

Virtual screening of norbornadiene-based molecular solar thermal energy storage systems using a genetic algorithm

Nicolai Ree, Mads Koerstz, Kurt V. Mikkelsen,* and Jan H. Jensen*

*Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100
Copenhagen Ø, Denmark*

E-mail: kmi@chem.ku.dk; jhjensen@chem.ku.dk

Abstract

We present a computational methodology for the screening of a chemical space of 10^{25} substituted norbornadiene molecules for promising kinetically stable molecular solar thermal (MOST) energy storage systems with high energy densities that absorb in the visible part of the solar spectrum. We use semiempirical tight-binding methods to construct a dataset of nearly 34,000 molecules and train graph convolutional networks to predict energy densities, kinetic stability, and absorption spectra and then use the models together with a genetic algorithm to search the chemical space for promising MOST energy storage systems. We identify 15 kinetically stable molecules, five of which have energy densities greater than 0.45 MJ/kg and the main conclusion of this study is that the largest energy density that can be obtained for a single norbornadiene moiety with the substituents considered here, while maintaining a long half-life and absorption in the visible spectrum, is around 0.55 MJ/kg.

Introduction

Today, the production of solar and wind energy have become more profitable than non-renewable alternatives. However, daily and seasonal variations in the renewable energy production as well as large variations in the power demands are two serious challenges for a sustainable energy eco-system. Storing excess power and using this energy in peak times is therefore absolutely crucial,^{1,2} but it requires affordable large-scale energy storage systems. One technology that attempts to solve this problem is closed-cycle **MO**lecular **S**olar **T**hermal (MOST) energy storage.³⁻¹³ Such systems relies on photochromic molecules or molecular photoswitches where stable reactants can interconvert to form metastable products using solar irradiation as the driving force. Thus, the solar energy can be stored as chemical energy until a subsequent exothermic reaction releases the captured energy. Depending on the storage lifetime, this energy release can either occur spontaneously or be controlled by e.g. a thermal activation, a heterogeneous catalyst, an electric potential, or light. To paraphrase, MOST systems are able to harvest and store solar energy, which can later be released as clean thermal energy for space heating or heating of domestic water. In fact, the thermal energy can be stored without the need for thermal insulation, which enables prolonged storage times compared to alternative thermal energy storage systems.

One of the most promising MOST systems is the norbornadiene/quadracyclane (NBD/QC) couple (Figure 1), which was introduced as a potential MOST system in 1961 by Dauben and Cargill,¹⁴ although the isomerization reaction of a NBD/QC dicarboxylic acid system was already observed in 1954 by Cristol and Snell.^{15,16} Since then, the system has been extensively studied due to its high energy density of almost 1 MJ/kg, which is estimated to be the fundamental upper limit of MOST systems and comparable to Li-ion batteries.⁷

However, the absorption spectrum of NBD lies in the UV region with absorption onset at 267 nm, and has therefore no overlap with the spectrum of solar radiation.¹² Several studies have shown that it is possible to obtain a large redshift in the absorption spectrum and better quantum yield, but at the expense of a decrease in the energy density due to the added weight of the chromophore.¹⁷⁻²⁴ For example, in a recent perspective Orrego-Hernández *et al.*²⁴ highlight seven examples of NBD systems with absorption onsets of 362-466 nm, but energy densities of 0.10-0.56 MJ/kg - compared to 0.97 MJ/kg for unsubstituted NBD. Furthermore, four of the seven molecules have half-lives of a week or less, which make them unsuitable for long-term energy storage.

In this study we combine quantum chemical calculations, machine learning, and a genetic algorithm to search chemical space of roughly 10^{26} NBD/QC derivatives defined in Figure 2 for optimal MOST candidates. Our results suggest that the largest energy storage value that can be obtained with these substituents, while maintaining a long half-life and absorption in the visible spectrum, is around 0.55 MJ/kg.

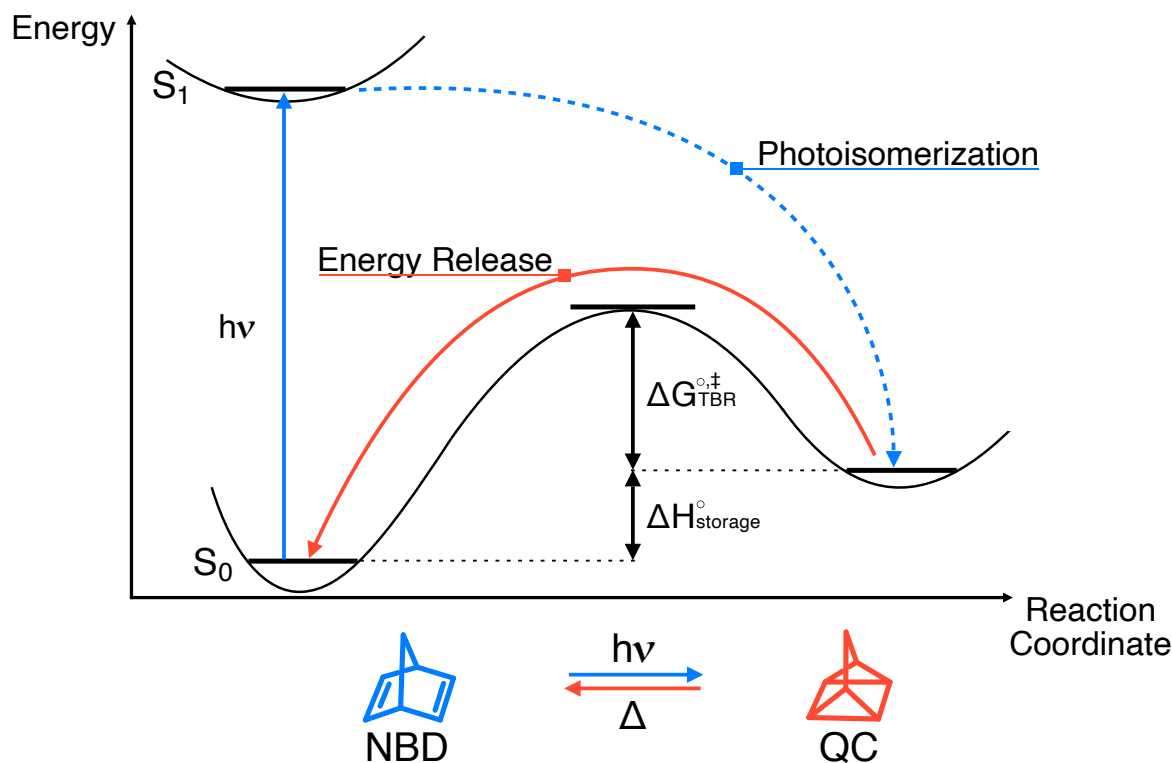


Figure 1: Energy diagram of the NBD/QC system.

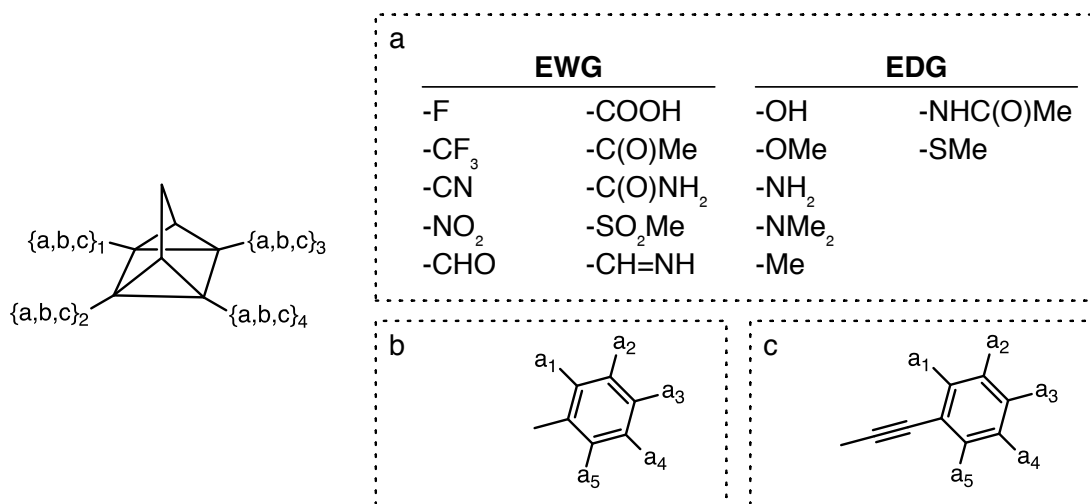


Figure 2: A representation of the chemical space investigated using machine learning, showing that the NBD/QC motif can be functionalized on four positions with three different groups ("a", "b", and "c"). The group "a" includes electron withdrawing groups (EWGs) and electron donating groups (EDGs) as well as hydrogen, and accounts for a total of 18 different substituents. There are roughly $18 + 2(\frac{1}{2}18^5) = 1.9\text{M}$ substituents and $\frac{1}{4}(1.9\text{M})^4 = 10^{25}$ different NBD systems.

