High throughput virtual screening of 200 billion molecular solar heat battery candidates

Mads Koerstz[1], Anders S. Christensen[2], Kurt V. Mikkelsen[1], Mogens Brøndsted Nielsen[1], and Jan H. Jensen[1]

[1]Department of Chemistry, University of Copenhagen, Denmark
[2]Department of Chemistry, University of Basel, Switzerland

April 23, 2019

Abstract

The dihydroazulene/vinylheptafulvene (DHA/VHF) thermocouple is a promising candidate for thermal heat batteries that absorb and store solar energy as chemical energy without the need for insulation. However, in order to be viable the energy storage capacity and stability of the high energy form (the free energy barrier to the back reaction) must be increased significantly. We use semiempirical quantum chemical methods, machine learning, genetic algorithms, and density functional theory to virtually screen roughly 200 billion substituted DHA molecules to identify promising candidates for further study. We identify three molecules with predicted energy densities of (0.34-0.36 kJ/g), which is significantly larger than the 0.14 kJ/g computed for the parent system. The free energy barriers to the back reaction are between 6.8 and 7.7 kJ/mol higher than the parent compound, which should correspond to half-lives of days - sufficiently long for many practical applications.

Introduction

The Sun is the most abundant source of energy, but periods of supply do not always match periods of demand. Therefore finding solutions for storing solar energy is one of the major challenges for a sustainable society. One approach is to employ light-induced isomerization of photoactive molecules[9] as shown schematically in Figure 1. Upon irradiation, a molecule is converted to a high-energy photo-isomer and upon a certain stimulus, the high-energy isomer returns to the original molecule, and the excess energy is released as heat. This corresponds to a closed energy cycle of light-harvesting, energy storage and release, with no emission of CO_2 or other chemical products. Such systems are termed molecular solar-heat batteries.

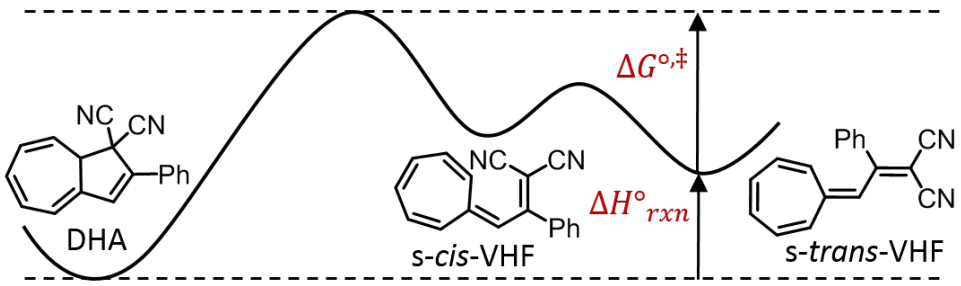


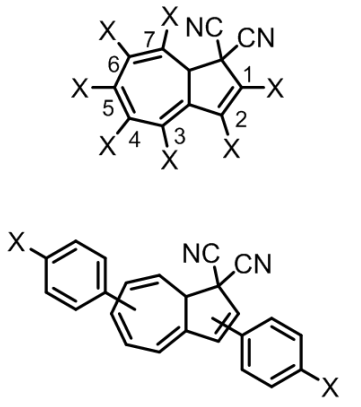
Figure 1: Schematic representation of the dihydroazulene/vinylheptafulvene (DHA/VHF) thermal heat battery. DHA is converted to VHF photochemically and the chemical energy (ΔH°_{rxn}) is released as heat when needed. The half life of VHA is determined by the free energy barrier of the back reaction to DHA. The molecule shown has been studied experimentally and if referred to as the "parent" molecule.

A suitable molecule 1) must absorb sunlight by converting to a higher energy form, 2) must absorb as much energy as possible (preferably ca 1 kJ/g), and 3) must be stable in the high-energy form for days or weeks. The key challenge is to test this complex set of demands for thousands of molecules in a computationally efficient and automated fashion. Dihydroazulenes (DHAs) are one class of promising candidates for solar heat batteries. The parent system (shown in Figure 1) absorbs at the right wavelength with good quantum yield. However, the energy density is only ≈0.1 kJ/g and the half life of VHF is only a few hours.[1] Removal of one cyano group increases the storage density to 0.25 kJ/g and the half-life to years.[3] Unfortunately, the back reaction could not be triggered without causing degradation. With both cyano groups present the storage density can be increased by up to 0.38 kJ/g but not without decreasing the half-life significantly.[12] The goal of this study is to identify substituted DHAs with both higher energy density and longer half life through high throughput virtually screening.

