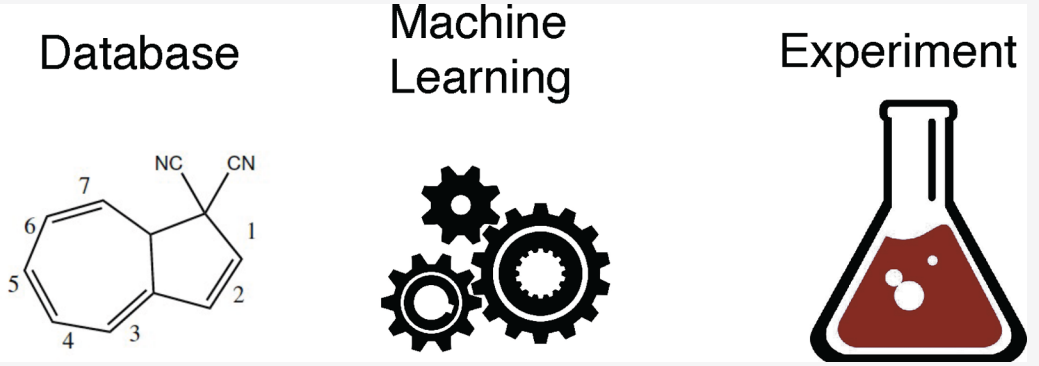
pubs.acs.org/JPCA Article

A Neural Network Approach for Property Determination of Molecular Solar Cell Candidates

Oliver Christensen,[†] Rasmus Dalsgaard Schlosser,[†] Rasmus Buus Nielsen, Jes Johansen, Mads Koerstz, Jan H. Jensen, and Kurt V. Mikkelsen*

Cite This: *J. Phys. Chem. A* 2022, 126, 1681−1688 Read Online

ACCESS | Metrics & More | Article Recommendations



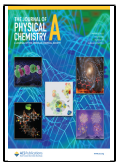
ABSTRACT: The dihydroazulene/vinylheptafulvene (DHA/VHF) photocouple is a promising candidate for molecular solar heat batteries, storing and releasing energy in a closed cycle. Much work has been done on improving the energy storage capacity and the half-life of the high-energy isomer via substituent functionalization, but similarly important is keeping these improved properties in common polar solvents, along with being soluble in these, which is tied to the dipole properties. However, the number of possible derivatives makes an overview of this combinatorial space impossible both for experimental work and traditional computational chemistry. Due to the time-consuming nature of running many thousands of computations, we look to machine learning, which bears the advantage that once a model has been trained, it can be used to rapidly estimate approximate values for the given system. Applying a convolutional neural network, we show that it is possible to reach good agreement with traditional computations on a scale that allows us to rapidly screen tens of thousands of the DHA/VHF photocouple, eliminating bad candidates and allowing computational resources to be directed toward meaningful compounds.

INTRODUCTION

The Sun is our most plentiful energy source, but as periods of abundant sunlight do not always match periods of demand, an important challenge for efficient exploitation of solar energy is storing it. One approach is the molecular solar−thermal (MOST) system,[1−5] where solar energy is stored in chemical bonds through light-induced isomerization of photoactive molecules (see Figure 1). Through a certain trigger, the stored energy is released as heat and the high-energy isomer returns to its original state. This makes for a closed cycle of solar energy harvesting, storage and release, without emission of CO_2 or other chemical byproducts.

A candidate system should preferably have a high energy storage density (with an upper limit of 1 MJ/kg for these systems),[6] be stable in the high-energy isomer for days or weeks, and absorb close to the peaks of the solar spectrum (≈500 nm). A promising system for use as a solar heat battery is the dihydroazulene/vinylheptafulvene (DHA/VHF) photoswitch. The dicyano parent system is shown in Figure 1. It absorbs well into the solar spectrum and has a high quantum yield $\Phi_{DHA\rightarrow VHF}$ = 0.55 in acetonitrile.[7] However, the energy density is ≈0.1MJ/kg, and the VHF half-life is only on the order of hours.[8] These properties can be tuned and improved through the use of electron donor−acceptor substituents on the azulene ring.[9−13] Due to the sheer number of combinations of substituents and substituent positions, the balancing act in improving one property without diminishing others, along with the lack of clear systematic guidelines for DHA/VHF functionalization presents a challenge to standard computational chemistry approaches. This is both in terms of the