Computational Methodology

We show that semiempirical tight-binding methods (SQM) can be used to identify molecules with high energy densities and thermal back reaction (TBR) barriers as well as suitable absorption maxima by benchmarking against DFT calculations and experiment. We use the SQM methods to construct a dataset that provided even coverage of the chemical space of interest and train graph convolutional neural networks to predict energy densities, TBR barriers, and absorption spectra and then use the models together with a genetic algorithm to search chemical space for promising MOST candidates.

Semiempirical tight-binding calculations

$\Delta H_{storage}^\circ$ and $\Delta G_{TBR}^{\circ,\ddagger}$ are approximated as the differences in electronic energy ($\Delta E_{storage}$ and ΔE_{TBR}^\ddagger) computed using GFN2-xTB.²⁵ For each NBD structure the QC structure is automatically generated using RDKit.²⁶ $5 + 5n_{rot}$ random conformations (where n_{rot} is the number of rotatable bonds in the molecule) are then generated for each structure using RDKit and optimized with GFN2-xTB. Optimizations that result in discrepancies between the input and output connectivity are discarded. The lowest energy conformers of NBD and QC are used to compute $\Delta E_{storage}$.

To compute the energy barrier of the thermal back reaction a concerted adiabatic scan is performed for the two breaking CC single bonds of QC-to-NBD, which are constrained to 20 values from 1.5 Å out to 2.2 Å, starting from the QC structure with the lowest energy. The highest energy structure and energy is used as an estimate for the transition state. The absorption spectra of NBD are computed by the sTDA-xTB method²⁷ using the lowest energy GFN2-xTB structure.

The entire process is automated and requires only SMILES strings of NBD derivatives as input (see flowchart in supporting information).

The machine learning model

Three undirected graph convolutional networks (GCNs)²⁸ are trained to reproduce the energy storage, absorption, and TBR barrier. The GCN model is essentially that implemented in DeepChem²⁹ and written in Python 3.6.8 using PyTorch version 1.2.0³⁰ and PyTorch Geometric version 1.3.2.³¹ The feature vectors of the nodes describe the atom type, number of directly bonded neighbors, number of hydrogen atoms attached to the atom, formal charge, hybridization, and aromaticity, while the feature vectors of the edges include bond orders as well as information about conjugation and presence in a ring system. The GCN uses two graph convolutional layers, with 128 channels in the first layer and 64 in the second. After the final graph convolutional layer, a global max-pooling layer creates a representation with 64 values, which is fed to the feed-forward neural network. The feed-forward network has two hidden layers with 64 nodes in each layer. The final layer outputs a single value i.e. the

target value predicted by the network.

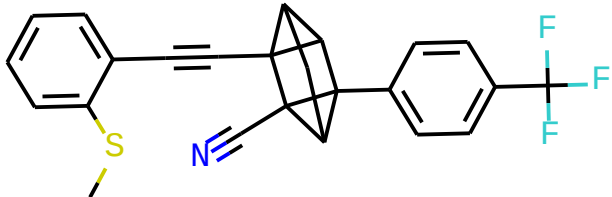


Figure 3: An example of a NBD/QC derivative included in the chemical space defined in Fig. 2. The gene for this system is "[['C',17,0,0,0,0],['A',3,0,0,0,0],['A',0,0,0,0,0],['B',0,0,2,0,0]]", and the smile is "CSc1ccccc1C#CC12C3CC4C(c5ccc(C(F)(F)F)cc5)(C31)C42C#N".

The genetic algorithm

In a previous study³² we have shown that chemical space of MOST candidates can be searched efficiently by genetic algorithms and we apply a similar approach here. The chemical space shown in Figure 2 is encoded by a gene like that shown in Figure 3. Each of the four positions is represented by a label ("A", "B", and "C") followed by five integers ranging from 0 to 17. The labels represent the three types of ligands shown in Figure 2a, b, and c, respectively, while the five integers represent the five different positions on the phenyl rings shown in Figure 2b and c (if the label is "A", only the first digit is read). The GA search generates an initial population of 300 genes corresponding to random singly substituted NBD/QC systems (phenyl and phenylacetylene groups are also singly substituted and random mutations are applied if duplication occurs) after which a new generation of same size is created from crossovers between two parent genes. In the crossover, between one and three ligands from the first parent are randomly selected and combined with the remaining ligands from the second parent to produce a child. Furthermore, random mutations of the children occurs with a mutation rate of 25 %. These mutations allows one of the ligands to be modified by either changing the ligand type ("A", "B", or "C" from Figure 2) or one of the attached "A" ligands. The different parents are selected with a probability that is proportional to their score (roulette selection). This score is computed as sum of three thresholded-linear functions³³ ranging from 0 to 1 with thresholds of 0.6 MJ/kg, {375, 400, 450, 525} nm, and 250 kJ/mol for the energy density, absorption, and TBR barrier, respectively, predicted by the GCN models. To favour a redshift of the absorption spectrum, we multiplied the absorption score by two resulting in a maximum GA score of four. Hereafter, 300 of the worst scoring genes are discarded, which results in a new population of 300 genes. Each GA search runs for 300 generations and the final output is the 50 highest scoring molecules of the final population. We select four different absorption thresholds resulting in four runs of a thousand GA searches to also promote higher energy densities and TBR barriers.

Results and discussion

Benchmarking the tight binding calculations

We benchmark the tight binding calculations against DFT on 31 different NBD/QC systems with promising properties collected from the literature.^{17–20,34} Based on a recent benchmark study by some of us,³⁵ we use M06-2X/6-311+G(d) for the energy density, and the same method is also used for benchmarking the TBR barrier estimates.

Figures S1 and S2 compares the GFN2-xTB storage energy and barriers of the back reaction to the corresponding M06-2X/6-311+G(d) values. While GFN2-xTB significantly underestimates the storage energy and overestimates the barrier heights both properties are strongly correlated with the DFT results, with R^2 values of 0.81 and 0.70 and Pearson’s R values of 0.90 and 0.84, respectively. GFN2-xTB is thus capable of identifying molecules with high energy densities and low barrier heights for further study using DFT.

The sTDA-xTB UV-Vis spectrum of a charge-tagged NBD/QC carboxylate is benchmarked against the experimental result by Ugo *et al.*³⁴ and several DFT calculations (Figure S3). The relevant peak at 315 nm is blueshifted by just 10 nm by sTDA-xTB. In fact, sTDA-xTB performs better than 13 out of the 16 DFT functionals tested in a benchmark study,³⁵ while requiring only a few seconds instead of several hours.