Computational Methodology

Semiempirical calculations

For high throughput virtual screening (HTVS) ΔH°_{rxn} is approximated as the difference in electronic energy (ΔE_{rxn}) computed using GFN2-xTB[5] while $\Delta G^{\circ,\ddagger}$ is approximated as the difference in heat of formation ($\Delta\Delta H^{\ddagger}_{f}$) computed using PM3[13] (collectively referred to as "SQM" hereafter). GFN2-xTB is chosen due to its computational efficiency, while PM3 is chosen because it is available in both ORCA 4.0[10] and Gaussian09[4] (see below). We screen 41 different substituents and 7 possible substituent positions (Figure 2) Structures for all (35,588) singly and doubly substituted DHA motifs are generated using RDKit [8]. For each DHA structure the VHF structure is automatically generated using RDKit. $5 + 5n_{rot}$ random conformations are generated using RDKit and optimized using GFN2-xTB. Optimizations that resulted in discrepancies between the input and output connectivity is discarded. The lowest energy conformers of DHA and VHF are used to compute ΔE_{rxn}.



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: The substituents and positions considered in this study. The substituents are separated into electron withdrawing groups (EWG) and electron donating groups (EDG). There are a total of 42 different substituents counting hydrogen and phenyl resulting in more than 200 billion molecules (slightly less than 42^7 due to permutational symmetry).

To compute the energy barrier a adiabatic scan is performed (using ORCA) for the breaking CC bond, which is constrained to 12 values out to 3.5Å starting from the DHA structure with the lowest energy. The highest energy structure is used as a starting point for a transition state (TS) search using PM3 in Gaussian09 while computing the Hessian in each step. The TS-search converged successfully for 32,623 structures. The remaining molecules are considered implicitly in the genetic algorithm search as described below. For the molecules where the TS search converged it is verified that the normal mode associated with the imaginary frequency lies along the reacting bond. The low energy VHF GFN2-xTB conformer is reoptimized using PM3, and the PM3 barrier is computed. This TS connects DHA with with s-*cis*-VHF while the lowest energy VHF conformer is usually s-*trans*-VHF, so the implicit assumption s-*cis*-VHF and s-*trans*-VHF are in thermal equilibrium, i.e. that the barrier between the two VHF conformations is lower than the barrier between s-*cis*-VHF and DHA (Figure 1). The calculations require about 5 calendar days using ca 300 CPU-cores and involve ≈1.6 million separate geometry optimizations.

DFT refinement

Select structures are investigated further at the M06-2X[14]/6-31G(d)[7] level of theory (using Gaussian09) in one of two ways. In the first approach, the SQM structure with lowest energy is reoptimized for DHA, VHF, and the TS and the electronic energy is used to compute the storage density and barrier. In the second approach a systematic conformer search is performed using SQM by rotating each rotateable bond by ±120° starting from the lowest energy structure found using the RDKit conformer generating algorithm. Each structure is energy-minimized and in the case of the TSs the reacting bond is constrained. The minimized TS structure is then used as an initial guess for an

unconstrained TS search. All unique conformers (conformers with Torsion Fingerprint Deviation[11] that are less than 0.001 are considered identical) are then reoptimized with M06-2X/6-31G(d) and the structures with the lowest free energy are used to compute the storage density and barrier.

Machine learning models and genetic algorithms

A molecule is represented by positional one-hot encoding, i.e. vector with 287 (41×7) binary elements. This representation was used to train three different machine learning (ML)-models. Linear regression and kernel ridge regression as implemented in Scikit-learn and a neural network as implemented with Pytorch. We found very little difference in performance of the three methods and focus on linear regression in the remaining part of the paper.

The linear regression model is used for genetic algorithm (GA) searches of chemical space. Each gene is represented by the same 287-bit binary vector used for the ML-model. Each of the seven one-hot encoded ligand-representation represents one of the seven bases in each gene. Crossover is performed by choosing a random cut point between bases for two parent genes and recombining. Mutations are performed by choosing a base in a gene at random and the changing the ligand. Parents are chosen with a probability that is proportional to their score. The score is computed using the sum of two Treshholded-Linear functions described by Brown *et al.*[2] with respective thresholds of 1.0 kJ/g and 178 kJ/mol for the storage density and barrier height. Each GA search is run for 100 generations using a mating pool and population size of 100 and a mutation rate of 1%. A new population is constructed from the highest scoring molecules of the current and previous population. The result of the GA search is the highest scoring molecule in the final population after 100 generations.