Received: January 16, 2022
Revised: February 18, 2022
Published: March 4, 2022



https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

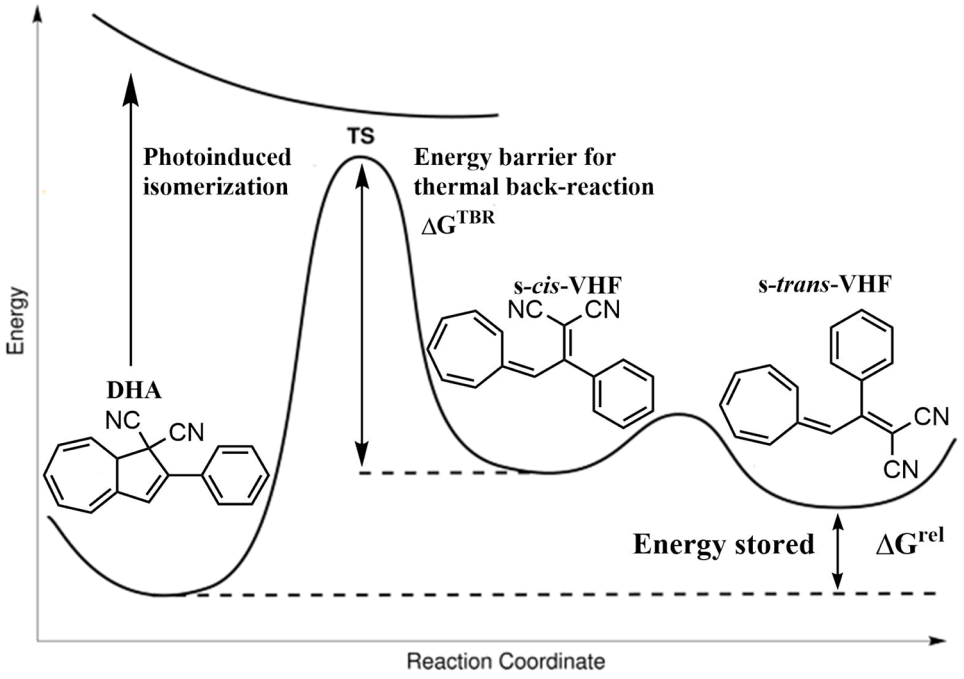


Figure 1. Energy level diagram of the dihydroazulene/vinylheptafulvene (DHA/VHF) solar heat battery. DHA photoconverts to to the high-energy s-*cis*-VHF, which is in thermal equilibrium with the more stable s-*trans*-VHF. The stored chemical energy (ΔG^{rel}) is released as heat upon a trigger. The VHF half-life is determined by the energy barrier (ΔG^{TBR}) for the thermal back-reaction (TBR).

immense computation time for calculating all systems and in terms of separating the wheat from the chaff and finding the promising candidates.

In this study, we use computational chemistry methods and develop a machine learning model to investigate the thermochemical and dipole properties of tens of thousands of DHA/VHF derivatives with electron-acceptor and electron-donor substituents on the azulene ring, in order to improve the system for use as a solar heat battery. The computational complexity is not only directly tied to the system size but also to the energy landscape traversed by the chosen optimization routine. For a single, small, simple compound, this is not an obstacle, as a calculation may be performed in a matter of seconds to hours depending on the level of theory, and it may be terminated manually if convergence becomes unlikely. For a large database of such systems, the time complexity can quickly become unmanageable, even on a high-performance cluster. On the contrary, a machine learning model bears a time complexity tied almost exclusively to the size of the model, and a small scale model can be run on even low-end personal computers, with good-to-reasonable speed. It thus constitutes a valuable supplement to problems of this nature.

With traditional approaches to compounds such as the DHA/VHF system, one would calculate the properties of a few systems and then do a detailed analysis and discussion of what chemical changes led to what changes in the properties. By exploring the full chemical space of synthetically accessible DHA/VHF systems with simple substituents, we are instead able to predict the best individual systems among thousands upon thousands of candidates within our broad feature space. Instead of deeply analyzing the chemical properties of a few systems, we are instead using deep learning to analyze the energy and dipole properties of all systems within our group, selecting the most promising solar heat battery candidates.

For the thermochemical properties, the DHA/VHF system's energy storage capacity ΔG^{rel} is approximated as the electronic energy difference between the DHA and lowest-energy corresponding VHF structure, $\Delta E^{rel} = E_{VHF} - E_{DHA}$. The thermal back-reaction (TBR) barrier ΔG^{TBR} is a measure of how long the system can hold the energy, and is calculated as the Gibbs energy difference between the transition state and corresponding *s-cis*-VHF, $\Delta G^{TBR} = \Delta G_{TS} - \Delta G_{s-cis-VHF}$. Our model predicts the barrier for the elementary step from the *s-cis*-VHF to the transition state.

For the dipole properties, we are investigating the individual dipoles of the DHA, VHF and the transition state (TS) molecules, but also the dipole analogues to the energetic properties, meaning the *dipole storage* $\Delta p^{rel} = p_{VHF} - p_{DHA}$ and *dipole barrier* $\Delta p^{TBR} = p_{TS} - p_{s-cis-VHF}$. This gives insight into the solvation energy of the photoswitches, the projected solubility of the compounds, and how the thermochemical properties change with the addition of solvent or increasing solvent polarity. It also indicates the degree of physical adsorption between the DHA/VHF system and nanoparticles, suggesting the magnitude of the nanoparticle−photoswitch dipole moment interaction. As we are interested in working with polar solvents, we generally want the dipole moments to be higher than for the parent system. It is also ideal for the DHA to have a higher dipole moment than VHF, as DHA will then be more stabilized than VHF in increasingly polar solvents, increasing the energy storage capacity. The same thing holds true for s-*cis*-VHF and the transition state, as this will increase the TBR barrier value. In practice, the VHF usually has a higher dipole moment than the DHA due to greater conformational freedom, while the transition state usually has a higher dipole moment than the s-*cis*-VHF due to greater zwitterionic character, meaning that the challenge becomes to not have a significant difference in dipole moment between DHA and VHF for energy storage capacity and

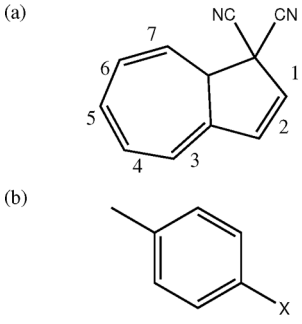
https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

between s-*cis*-VHF and TS for TBR barrier optimization in polar solvents.

Previous work has been done on a similar database, focusing directly on the thermochemical properties.[14] Our focus here is predicting the dipole moments and how they relate to the thermochemical properties.