Training the machine learning models

For the dataset, different singly and doubly substituted NBD/QC systems are selected to get an even coverage of chemical space where every small ligand (Figure 2a) is represented at every position on the phenyl rings and NBD scaffold a roughly equal number of times. The singly substituted NBD/QC systems include all directly attached EWGs and EDGs as well as phenyl and phenylacetylene groups with all combinations of up to three EWGs and EDGs in the ortho-, meta-, and para-position (resulting in a total of $18^3 \cdot 2 + 18 = 11,682$ singly substituted systems). The doubly substituted NBD/QC systems (substituted in position 1 and 2, 1 and 3, or 1 and 4, see Figure 2) included all combinations of the EWGs, EDGs, and single ortho-, meta-, or para-substituted phenyl and phenylacetylene groups (resulting in a total of $\binom{((17 \cdot 3 + 1) \cdot 2 + 17) + 2 - 1}{2} \cdot 3 = 22,143$ double substituted systems). Hence, the dataset consisted of 33,825 unique NBD/QC systems. Of these, 119 of the GFN2-xTB calculations failed and are omitted from the dataset. This results in a total of 33,706 NBD/QC systems, which are split 80/20 for training and test set, respectively. A 5-fold cross-validation is used to train the different GCNs for 100 epochs with a batch size of 512 and using the Adam optimizer³⁶ with a learning rate of 0.01 on a MSE loss (see learning curves in supporting information). To prevent overfitting, the trained GCN with the overall best validation loss for each property are saved.

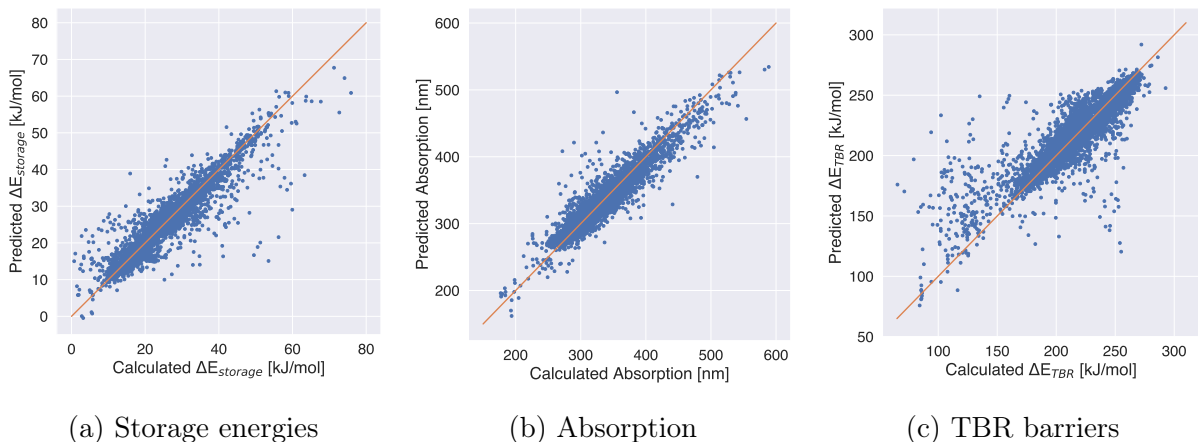


Figure 4: Predicted storage energies, absorption wavelengths for the first absorption peak in the electronic absorption spectrum of the NBD isomer, and thermal back-reaction (TBR) barriers obtained by graph convolutional networks (GCNs) versus calculated results at the GFN2-xTB level of theory. The GCNs was trained on 26,965 NBD/QC systems using 5-fold cross-validation and tested on yet another 6,741 NBD/QC systems for which the results are presented in (a), (b), and (c). The orange line represents the line of equality.

Figure 4 shows the GCN-predicted vs GFN2-xTB values for the storage energy, absorption, and TBR barriers for the test set. The respective mean absolute errors are 1.66 ± 2.46 kJ/mol, 8.62 ± 10.41 nm, and 7.62 ± 11.17 kJ/mol. In the case of the TBR barriers the most significant deviations are for molecules with low barriers, meaning that the GCN predictions could lead to a significant number of false positives with respect to the barrier. However, such false positives can be efficiently eliminated by subsequent GFN2-xTB calculations and does not present a significant problem.

Identifying promising candidates using a genetic algorithm

The 50 top-scoring molecules from the final populations of 4,000 GA searches are selected, which results in 1,234 unique NBD/QC systems. The energy storage, TBR barrier height, and absorption spectra of these molecules are then recomputed using GFN2-xTB and sTDA-xTB, respectively. Then, the 230 systems with energy densities, absorption spectra, and TBR barriers above 0.2 MJ/kg, 350 nm, and 160 kJ/mol, respectively, are reoptimised at the M06-2X/6-311+G(d) level of theory (using the lowest energy GFN2-xTB structures as starting point) followed by vibrational analysis to ensure correct convergence. TDDFT calculations on all 230 molecules reveal that 116 molecules have absorption maxima above 350 nm. The energy densities for these 116 molecules are then computed and the 38 molecules with energy densities higher than 0.4 MJ/kg are selected for barrier computations. The TS search fail for four of the 38 molecules while the TBR barrier is higher than 150 kJ/mol for 15 of the 34 molecules (shown in Table 1).

Label	Structure	$\Delta H_{\text{storage}}^{\circ}/M$ [MJ/kg]	λ [nm]	$\Delta G_{\text{TBR}}^{\circ,\ddagger}$ [kJ/mol]
1		0.56	357	162
2		0.55	361	165
3		0.49	378	163
4		0.48	353	167
5		0.46	358	157
6		0.43	358	161
7		0.43	376	167
8		0.42	359	159
9		0.42	371	219
10		0.41	386	236

11		0.41	426	155
12		0.41	352	160
13		0.41	412	160
14		0.41	397	153
15		0.40	376	153
		0.70	204	162

Table 1: All NBD/QC systems with properties above 0.4 MJ/kg, 350 nm, 150 kJ/mol for the energy density, absorption, and TBR barrier with respect to calculations at the M06-2X/6-311+G(d) level of theory. Moreover, the calculated properties of the NBD/QC parent system at the same level of theory are given as a reference in the bottom of the table.

Five of the molecules have energy densities greater than 0.45 MJ/kg, which is similar to the highest values observed so far for NBD-dimer systems and considerably higher than any previously observed for NBD-monomers.²⁴ The absorption spectra of these five molecules are shown in Figure 5. However, most of the molecules have one or more amine and hydroxy groups directly attached to the NBD scaffold and may undergo keto-enol or imine-enamine tautomerisation, which may adversely impact the properties. The main conclusion of this study is that the largest energy storage value that can be obtained for a single NBD/QC moiety with these substituents, while maintaining a long half-life and absorption in the visible spectrum, is around 0.55 MJ/kg.

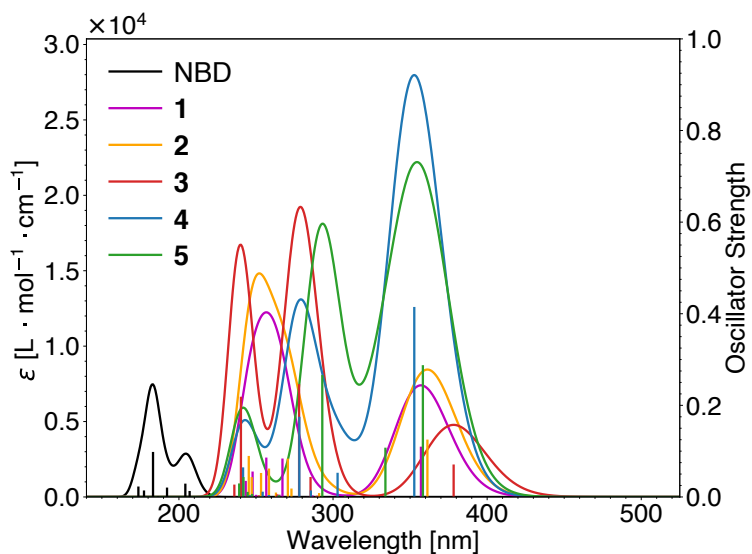


Figure 5: UV-Vis representation of the NBD isomers with energy densities greater than 0.45 MJ/kg (**1-5**). Furthermore, the UV-Vis spectrum of the parent NBD isomer is given as a reference (NBD). All the spectra are obtained using the time-dependent analog of M06-2X/6-311+G(d) and simulated using Eq. S1 from the supporting information.