Results and Discussion

The size of the chemical space depicted in Figure 2 is roughly 200 billion molecules. We start by screening all 35,588 singly and doubly substituted DHAs using semiempirical methods (SQM). Promising candidates are then further tested using DFT. The SQM data is then used to train and test a ML-method to predict storage energies and barriers and this ML-method is used together with a genetic algorithm to search the entire chemical space. Promising candidates are further refined with SQM and then DFT.

Benchmarking the virtual screening approach

To benchmark the accuracy of the SQM reaction energies and barrier heights we choose 20 random molecules among the 100 molecules with largest storage energies, 20 random molecules out of the 100 with smallest storage energies, and 60 random molecules from the remaining molecules. For all 100 molecules we optimized the low energy GFN2-xTB

DHA, VHF, and TS structures with M06-2X/6-31G(d) and computed the storage energies and back reaction barriers. Figure 3(a) and (b) shows the reaction energy vs. the barrier computed with DFT and and SQM, respectively.

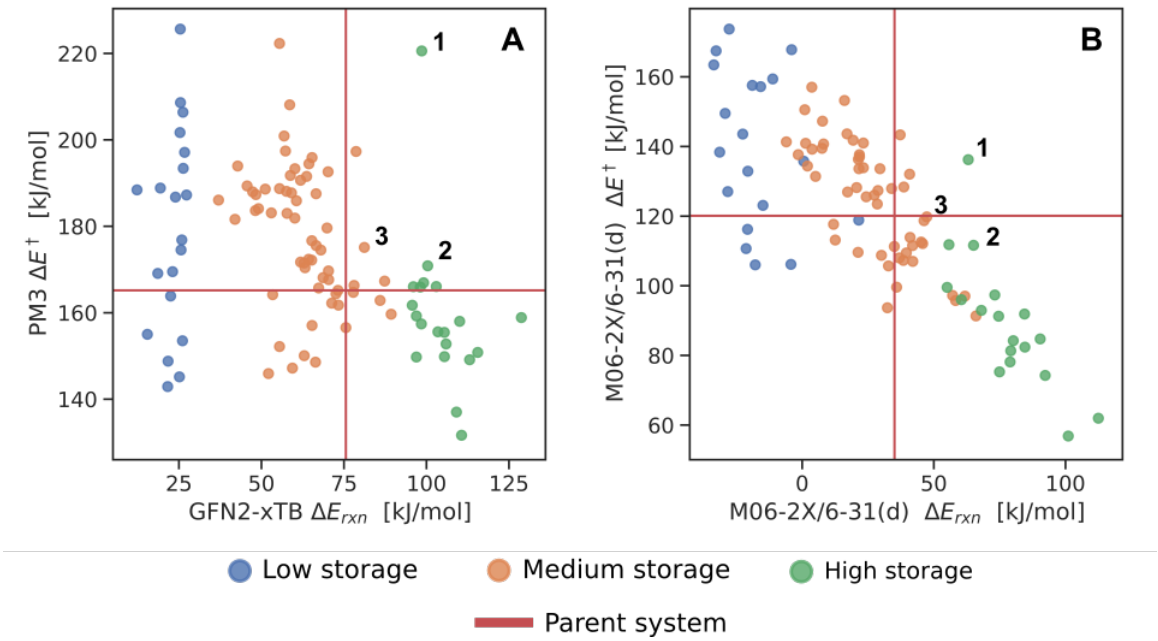


Figure 3: Electronic barrier heights and storage energies computed using (a) SQM and (b) DFT for 100 randomly selected doubly-substituted molecules. The vertical and horizontal lines mark the storage energy and barrier height for the parent compound shown in Figure 1. See text for further details.

The main question is whether SQM can reliably identify molecules for which the reaction energies and barriers height are considerably larger than the corresponding values for the parent compound (indicated by the red lines). Figure 3(b) shows that SQM identifies one such molecule (labelled "1") for which this is clearly the case and this molecule is also the most promising candidate when using DFT. SQM also identifies two other molecules with somewhat larger storage energy and slightly larger barrier ("2") or somewhat larger barrier and slightly larger storage energy ("3"). However, DFT predicts that the barriers for these molecules are either nearly identical to, or lower than, that of the reference compound. In general, PM3 tends to significantly overestimate the barrier relative to the reference compound, while GFN2-xTB tends to somewhat underestimate the storage energy (i.e. give fewer points to the right of the vertical line) and we will use this finding when selecting promising molecules in the next subsection.