METHODS

Semiempirical Calculations. To create the data set, 41 different substituents (comprised of 20 functional groups plus their para-substituted phenyl variant) and 7 substituent positions were chosen; see Figure 2. The method used to generate the structures was similar to the one performed in the study by Kromann et al.[15] and Koerstz et al.[14] All combinations of singly and doubly substituted compounds were generated from SMILES strings using RDkit.[16] This leads to a total of $(41{\cdot}7) + 41^2\sum_{n=1}^{6} n = 35\,588$ systems (each including a DHA, VHF and TS structure). A simple conformational search was then performed, again using RDkit, consisting of generating $5 + 5n_{rot}$ start geometries, where n_{rot} is the number of rotatable bonds. DHA and VHF structures were optimized using the GFN2-xTB method, while VHF and TS structures were optimized using the PM3 method. For the transition states, the highest energy structure for each system is used as a starting point for a transition state (TS) search using PM3[17] in Gaussian16,[18] calculating the Hessian in each step. The transition state was then verified using a script, by checking whether its imaginary frequency corresponds to the normal mode lying along the reacting C−C bond. This transition state connects the s-*cis*-VHF molecular structure with the DHA. The lowest-energy VHF conformer is usually s-*trans*-VHF, so it is here assumed that s-*cis*-VHF and s-*trans*-VHF are in thermal equilibrium, i.e., that the energy barrier between s-*cis*-VHF and s-*trans*-VHF is much lower than that between DHA and s-*cis*-VHF. Of the 35 588 systems, 32 623 converged.



EWG	EDG
-[F, Cl, Br]	-OH
$-CF_3$	-OMe
-CN	$-NH_2$
$-NO_2$	$-NMe_2$
-CHO	-Me
$-CO_2H$	-NHC(O)Me
-C(O)Me	-SMe
$-C(O)NH_2$	-
-CCH	-
$-SO_2Me$	-
-CH=NH	-

Figure 2. List of tested substituents (not including phenyl and hydrogen), sorted according to whether they are electron-withdrawing (EWG) or electron-donating groups (EDG). *Me* is shorthand for methyl. Part a shows the azulene ring with the seven tested substituent positions. Part b shows the para-substituted phenyl substituents, where X is a substituent in the table.

Only singly and doubly substituted systems were considered here. This is due to them being the most synthetically accessible. Furthermore, as the number of substituents on the azulene ring increases, steric effects become much more pronounced, and intermolecular reactions between functional groups become much more likely, for example, between amino and aldehyde groups.

The energy storage capacity ΔG^{rel} is calculated using the semiempirical quantum method GFN2-xTB,[19] where it is approximated as the difference in electronic energy ΔE^{rel} between the DHA and the corresponding VHF conformer. The TBR barrier ΔG^{TBR} is calculated using PM3[17] in Gaussian16[18] as the difference in Gibbs free energy between the transition state and the corresponding s-*cis*-VHF structure. This is because only s-*cis*-VHF, and not s-*trans*-VHF, has the correct stereochemistry to rejoin the ring and convert back to DHA. For the converged transition states, their vibrational frequencies were calculated to ensure that the structure corresponded to a first-order saddle point and that the normal mode associated with the imaginary frequency corresponds to the ring-opening/ring-closing of the DHA/VHF system.

Convolutional Neural Network Algorithm. In the original data structure, each system is described by a positional matrix with an integer value to represent the substituent. There is no correlation between this value and the substituent

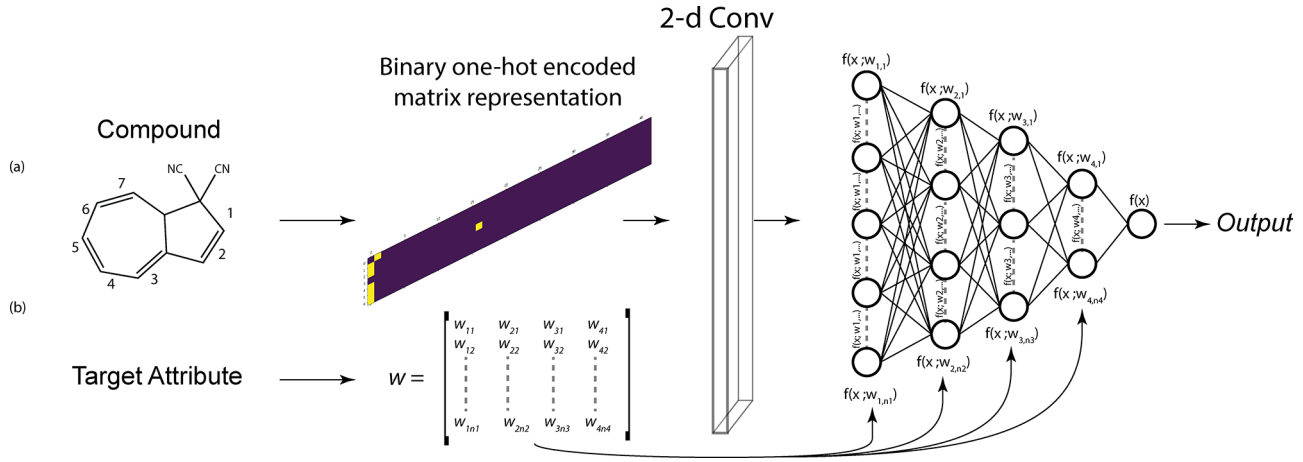


Figure 3. Basic flow of the modular neural network algorithm. The input takes a compound and a target attribute; the compound is transformed into a binary representation, which is encoded into a matrix such that each substituent position is associated with a row; the target attribute initializes retrieval of a set of stored weights and associated layer shapes. For each target attribute, the network can reconfigure the underlying shape of the dense layers, modify drop-out rates (for training only), and it can add or subtract filters from the convolutional layer. The stored weights include an output scaler that transforms the output value into a meaningful property.

https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

characteristic, so to minimize number bias, we transform the data via seven-hot encoding, creating a vector with 7 × 42 binary elements, where each chunk of 42 is a one-hot encoded representation of a tested substituent (see Figure 2) plus H at a specific position on the azulene ring. This is a common method for categorical data. Because each substituent may be replicated at seven different positions, we considered that a matrix representation might better translate the positional influence of each substituent, to which a convolutional layer could possibly outperform a more straightforward neural network (NN).