Conclusions

We present a computational methodology for the screening of a chemical space of 10^{25} substituted NBD molecules for promising MOST systems with high energy density and TBR barrier that absorb in the visible part of the solar spectrum. We show that semiempirical tight-binding methods (SQM) can be used to identify molecules with high energy densities and thermal back reaction (TBR) barriers as well as suitable absorption maxima by benchmarking against DFT calculations and experiment. We use the SQM methods to construct a dataset of nearly 34,000 molecules that provided even coverage of the chemical space of interest and train graph convolutional networks to predict energy densities, TBR barriers, and absorption spectra and then use the models together with a genetic algorithm to search chemical space for promising MOST candidates.

The 50 top-scoring molecules from the final populations of 4,000 GA searches are selected, which results in 1,234 unique NBD/QC systems. Then M06-2X/6-311+G(d) TDDFT calculations are performed on the 230 systems with SQM-predicted energy densities, absorption spectra, and TBR barriers above 0.2 MJ/kg, 350 nm, and 160 kJ/mol, respectively. TDDFT calculations on all 230 molecules reveal that 116 molecules have absorption maxima above 350 nm. The energy densities for the 116 with absorption maxima above 350 nm are then computed and the 38 molecules with energy densities higher than 0.4 MJ/kg are selected for barrier computations. The final results are 15 molecules with a TBR barrier higher than 150 kJ/mol.

Five of the molecules have energy densities greater than 0.45 MJ/kg and the main conclusion of this study is that the largest energy density that can be obtained for a single NBD/QC moiety with these substituents, while maintaining a long half-life and absorption in the visible spectrum, is around 0.55 MJ/kg.

Data Availability

Additional figures and tables can be found in supporting information. The code and data are available at <https://sid.erda.dk/sharelink/DeRV97z1Nz>

Acknowledgement

NR thanks H.C. Ørsted Selskabet and Ørsted A/S for financial support in terms of the Ørsted Scholarship 2018. KVM thanks the Center of Exploitation of Solar Energy for financial support and FTP.

References

- (1) Rugolo, J.; Aziz, M. J. *Energy Environ. Sci.* **2012**, *5*, 7151–7160.
- (2) Duffy, P.; Fitzpatrick, C.; Conway, T.; Lynch, R. P. *Issues in Environmental Science and Technology*; Royal Society of Chemistry, 2018; pp 1–41.
- (3) Yoshida, Z. *J. Photochem.* **1985**, *29*, 27–40.
- (4) Bren, V. A.; Dubonosov, A. D.; Minkin, V. I.; Chernoiyanov, V. A. *Russ. Chem. Rev.* **1991**, *60*, 451–469.
- (5) Dubonosov, A. D.; Bren, V. A.; Chernoiyanov, V. A. *Russ. Chem. Rev.* **2002**, *71*, 917–927.
- (6) Kanai, Y.; Srinivasan, V.; Meier, S. K.; Vollhardt, K. P. C.; Grossman, J. C. *Angew. Chem. Int. Ed.* **2010**, *49*, 8926–8929.
- (7) Kucharski, T. J.; Tian, Y.; Akbulatov, S.; Boulatov, R. *Energy Environ. Sci.* **2011**, *4*, 4449–4472.
- (8) Lennartson, A.; Roffey, A.; Moth-Poulsen, K. *Tetrahedron Lett.* **2015**, *56*, 1457–1465.
- (9) Gurke, J.; Quick, M.; Ernsting, N. P.; Hecht, S. *ChemComm* **2017**, *53*, 2150–2153.
- (10) Dong, L.; Feng, Y.; Wang, L.; Feng, W. *Chem. Soc. Rev.* **2018**, *47*, 7339–7368.
- (11) Edel, K.; Yang, X.; Ishibashi, J. S. A.; Lamm, A. N.; Maichle-Mössmer, C.; Giustra, Z. X.; Liu, S.-Y.; Bettinger, H. F. *Angew. Chem. Int. Ed.* **2018**, *57*, 5296–5300.
- (12) Sun, C.-L.; Wang, C.; Boulatov, R. *ChemPhotoChem* **2019**, *3*, 268–283.
- (13) Nielsen, M. B.; Ree, N.; Mikkelsen, K. V.; Cacciarini, M. *Russ. Chem. Rev.* **2020**, *89*, 573–586.
- (14) Dauben, W. G.; Cargill, R. L. *Tetrahedron* **1961**, *15*, 197–201.

- (15) Cristol, S. J.; Snell, R. L. *J. Am. Chem. Soc.* **1954**, *76*, 5000–5000.
- (16) Cristol, S. J.; Snell, R. L. *J. Am. Chem. Soc.* **1958**, *80*, 1950–1952.
- (17) Dreos, A.; Wang, Z.; Udmark, J.; Ström, A.; Erhart, P.; Börjesson, K.; Nielsen, M. B.; Moth-Poulsen, K. *Adv. Energy Mater.* **2018**, *8*, 1703401.
- (18) Quant, M.; Hamrin, A.; Lennartson, A.; Erhart, P.; Moth-Poulsen, K. *J. Phys. Chem. C* **2019**, *123*, 7081–7087.
- (19) Wang, Z.; Roffey, A.; Losantos, R.; Lennartson, A.; Jevric, M.; Petersen, A. U.; Quant, M.; Dreos, A.; Wen, X.; Sampedro, D.; Börjesson, K.; Moth-Poulsen, K. *Energy Environ. Sci.* **2019**, *12*, 187–193.
- (20) Jevric, M.; Petersen, A. U.; Mansø, M.; Kumar Singh, S.; Wang, Z.; Dreos, A.; Sumby, C.; Nielsen, M. B.; Börjesson, K.; Erhart, P.; Moth-Poulsen, K. *Chem.: Eur. J* **2018**, *24*, 12767–12772.
- (21) Quant, M.; Lennartson, A.; Dreos, A.; Kuisma, M.; Erhart, P.; Börjesson, K.; Moth-Poulsen, K. *Chem. Eur. J.* **2016**, *22*, 13265–13274.
- (22) Mansø, M.; Kilde, M. D.; Singh, S. K.; Erhart, P.; Moth-Poulsen, K.; Nielsen, M. B. *Phys. Chem. Chem. Phys.* **2019**, *21*, 3092–3097.
- (23) Mansø, M.; Tebikachew, B. E.; Moth-Poulsen, K.; Nielsen, M. B. *Org. Biomol. Chem.* **2018**, *16*, 5585–5590.
- (24) Orrego-Hernández, J.; Dreos, A.; Moth-Poulsen, K. *Acc. Chem. Res.* **2020**, *53*, 1478–1487.
- (25) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.
- (26) RDKit: Open-source cheminformatics. <http://www.rdkit.org> (version 2019.09.3).
- (27) Grimme, S.; Bannwarth, C. *J. Chem. Phys.* **2016**, *145*, 054103.

- (28) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. International Conference on Learning Representations (ICLR). 2017.
- (29) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; <https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>.
- (30) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. NIPS-W. 2017.
- (31) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. ICLR Workshop on Representation Learning on Graphs and Manifolds. 2019.
- (32) Koerstz, M.; Christensen, A. S.; Mikkelsen, K. V.; Nielsen, M. B.; Jensen, J. H. *PeerJ Phy. Chem.* **2021**, *3*, e16.
- (33) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.
- (34) Jacovella, U.; Carrascosa, E.; Buntine, J. T.; Ree, N.; Mikkelsen, K. V.; Jevric, M.; Moth-Poulsen, K.; Bieske, E. J. *J. Phys. Chem. Lett.* **2020**, *11*, 6045–6050.
- (35) Ree, N.; Mikkelsen, K. V. *Chem. Phys. Lett.* **2021**, 138665.
- (36) Kingma, D. P.; Ba, J. L. Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). 2015.

Supporting Information

Estimation of chemical space size

There are 18 different small substituents (including H) as shown in Figure 2a. There are five possible substitution sites on the phenyl ring (Figure 2b-c), so there is on the order of 18^5 different substituted phenyl rings but since the ortho- and para-positions are symmetrically equivalent this number should be reduced by roughly a factor of two. Thus there are roughly $2(\frac{1}{2}18^5)$ phenyl substituents (with and without the ethyl linker) plus 18 non-phenyl substituents (including H). These 1.9M substituents can be placed on four different, but symmetrically equivalent, sites in the NBD scaffold, resulting in roughly $\frac{1}{2}(1.9M)^4$ different NBD systems.