The GFN2-xTB storage energy calculation is based on the lowest energy DHA and VHF structures found by optimizations of $5+5n_{rot}$ geometries generated using RDKit. We test the accuracy of this approach by generating all conformers by systematically rotating each rotateable bond by $\pm 120°$ for the 100 molecules-subset. We find (Figure S2) that the "$5+5n_{rot}$-approach" works really well for most molecules (including the 20 with high storage energy) and, if anything, overestimates the storage energy (i.e. at worst

we will have some false positives).

Screening singly and doubly substituted molecules

Figure 4(a) shows a plot of the the SQM barriers plotted vs the storage densities for 32,623 singly and doubly substituted DHA/VHF couples. We want to select roughly 100 of the most promising molecules for further study with M06-2X/6-31G(d). As discussed in the previous subsection, PM3 tends to overestimate the barrier relative to the parent molecule (the horizontal line) so the barrier should be significantly higher than that. Similarly, GFN2-xTB tends to underestimate the reaction energy relative to the parent system (the vertical line) and, hence, the energy storage density, so molecules with only somewhat higher storage density are potentially promising candidates. After some experimentation we found that cutoffs of 178 kJ/mol and 0.33 kJ/g leads to 109 molecules for further study using DFT (green box in Figure 4(a)). For all 109 molecules we optimize the lowest energy GFN2-xTB DHA, VHF, and TS structures with M06-2X/6-31G(d) and the results are shown in Figure 4(b). As expected, the majority of the molecules have higher energy storage densities than the parent compound, but many have significantly lower barriers.

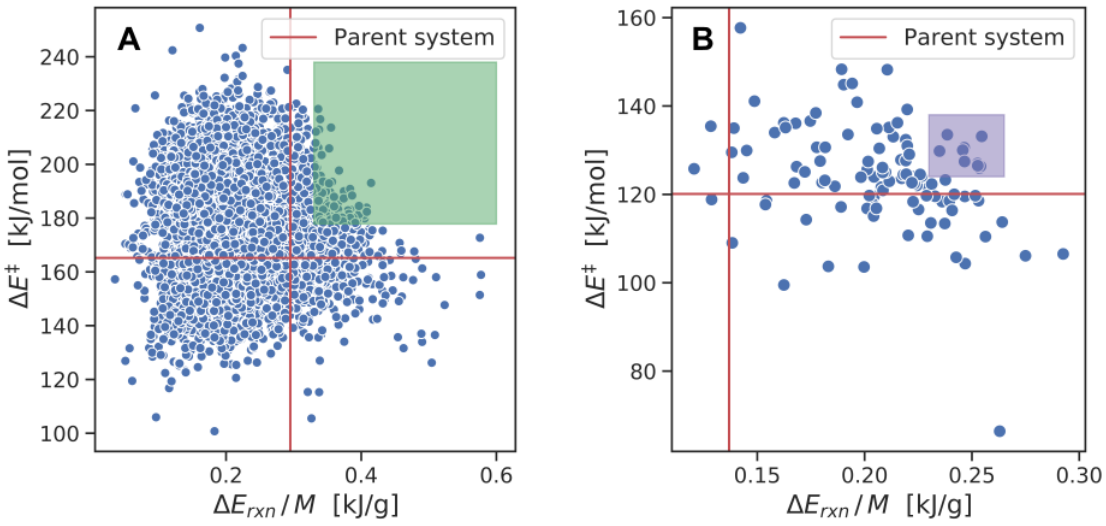


Figure 4: (a) Electronic barrier heights and storage energies computed for 32,623 singly- and doubly substituted molecules using SQM. (b) Electronic barrier heights and storage energies computed for the 109 molecules in the green box in (a) computed using M06-2X/6-31G(d). The geometry optimizations are started from the SQM structures and the purple box highlights 10 molecules selected for further study (Table 1). The vertical and horizontal lines mark the storage energy and barrier height for the parent molecule shown in Figure 1.

From this set we choose 10 molecules (the purple box in Figure 4(b), and Table 1) for more thorough examination by performing a systematic conformational search and

computing enthalpies and free energies for the lowest energy conformers using M06-2X/6-31G(d).

The resulting storage densities and free energy barrier heights are shown in Table 1. Five of the molecules are predicted to have an almost two-fold increase in storage density (0.24-0.25 kJ/g) compared to the parent system and all but one of these have predicted back reaction-barriers that are between 4.7 and 16.5 kJ/mol higher than the parent compound.