The NN algorithm is constructed in two parts and is outlined in Figure 3. The macro layer consists of the overall node architecture of the NN itself and a sublayer that stores the node weights. It receives the one-hot encoded system representation as input and a target attribute. For each target attribute, the network accesses a stored set of weights for every node to a trained set for that specific target attribute. Each weight set includes the ability to modify the underlying shape of each layer, change the dropout rates (for training only), and subtract or add filters to the 2-dimensional convolutional layer. The downside to this is that initial creation of the weight set requires individual training for each target attribute. However, given that a single weight set can be computed in a matter of minutes, we do not see this as an obstacle in the current state. The initial architecture was a sequence of four dense node layers and a single output unit, with each sequential layer having 256, 128, 64, and 16 nodes, respectively. We use the terminology *dense* to indicate that each node in a layer is connected to each node in the next layer (that is to say they are fully connected). For the standard NN variant an input layer of size 294 is used to account for all possible input features. For the convolutional variant, a 2-dimensional convolutional layer is prefixed to the dense layer section, replacing the standard input layer.

Because the underlying structure of both convolutional and dense layers can be considered unique hyperparameters, we have implemented a combination of a randomized search with a cross-validation routine.[20] Each routine was independently run and allowed to compete, and the best scoring set of parameters were probed further in a narrow region using the randomized search routine. These optimizations were run using the training set exclusively. We will refer to the three network architectures as DenseNN, ConvNN, and the OptConvNN (optimized ConvNN). For model training, the data set was split 80:20 into training and test set respectively, with a 5-fold cross-validation procedure for the training and validation set. Attribute training data is processed via a normalization scaler, and each scaler is implemented as an optional suffix to the NN output. Model training used the log cosh loss function with the absolute error |*E*| as a metric for minimization:

$$\text{loss}(|E|) = \sum_{i=1}^{n} \log(\cosh(|E_i|)) \tag{1}$$

where $E = ((y_1 - x_1), (y_2 - x_2), ..., (y_i - x_i))$ is the point-value error vector of true y and predicted x values in the training set. A stochastic first-order gradient descent algorithm was utilized as the optimizer.[21] As the ML approach is made to directly compete with quantum chemical computations, it is relevant to consider training and optimization time. For this reason, the networks were purposefully small, and trained using small batch sizes, early stopping conditions, and reduced learning rate conditions. The full parameter overview can be found in the Data Availability.

RESULTS AND DISCUSSION

Each of the three neural network architectures was trained for all target attributes. Table 1 summarizes the mean and standard deviation on the error ϵ for two select target attributes. as calculated by $\epsilon = (y_{\text{predict}} - y_{\text{true}})/y_{\text{true}}$, the relative error is thus a unitless quantity corresponding to a fraction of the total value, and is a direct measure of performance. When reporting the mean and median error, we calculate the relative error of the absolute residual, to avoid averaging between both positive and negative values, which would lead to a falsely improved performance. The addition of the 2-dimensional convolutional layer alone reduces the error by an order of magnitude, and the value is further refined by optimizing the underlying network architecture. For the VHF energy (Figure 4A), the DenseNN exhibits a multimodal distribution, suggesting that some systems are grouped into distinct subsets that decrease the DenseNN accuracy power substantially. DHA similarly exhibits a negative trailing distribution, though less distinctly grouped, with straggling outliers reaching above 10% error. Upon addition of the convolutional layer alone both distributions are immediately narrowed and refined (Figure 4, B and C), achieving standard deviations $\sigma_{std} < 1\%$. We interpret this to mean that the Convolutional Neural Network conveys clearer information of substituent type and position by the addition of the Convolutional layer alone. Yet the standard Neural Network still falls within a reasonable relative error. Optimization of the ConvNN dense layer architecture further refines these results, corrects the outliers, and in both cases reduces the standard deviation by more than 40%, narrowing the distribution around the peak at 0.

Table 1. Progression of Prediction Accuracy with the Addition of a 2D Convolutional Layer[a]

	$\lvert\epsilon\rvert \pm \sigma_{\text{std}}$	
attribute	VHF energy	DHA energy
DenseNN	0.0365 ± 0.0320	0.0488 ± 0.0361
ConvNN	0.0066 ± 0.0069	0.0096 ± 0.0089
OptConvNN	0.0046 ± 0.0040	0.0039 ± 0.0032

[a]We observe a significant refining of the prediction accuracy when the 2-dimensional convolutional layer is added. Further refinement, though less drastic, is possible with optimization of the underlying network architecture. Error is calculated as $|\epsilon| = (|y^{\text{predict}} - y^{\text{true}}|)/y^{\text{true}}$. Standard deviation σ_{std} is calculated for the absolute value of the residuals.