Supplementary figures and tables

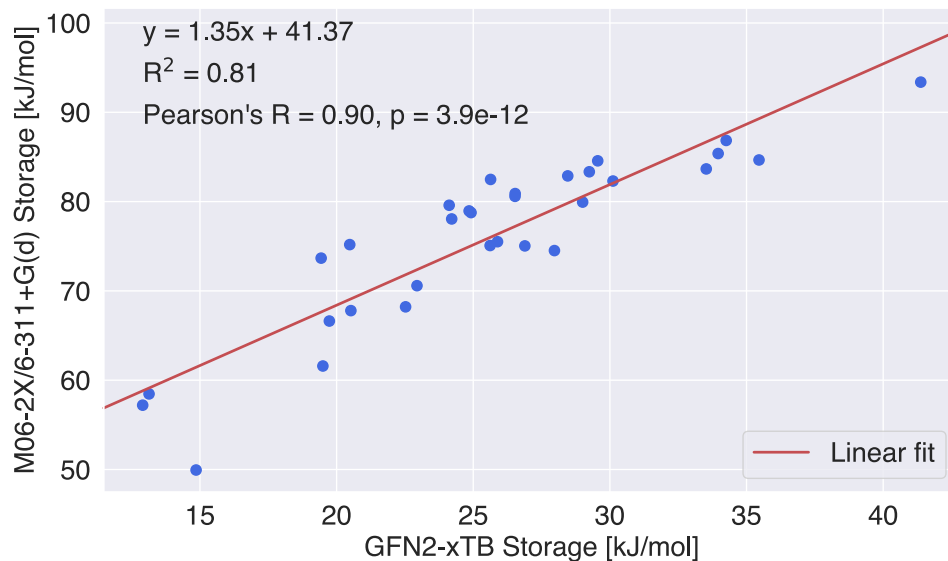


Figure S1: Initial storage energies check

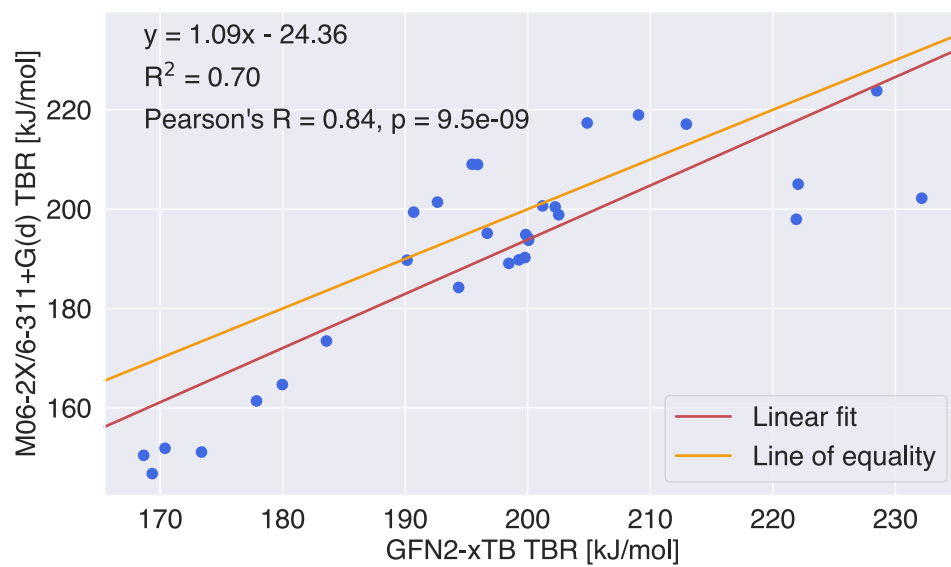


Figure S2: Initial TBR barriers check.

Table S1: Information about the systems used in Figure S1 and Figure S2 including references to experimental studies examining the systems.

Reactant	Product	GFN2-xTB Storage [kJ/mol]	M06-2X/6- 311+G(d) Storage [kJ/mol]	GFN2-xTB TBR [kJ/mol]	M06-2X/6- 311+G(d) TBR [kJ/mol]	Reference
N#CC1=C(c2ccc(C(O))=O)cc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc(C(O))=O)cc1	41.37	93.4	183.58	173.5	Jacovella et al. ³⁴
O=C(C1=C(c2ccc2)C2C=CC1C2)C(F)F	O=C(C(F)F)C12C3CC4C(C31)C42c1ccc1	20.52	67.8	170.41	151.9	Dreos et al. ¹⁷
O=C(C1=C(c2ccc2)C2C=CC1C2)C(F)F	O=C(C(F)F)C12C3CC4C(C31)C42c1ccc1	19.73	66.6	173.38	151.1	Dreos et al. ¹⁷
O=C(C1=C(c2ccc2)C2C=CC1C2)C(F)F	O=C(C(F)F)C12C3CC4C(C31)C42c1ccc1	19.50	61.6	168.36	146.8	Dreos et al. ¹⁷
CC1(C)C2C=CC1C(c1ccc1)=C2C(=O)C(F)F	CC1(C)C2C3C4C1C4(c1ccc1)C32C(=O)C(F)F	14.86	49.9	168.67	150.5	Dreos et al. ¹⁷
CC(C)=C1C2C=CC1C(c1ccc1)=C2C(=O)C(F)F	CC(C)=C1C2C3C4C1C4(c1ccc1)C23C(=O)C(F)F	12.90	57.2	177.87	161.4	Dreos et al. ¹⁷
O=C(C1=C(c2ccc2)C2C=CC1C2)C(F)F	O=C(C(F)F)C12C3C(=C(c4ccc4)C4C(C31)C42c1ccc1	13.14	58.5	179.98	164.7	Dreos et al. ¹⁷
Nc1ccc(C2=C(c3ccc(C(F)F)cc3)C3C=CC2C3)cc1	Nc1ccc(C23C4CC5C(C42)C53c2ccc(C(F)F)cc2)cc1	26.89	75.0	204.85	217.4	Dreos et al. ¹⁷
N#CC1=C(C#Cc2ccc(N)cc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42C#Cc1ccc(N)cc1	33.53	83.7	192.63	201.4	Dreos et al. ¹⁷
N#CC1=C(C#Cc2ccc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42C#Cc1ccc1	29.00	79.9	195.90	209.0	Quant et al. ¹⁸
CN(C)c1ccc(C#CC2=C(C#N)C3C=CC2C3)cc1	CN(C)c1ccc(C#CC23C4CC5C(C42)C53C#N)cc1	35.46	84.7	190.69	199.4	Quant et al. ¹⁸
N#CC1=C(c2ccc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc1	24.85	79.0	202.53	198.9	Quant et al. ¹⁸
C1=CC2CC1C(c1ccc1)=C2c1ccc1	C1ccc(C23C4CC5C(C42)C53c2ccc2)cc1	25.62	75.1	-	-	Quant et al. ¹⁸
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	28.46	82.9	199.77	190.3	Wang et al. ¹⁹
N#CC1=C(c2ccc(N+)(=O)[O-])cc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc(N+)(=O)[O-])cc1	19.43	73.7	195.46	209.0	Jevric et al. ²⁰
N#CC1=C(c2ccc(N+)(=O)[O-])cc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc(N+)(=O)[O-])cc1	20.48	75.2	222.06	205.0	Jevric et al. ²⁰
N#CC1=C(c2ccc(F)cc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc(F)cc1	24.12	79.6	221.91	198.0	Jevric et al. ²⁰
N#CC1=C(c2ccc2F)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc2F	27.97	74.5	232.15	202.2	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1F	COc1ccc(C23C4CC5C(C42)C53C#N)cc1F	26.53	80.9	200.07	193.7	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	22.94	70.6	228.48	223.8	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	25.63	82.5	201.21	200.7	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	29.55	84.6	198.46	189.1	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	25.89	75.5	209.04	219.0	Jevric et al. ²⁰
COc1ccc(C2=C(C#N)C3C=CC2C3)cc1	COc1ccc(C23C4CC5C(C42)C53C#N)cc1	30.11	82.3	196.69	195.2	Jevric et al. ²⁰
N#CC1=C(c2ccc3ccc23)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc3ccc23	26.53	80.6	200.07	194.1	Jevric et al. ²⁰
N#CC1=C(c2ccc3ccc23)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc3ccc23	22.52	68.2	212.94	217.1	Jevric et al. ²⁰
N#CC1=C(c2ccc3ccc23)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc3ccc23	24.92	78.8	199.84	194.9	Jevric et al. ²⁰
CN(C)c1ccc(C2=C(C#N)C3C=CC2C3)cc1	CN(C)c1ccc(C23C4CC5C(C42)C53C#N)cc1	34.26	86.8	194.36	184.3	Jevric et al. ²⁰
CC(C)(C)Sc1ccc(C2=C(C#N)C3C=CC2C3)cc1	CC(C)(C)Sc1ccc(C23C4CC5C(C42)C53C#N)cc1	24.21	78.1	202.24	200.4	Jevric et al. ²⁰
N#CC1=C(c2ccc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc2	33.96	85.4	190.15	189.7	Jevric et al. ²⁰
N#CC1=C(c2ccc2)C2C=CC1C2	N#CC12C3CC4C(C31)C42c1ccc2	29.25	83.3	199.28	189.8	Jevric et al. ²⁰