With the hyperparameter optimized convolutional neural network, the required sample size is not only small, but requires little training time. In our test, training the OptConvNN on 20 000 systems required 4 min on an 8 thread system with 16 GB of memory. Figure 4D shows the training set size dependency for the DHA energy, calculated using GFN2-xTB. We see that the loss function converges around 15 000 systems, or around 50% of the full data set. It is expected that larger data sets result in better accuracy of the model, but it is of interest to reduce the number of required computations before a model is *good enough* to be used for qualitative predictions. In our case, a sufficient accuracy for qualitative prediction is achieved as early as 8000 systems,

https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

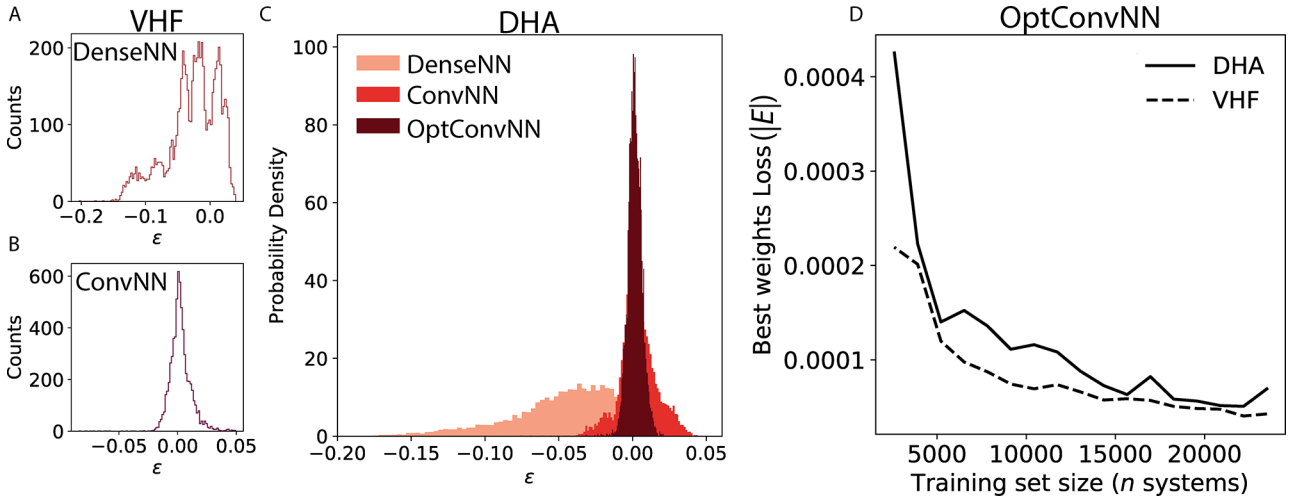


Figure 4. Neural network error distribution and training set size dependency. Error $\epsilon = (y_{\text{predict}} - y_{\text{true}})/y_{\text{true}}$ for (A and B) the VHF and (C) the DHA energy predicted from the test set. DenseNN is the Dense Neural Network, ConvNN is the Convolutional Neural Network and OptConvNN is the Convolutional Neural Network with hyperparameter optimization. (D) log cosh loss function at best epoch for the OptConvNN training on the DHA and VHF energy as a function of the number of systems in the training set.

Table 2. Overview of Relative Absolute Error $|\epsilon| = (|y_{predict} - y_{true}|)/y_{true}$ (Mean and Median) on the Test Set for Investigated Dipole Moments, *D*, in Debye as Target Parameters and Energy Values, *E* (Gibbs-Free for PM3, Electronic for GFN2-xTB Calculations), in kJ/mol[a]

target	$\epsilon_{D,\text{mean}}$	$\epsilon_{D,\text{median}}$	$\epsilon_{E,\text{mean}}$	$\epsilon_{E,\text{median}}$
DHA, GFN2-xTB	0.096	0.046	0.004	0.009
VHF, GFN2-xTB	0.111	0.054	0.005	0.007
VHF, PM3	0.135	0.072	0.012	0.010
TS, GFN2-xTB	0.042	0.010	–	–
TS, PM3	0.031	0.003	0.025	0.016
storage, GFN2-xTB	1.114 (0.147)	0.185	0.042 (0.006)	0.028
barrier, PM3	0.932 (0.139)	0.144	0.017 (0.028)	0.011

[a]Values in parentheses are achieved via error propagation rather than direct optimization of the parameter.