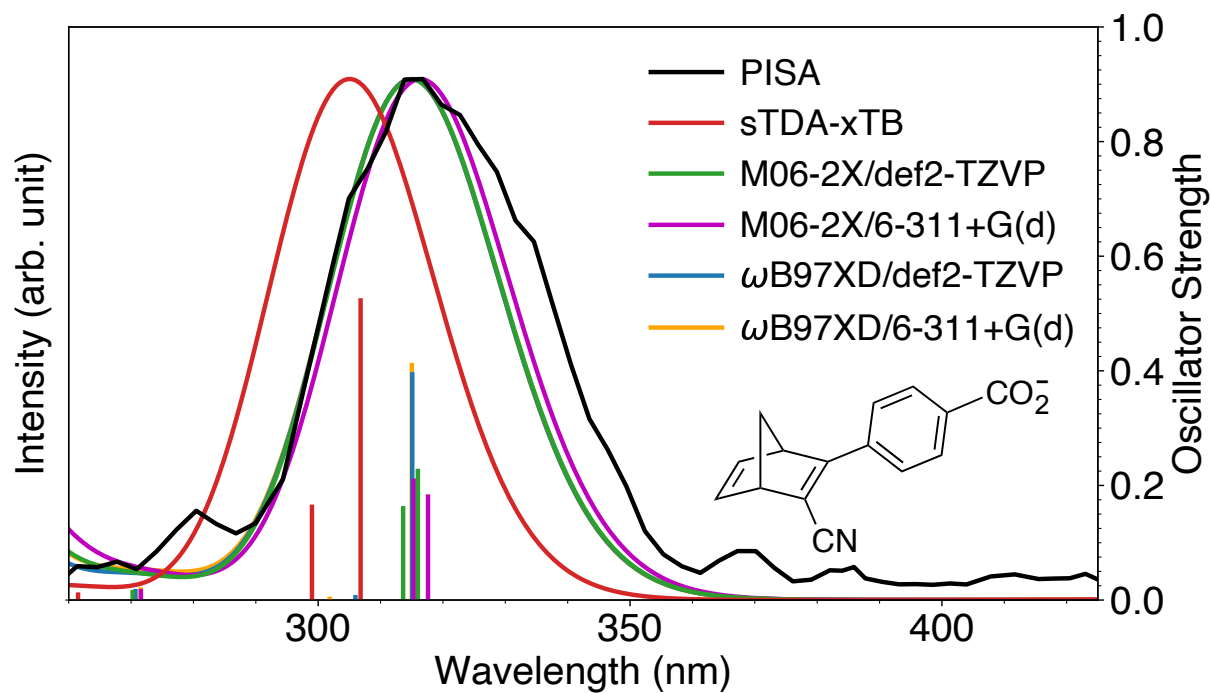


Figure S3: Initial UV-Vis check.

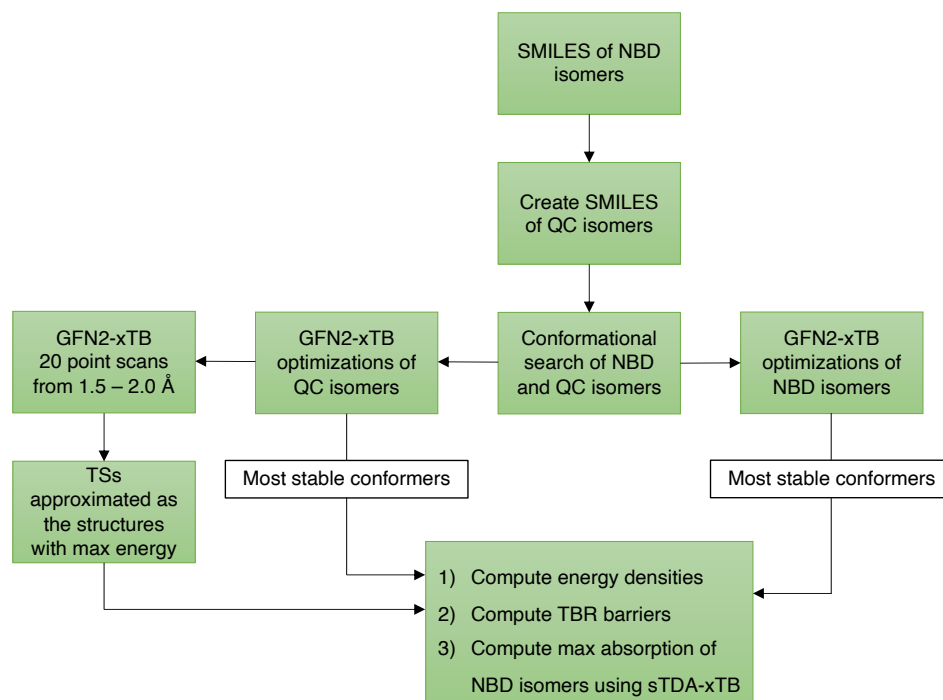


Figure S4: Flowchart.

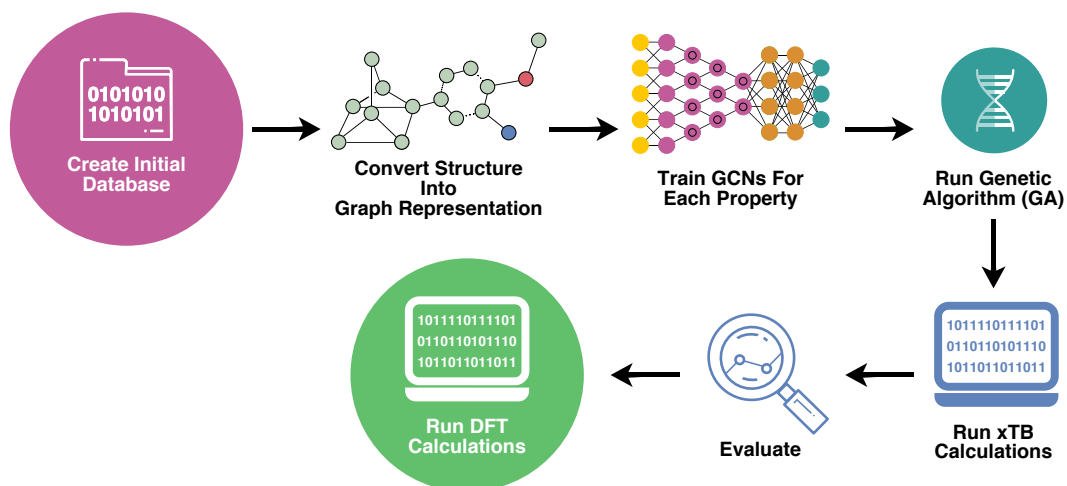


Figure S5: The complete procedure for high-throughput virtual screening of MOST systems based on the NBD/QC photoswitch.