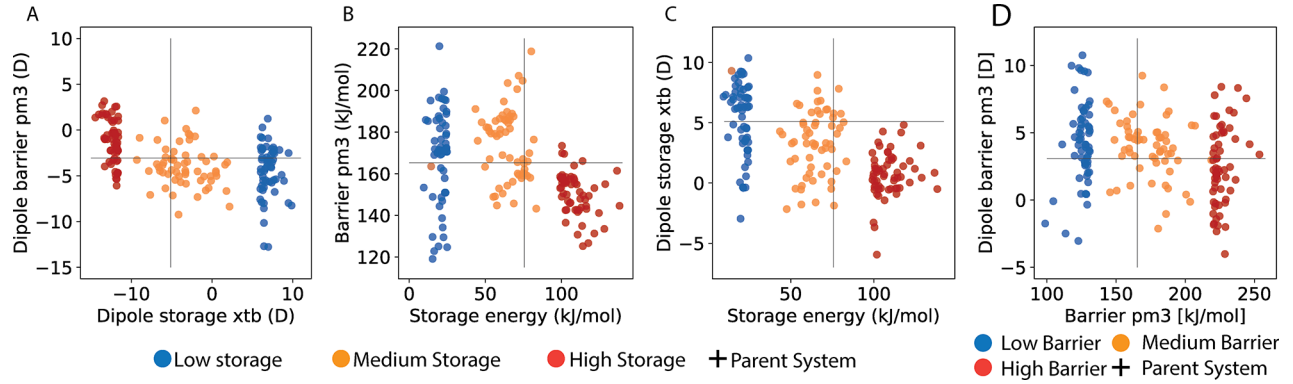


Figure 5. Select attributes for the top 50, bottom 50, and 50 randomly selected molecules. (A) Dipole storage (GFN2-xTB) vs dipole barrier (PM3) for the top 50, bottom 50 and 50 randomly selected molecules. (B) Energy storage ((GFN2-xTB)) vs TBR barrier (PM3). (C) Energy storage ((GFN2-xTB)) vs dipole storage (GFN2-xTB). (D) TBR barrier (PM3) vs dipole barrier (PM3).

corresponding to less than 25% of the data set. This indicates that expanding the data set, for example by including new substituents or substituent positions, will not require many new calculations for the model to the able to predict on the new type.

Hyperparameter optimization of the ConvNN revealed expectedly different dense layer node numbers for the three tested target attributes. Though constrictions were applied such that the general four-layer architecture could not be changed, the node count in each layer could be varied freely together with dropout rates, and we typically observed increased node counts in the third and sometimes fourth layer, while simultaneously seeing increasing dropout rates. In no cases did we see a model that kept the initial descending node arrangement as outlined in the example in Figure 3. Because of the clear advantage to interpreting the systems via a data transformation and convolutional layer processing, we present further target attributes derived via this model only.

https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

Table 2 shows the relative errors |ϵ| for select components estimated by the OptConvNN and validated on the true values.

These errors are small enough that energy and dipole properties can be qualitatively predicted with significant accuracy, allowing for usage of this method in selecting only the most promising DHA/VHF systems for further analysis as solar heat battery candidates. We note that our NN model is better at predicting the energy and dipole properties of individual DHA and VHF structures, rather than trying to directly predict the difference between two conformers in the form of the energy/dipole storage or barrier. For example, the mean dipole storage relative error for the xTB calculations is 1.1143, where the model trains directly on the VHF dipole value minus the DHA dipole value. However, calculating the same value via error propagation, i.e., training on the VHF and DHA dipole values separately, then subtracting the predicted DHA dipole from the corresponding predicted VHF dipole and propagating the errors, gives only a mean dipole storage relative error of 0.1464. Another thing to note is that our ϵ on the dipole properties are significantly higher than those of the energy properties. This is primarily due to the dipole moment varying to a greater extent than the energy, being very sensitive to conformational changes, and thus making it harder to predict. Nevertheless, these errors are already much smaller than what is needed for qualitative predictions on whether a system will perform better or worse in polar solvents. In Figure 5, from our selection of 32 623 optimized molecular systems, we plot 50 random molecules and then the top and bottom 50 in terms of the x-axis property, either dipole storage, energy storage capacity, or energetic TBR barrier values. This shows the value of a large statistical basis in that we are easily able to find and select promising DHA/VHF candidates with high reaction energies, barrier heights or dipole shifts for further optimization and research at a higher level of theory. One problem one often runs into when designing photoswitch derivatives is that optimizing one property, such as the storage capacity, leads to a decrease in another property, such as the TBR barrier. Of particular interest for this is the first quadrant, where we manage to improve two separate properties simultaneously. Here we have results for all tested combinations, meaning that these systems improve on the two tested properties simultaneously and thus may be worth further study in terms of photoswitch optimization. Figure 5C shows only a few systems that manage to have both a higher storage energy and dipole storage compared to the parent system, and none are among the top candidates. However, testing all the systems rather than just the 150 shown on the figure will show a greater amount of candidates. In the same vein, Figure 5D shows a great amount of systems with both higher TBR barrier values and higher dipole barrier values than the dicyano parent system. Thus, there seems to be a large amount of potential candidate systems with a longer VHF half-life than the parent, but which keep or even improve on that longevity in polar solvents. For reference, according to transition state theory, a 10 kJ/mol increase in the activation energy corresponds to a 57-fold increase in half-life, going from roughly 3 h to 2 days for the parent (depending on solvent) to around 8 days to 3 months, which is a VHF half-life enough for many practical applications. Parts A and B of Figure 5 show the same trends, with some promising systems in the first quadrant that manage either both higher dipole storage and dipole barrier, or both higher energy storage and TBR, than the parent system.

A common ML metric is the generalization to novel input types. As a proof of concept of the adaptability of the ML model, we tried to get it to predict the values of triply and quadruply substituted systems with the same possible combinations of substituents and positions as our data set. These types of systems are not part of the training set and thus are something the model has never encountered before. These systems are less synthetically accessible than singly and doubly substituted systems. In addition, they often have problems with steric effects and intermolecular reactions between functional groups due to the amount of substituents on the ring.[14] As they also take much longer to calculate, we restricted our test to 12 systems, 6 of each. These are shown primarily as further proof of the strength of our model, rather than as likely candidates for further optimization.

The horizontal and vertical line mark the values for the parent compound shown in Figure 1.

The relative mean and median errors are shown in Table 3. While these errors are obviously for a small sample of systems and thus not perfectly representative, along with the errors being larger than those of the test set especially for the quadruply substituted systems, they nevertheless show that the model is still capable of making qualitative predictions of systems with no direct analogues in its training set of singly and doubly substituted systems. This indicates the adaptability of the model, and that it would not require a large amount of extra computational calculations for our model to train on before it is capable of qualitatively predicting its properties. This is also strongly indicated by Figure 4D, which shows that the loss function converges quickly and that a small training set size is sufficient, especially for qualitative description. Figure 4D shows the convergence of the loss function with respect to the size of the training set, i.e., how big a fraction of systems out of the possible combinations we would actually need to calculate in order to achieve predictive accuracy. The loss function converges around 15 000 systems, or around 50% of the full data set, although accuracy good enough for qualitative prediction is achieved much earlier. This indicates that the model is robust and capable of predicting the properties of systems without requiring a ton of computations for it to train on. This also indicates that expanding the data set, for example by including new substituents or substituent positions, will not require many new calculations for the model to be able to predict on the new type.

These could include triply substituted systems, a new ligand, or a different substituent position. The latter is perhaps the most interesting, considering that replacing one of the two CN substituents on the DHA/VHF system with a H atom more

Table 3. Overview of Relative Absolute Error $|\epsilon| = (|y_{predict} - y_{true}|)/y_{true}$ (Mean and Median) for Investigated Dipole Moments, D, in Debye as Target Parameters and Energy Values, E (Gibbs-free Energy for PM3 and Electronic Energy for GFN2-xTB Calculations)[a]

target	$\epsilon_{D,\mathrm{mean}}$	$\epsilon_{D,\mathrm{median}}$	$\epsilon_{E,\mathrm{mean}}$	$\epsilon_{E,\mathrm{median}}$
DHA, GFN2-xTB	0.223	0.204	0.143	0.157
VHF, GFN2-xTB	0.269	0.253	0.142	0.148
storage, GFN2-xTB	(0.302)	(0.369)	(0.202)	(0.216)
barrier, PM3	0.183	0.158	–	–

[a]Values in parentheses are achieved via error propagation rather than direct optimization of the parameter.

https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688

than doubles the energy storage density compared to the dicyano parent, and increases the TBR barrier to the extent that the back-reaction is effectively put on hold under standard conditions.[12,22] This would allow us to consider substituents that tend to increase the energy storage density and decrease the activation energy, with the latter actually being beneficial in order for the back-reaction to occur.

We observe based on Figure 5C a correlation between small dipole storage values and large storage energies. This shows that when designing MOST devices based on DHA/VHF one should ensure that the change of dipole moments from DHA to VHF should be as low as possible. We note that electron-withdrawing groups and electron-donating groups have to be placed at specific positions on the DHA/VHF system in order to ensure that. We have observed large storage energies for electron withdrawing groups at position 3 and electron donating groups at position 2. Based on Figure 5D we see that large energy barriers correlate with dipole barriers between 0 and 2. We note that we are able to obtain DHA/VHF systems with large energy barriers when electron withdrawing groups are placed at position 2 and electron donating groups are placed at position 1.

CONCLUSION

In conclusion, a database consisting of 32 623 DHA/VHF derivative systems were created, and it was used to train and validate a small-scale NN algorithm for prediction of dipole moments and energies. The NN was trained to determine specific molecular attributes. We find that the addition of a convolutional layer coupled with a data transformation to a binary matrix representation yields a significant refining of the target attributes, and an increase in accuracy by an order of magnitude in the best case scenario, with mean absolute errors ranging from <1% to 15%, depending on the property. This is low enough that qualitative predictions can select the most promising candidates. The worst case scenario was no reduction in error from the standard NN, and at no point did we observe worse performance by the addition of the convolutional layer. The NN subarchitecture was optimized with machine learning methods. On this basis, we were able to exclude roughly 95% of the candidates, which is directly analogous to reducing the computational requirements by the same percentage. We expect that a mutable NN architecture would be beneficial for a large chemical component space, as there are significant differences in the correlation between the many attributes. Having trained the convolutional NN only on single- and double-substituted systems, we employed it to predict the properties of triple- and quadruple substituent systems, and it performed well enough that a qualitative prediction should be useful. These systems are notoriously more difficult to compute, and could therefore be of special interest in terms of rapid property estimation. The success of the NN on these novel type systems seems to us a promising signal for applying ML to problems of this nature. This has use in prescreening molecules and finding promising MOST candidates, both from within the database and by using the database to predict properties of as of yet uninvestigated DHA/VHF derivatives. This database and algorithm are readily available for future work in prescreening and selecting promising MOST candidates for further refinement and for qualitatively calculating thermochemical and dipole properties for predicting solvent effects.

AUTHOR INFORMATION

Corresponding Author

Kurt V. Mikkelsen − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark;* orcid.org/0000-0003-4090-7697; Email: kmi@chem.ku.dk

Authors

Oliver Christensen − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark*

Rasmus Dalsgaard Schlosser − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark*

Rasmus Buus Nielsen − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark*

Jes Johansen − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark*

Mads Koerstz − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark*

Jan H. Jensen − *Department of Chemistry, University of Copenhagen, 2100 Copenhagen, Denmark;* orcid.org/0000-0002-1465-1010

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.2c00351

Author Contributions

†O.C. and R.D.S. contributed equally to this work.

Notes

The authors declare no competing financial interest.

Data Availability. The electronic structures generated and/or analyzed in the current study are not included, due to the large size of the data set. It is available from the corresponding author upon reasonable request. Supporting data are available online at the University of Copenhagen Electronic Research Data Archive, https://sid.erda.dk/sharelink/BOq5IYk6Zx

ACKNOWLEDGMENTS

K.V.M. acknowledges the Danish Council for Independent Research, DFF-0136-00081 B and the European Union's Horizon 2020 Framework Programme under Grant Agreement Number 951801 for financial support.