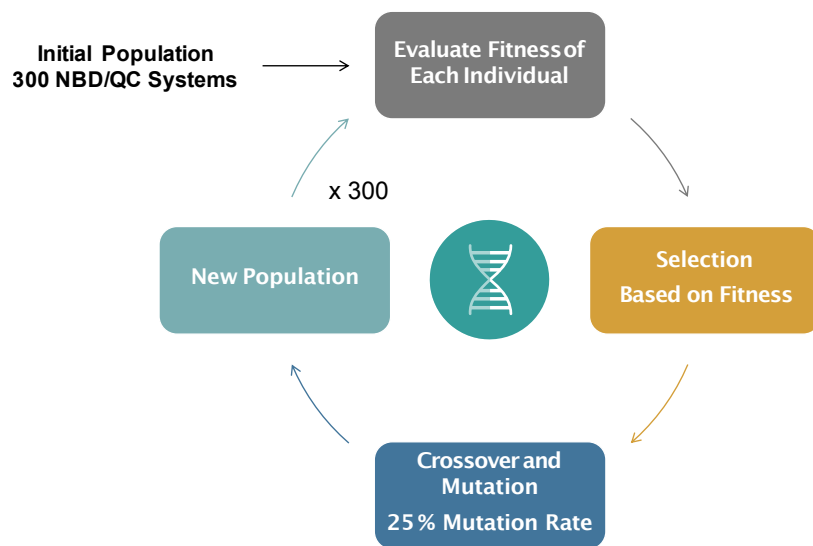


Figure S6: Illustration of the procedure for a single genetic algorithm (GA) search.

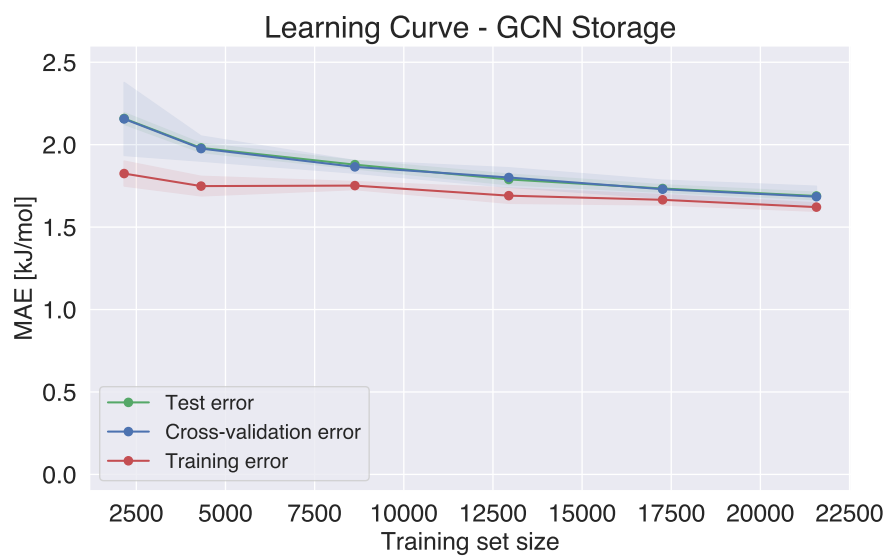


Figure S7: Learning curve - Storage energies.

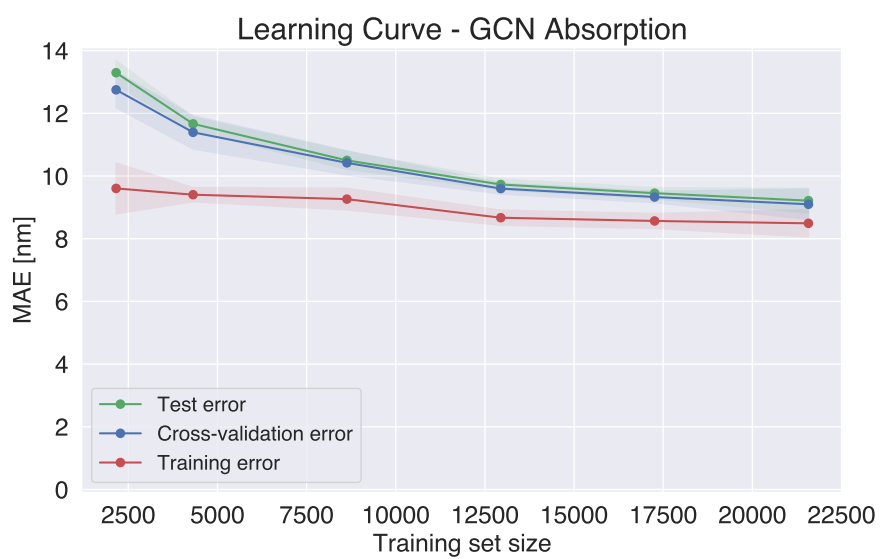


Figure S8: Learning curve - Absorption.

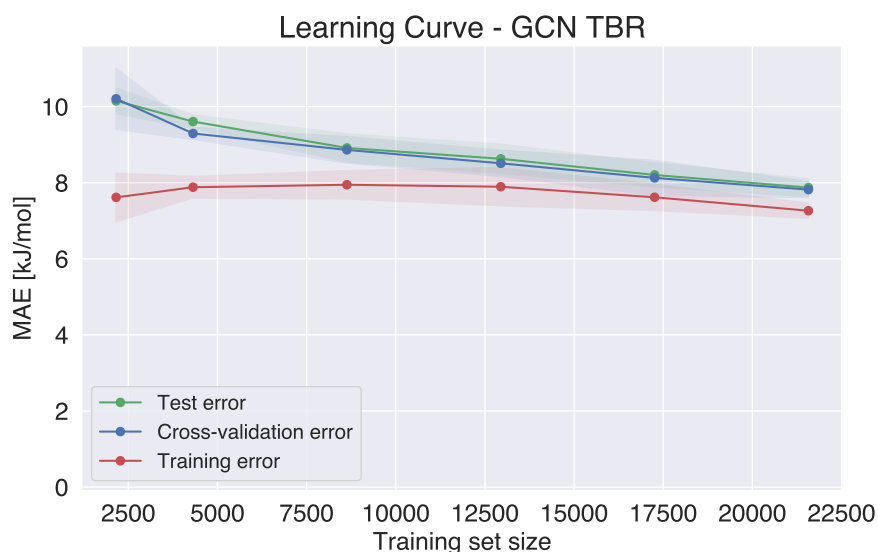


Figure S9: Learning curve - TBR barriers.

Procedure for obtaining the learning curves

- Perform a random shuffle of the training set containing 26,965 NBD/QC systems.
- Loop over a list to select the first [10%, 20%, 40%, 60%, 80%, 100%] of the training data.
 - Use 5-fold cross-validation to train the GCN on the selected subset for 100 epochs with a batch size of 512 and using the Adam optimizer with a learning rate of 0.01 on a MSE loss.
 - Test the performance on the validation set for every epoch and save the best model.
 - Use the best model to obtain the mean absolute error (MAE) on the training, validation and test sets.
 - Continue to the next fold and report the mean and standard deviation of the 5 models obtained in the cross-validation loop.
- Continue to the next subset.

The code for training the GCNs and creating the learning curves are available at <https://sid.erd.dk/sharelink/DeRV97z1Nz>

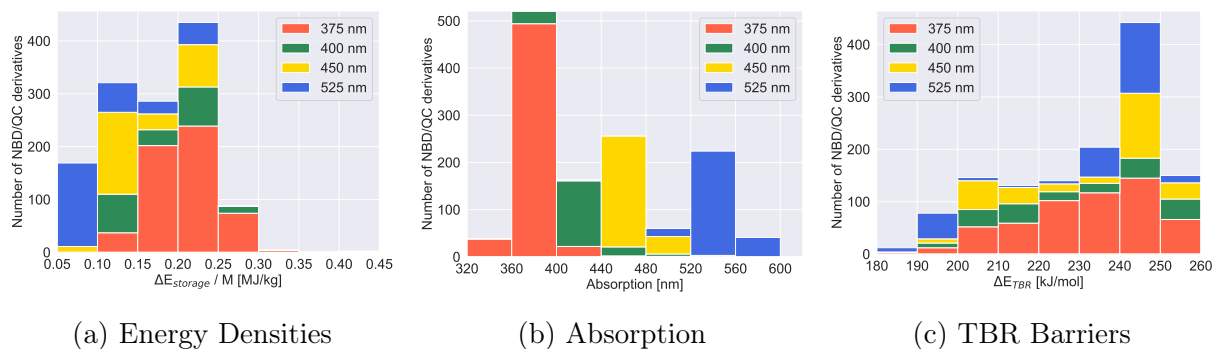
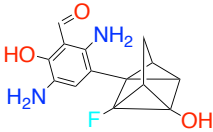
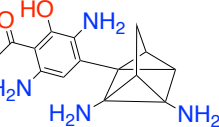
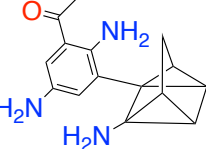
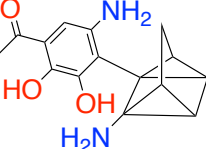
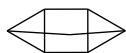


Figure S10: GA results. 1,306 structures, where 72 are shared between the four GA runs.

Table S2: TOP-1 candidates in each of the four GA runs with corresponding energy densities, $\Delta E_{\text{storage}}/M$, first absorption peak, λ , and activation barriers, ΔE_{TBR} . The percentage of how often the NBD/QC system was found is shown in the right-hand side column. The absorption threshold was varied in the four GA runs with (a); 375 nm, (b); 400 nm, (c); 450 nm, and (d); 525 nm. Moreover, the calculated properties of the NBD/QC parent system using GFN2-xTB are given as a reference in the bottom of the table.

Run	Structure	$\Delta E_{\text{storage}}/M$ [MJ/kg]	λ [nm]	ΔE_{TBR} [kJ/mol]	% of 1000 GA searches
(a)		0.24	369	240	72.5
(b)		0.25	417	210	60.8
(c)		0.22	451	221	27.7
(d)		0.21	525	207	2.6
		0.09	160	320	

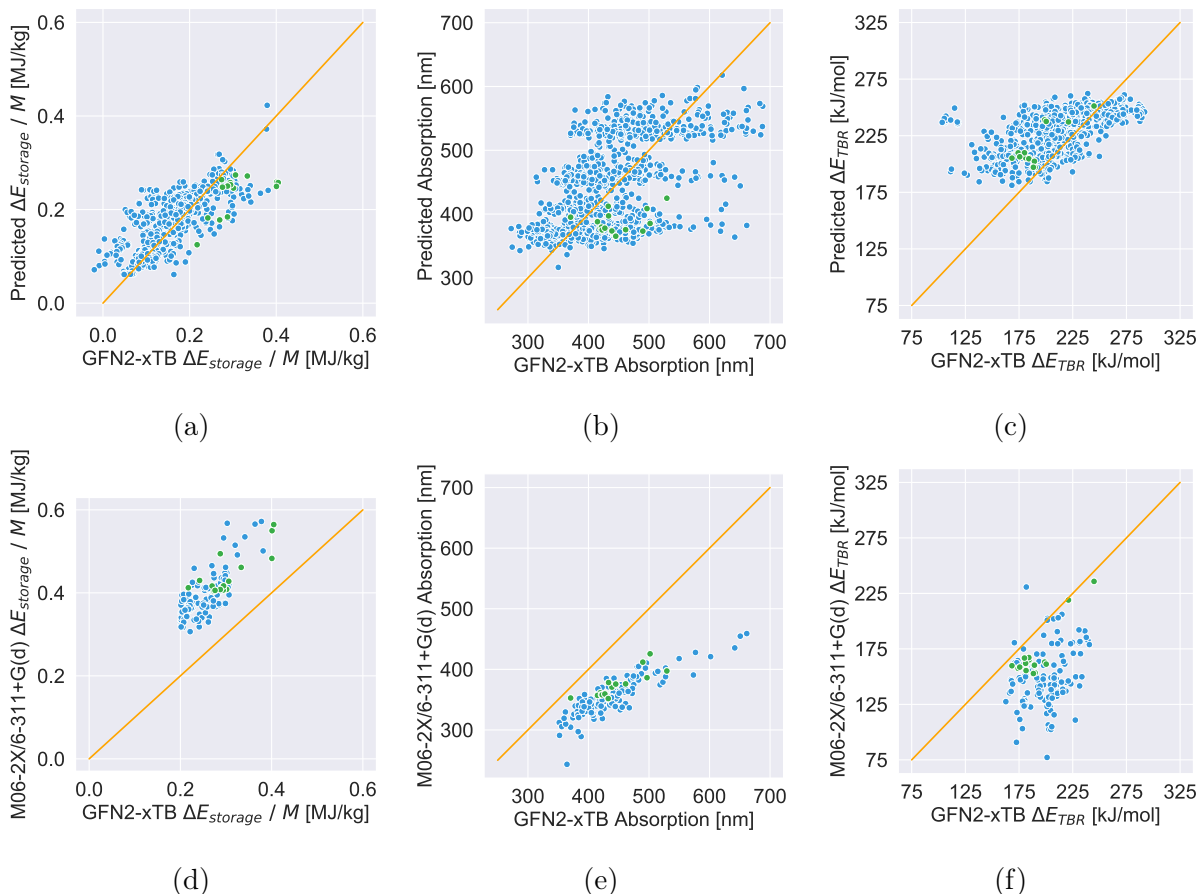


Figure S11: Comparison of the results for the investigated properties obtained by graph convolutional network (GCN), GFN2-xTB, and M06-2X/6-311+G(d). The top panels (a, b, and c) show the GCN vs xTB results for results of the 1,234 unique NBD/QC systems obtained from saving the 50 top-scoring molecules from the final populations of 4,000 GA searches. The bottom panels (d, e, and f) show the results of the NBD/QC systems with SQM-predicted energy densities, absorption spectra, and TBR barriers above 0.2 MJ/kg, 350 nm, and 160 kJ/mol, respectively. However, systems without a confirmed transition state structure at the DFT level of theory are omitted in the bottom panels, which limits the number of presented NBD/QC systems from 230 to 147. The green dots corresponds to the 15 NBD/QC systems shown in Table 1 and the orange line represents the line of equality.

Simulating UV-Vis Spectra

In this section, we present the expression used to simulate the UV-Vis spectra. The expression is derived from the assumption of Gaussian band shapes by applying Gaussian functions to convolute the calculated oscillator strengths. Thus, the UV-Vis spectra can be plotted as the extinction coefficient, ϵ , vs. the wavelength, λ , using the following equation.

$$\epsilon(\lambda) = \sum_{i=1}^n \epsilon_i(\lambda) = \sum_{i=1}^n k \cdot \frac{f_i}{\sigma} \cdot \exp \left[-4 \cdot \ln(2) \cdot \left(\frac{\frac{1}{\lambda} - \frac{1}{\lambda_i}}{\sigma \cdot 10^{-7}} \right)^2 \right] \quad (\text{S1})$$

where f_i is the calculated oscillator strength, λ_i is the corresponding wavelength in nm, λ is an independent variable defining the simulated spectrum, and σ is the standard deviation also known as the full width at half maximum of the Gaussian band (in these simulations $\sigma = 0.4 \text{ eV} = 0.4 \cdot 8065.54 \text{ cm}^{-1} = 3226.22 \text{ cm}^{-1}$). Furthermore, the constant k is given by

$$k = \frac{N_A \cdot e^2}{2 \cdot m_e \cdot c^2 \cdot \epsilon_0 \cdot \ln(10)} \cdot \sqrt{\frac{\ln(2)}{\pi}} = 2.1751 \cdot 10^8 \frac{L}{\text{mol} \cdot \text{cm}^2} \quad (\text{S2})$$

where N_A is Avogadro's constant, c is the speed of light, e is the elementary charge, m_e is the mass of an electron, and ϵ_0 is the vacuum permittivity.

Graphical TOC Entry