REFERENCES

(1) Moth-Poulsen, K.; Ćoso, D.; Börjesson, K.; Vinokurov, N.; Meier, S. K.; Majumdar, A.; Vollhardt, K. P. C.; Segalman, R. A. Molecular solar thermal (MOST) energy storage and release system. *Energy Environ. Sci.* **2012**, *5*, 8534−8537.

(2) Moth-Poulsen, K. Molecular Systems for Solar Thermal Energy Storage and Conversion. *In* Nielsen, M. B.: *Organic Synthesis and Molecular Engineering*; Wiley: 2013; pp 179−196;

(3) Lennartson, A.; Roffey, A.; Moth-Poulsen, K. Designing photoswitches for molecular solar thermal energy storage. *Tetrahedron Lett.* **2015**, *56*, 1457−1465.

(4) Wang, Z.; Roffey, A.; Losantos, R.; Lennartson, A.; Jevric, M.; Petersen, A. U.; Quant, M.; Dreos, A.; Wen, X.; Sampedro, D.; et al. Macroscopic heat release in a molecular solar thermal energy storage system. *Energy Environ. Sci.* **2019**, *12*, 187−193.

(5) Quant, M.; Lennartson, A.; Dreos, A.; Kuisma, M.; Erhart, P.; Börjesson, K.; Moth-Poulsen, K. Low Molecular Weight Norbornadiene Derivatives for Molecular Solar-Thermal Energy Storage. *Chemistry* **2016**, *22*, 13265.

(6) Kucharski, T. J.; Tian, Y.; Akbulatov, S.; Boulatov, R. Chemical solutions for the closed-cycle storage of solar energy. *Energy Environ. Sci.* **2011**, *4*, 4449−4472.

(7) Goerner, H.; Fischer, C.; Gierisch, S.; Daub, J. Dihydroazulene/vinylheptafulvene photochromism: Effects of substituents, solvent,

and temperature in the photorearrangement of dihydroazulenes to vinylheptafulvenes. *J. Phys. Chem.* **1993**, *97*, 4110−4117.
(8) Broman, S. L.; Brand, S. L.; Parker, C. R.; Petersen, M. Å.; Tortzen, C. G.; Kadziola, A.; Kilså, K.; Nielsen, M. B. Optimized synthesis and detailed NMR spectroscopic characterization of the 1, 8a-dihydroazulene-1, 1-dicarbonitrile photoswitch. *ARKIVOC* **2011**, *2011*, 51−67.
(9) Mogensen, J.; Christensen, O.; Kilde, M. D.; Abildgaard, M.; Metz, L.; Kadziola, A.; Jevric, M.; Mikkelsen, K. V.; Nielsen, M. B. Molecular Solar Thermal Energy Storage Systems with Long Discharge Times Based on the Dihydroazulene/Vinylheptafulvene Couple. *Eur. J. Org. Chem.* **2019**, *2019*, 1986−1993.
(10) Hansen, M. H.; Elm, J.; Olsen, S. T.; Gejl, A. N.; Storm, F. E.; Frandsen, B. N.; Skov, A. B.; Nielsen, M. B.; Kjaergaard, H. G.; Mikkelsen, K. V. Theoretical investigation of substituent effects on the dihydroazulene/vinylheptafulvene photoswitch: Increasing the energy storage capacity. *J. Phys. Chem. A* **2016**, *120*, 9782−9793.
(11) Olsen, S. T.; Elm, J.; Storm, F. E.; Gejl, A. N.; Hansen, A. S.; Hansen, M. H.; Nikolajsen, J. R.; Nielsen, M. B.; Kjaergaard, H. G.; Mikkelsen, K. V. Computational methodology study of the optical and thermochemical properties of a molecular photoswitch. *J. Phys. Chem. A* **2015**, *119*, 896−904.
(12) Koerstz, M.; Elm, J.; Mikkelsen, K. V. Benchmark Study of the Structural and Thermochemical Properties of a Dihydroazulene/Vinylheptafulvene Photoswitch. *J. Phys. Chem. A* **2017**, *121*, 3148−3154.
(13) Broman, S. L.; Nielsen, M. B. Dihydroazulene: from controlling photochromism to molecular electronics devices. *Phys. Chem. Chem. Phys.* **2014**, *16*, 21172−21182.
(14) Koerstz, M.; Christensen, A. S.; Mikkelsen, K. V.; Nielsen, M. B.; Jensen, J. H. High throughput virtual screening of 230 billion molecular solar heat battery candidates. *PeerJ. Phys. Chem.* **2021**, *3*, No. e16.
(15) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chem. Sci.* **2018**, *9*, 660−665.
(16) Landrum, G. *Rdkit documentation. Release 2013 1.1.79*.
(17) Stewart, J. J. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.
(18) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Petersson, G.; Nakatsuji, H.; et al. *Gaussian 16*; 2016.
(19) Grimme, S.; Bannwarth, C.; Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z= 1−86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(20) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(21) Dozat, T. *Incorporating nesterov momentum into adam, Workshop track - ICLR 2016*; https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ. 2016.
(22) Cacciarini, M.; Skov, A. B.; Jevric, M.; Hansen, A. S.; Elm, J.; Kjaergaard, H. G.; Mikkelsen, K. V.; Brøndsted Nielsen, M. Towards Solar Energy Storage in the Photochromic Dihydroazulene−Vinylheptafulvene System. *Eur. J. Org. Chem.* **2015**, *21*, 7454−7461.

https://doi.org/10.1021/acs.jpca.2c00351
J. Phys. Chem. A 2022, 126, 1681−1688