Figure 4(a) shows that there are three molecules with storage densities of nearly 0.6 kJ/g that we initially discounted because they are likely to have low barriers to the back reaction. To ensure that this is indeed the case we perform the same systematic conformational search and summarize the results in Table 2.

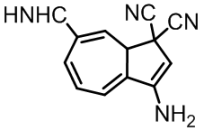
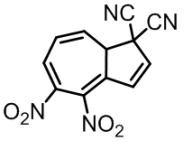
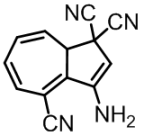
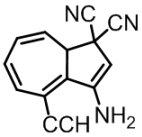
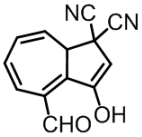
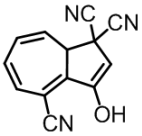
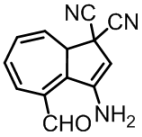
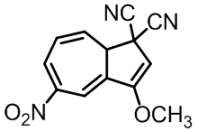
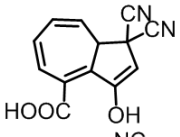
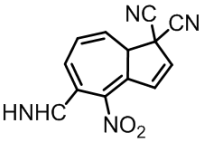
All three systems have an electron donating amino group at position 2 and an electron withdrawing group which allow for a hydrogen bond to the amino group in position 1. The three high energy density systems have very low back reaction barriers making them unsuitable for storage purposes. The low back reaction barrier is a consequence of the same properties that results in high storage energies. The amino group in position 2 have been shown to yield a large increase in storage energy, but also results in a decrease of the back reaction barrier [6]. The hydrogen bond between the amino group and the electron withdrawing group stabilizes the DHA system increasing the storage energy, also locks VHF in the s-*cis*-VHF conformer. This means that the most stable VHF conformer is structurally very similar to the transition state structure, making the back reaction barrier very small.

Screening all 200 billion molecules

In order to screen the entire chemical sub-space a faster estimation of storage energies and barriers is needed. We fit a linear regression model to SQM data obtained for 7,800 molecules with between one and six substituents (3000 molecules with 2 and 3 substituents, 1500 with four, 1000 with five, and 800 molecules with six substituents. This model is then tested on 1000 molecules in each category and we find that the accuracy decreases with substitution and ranges from 4.4 to 12.1 kJ/mol for storage energies and 5.5 to 17.6 kJ/mol for back reaction barriers (Figure S3). We also tested kernel ridge regression and a neural network, but these models are not significantly more accurate (data not shown).

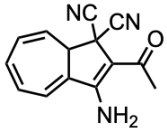
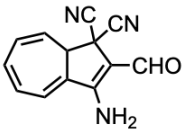
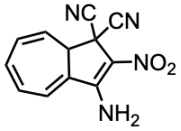
The ML-model is then used to perform 46,000 genetic algorithm (GA) searches for molecules with high storage density and barrier heights. The highest scoring molecule for each search is kept and this results in 15,522 unique molecules as shown in Figure 5(a). We recompute the storage density of these molecules using GFN2-xTB and select the 291 molecules with storage densities greater than 0.4 kJ/g (Figure 5(b)). We compute the storage density and barrier for these 291 molecules using a systematic conformer search and select 177 molecules for further refinement with M06-2X/6-31G(d) (Figure

Table 1: M06-2X/6-31G(d) predicted storage densities and back reaction barrier heights for the 10 molecules highlighted in Figures 4(b), based on the lowest free energy structures. The corresponding M06-2X/6-31G(d)-values for the parent molecule are 0.14 kJ/g and 119.1 kJ/mol, respectively.

Ranking	Structure	Name	$\Delta H^{\circ}_{rxn}/M$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
1		2-12:15-11:6	0.25	112.8
2		2-5:5-8:9	0.25	135.6
3		2-4:15-8:6	0.25	123.8
4		2-10:15-8:6	0.24	124.2
5		2-6:13-8:6	0.24	129.9
6		2-4:13-8:6	0.22	123.8
7		2-6:15-9:6	0.22	131.6
8		2-5:14-9:6	0.21	131.7
9		2-7:13-8:6	0.20	138.8
10		2-5:12-8:9	0.17	146.3

5(c)). As before, we first reoptimize the lowest energy structure of DHA, VHF, and the corresponding TS found with SQM using M06-2X/6-31G(d). Based in the DFT results

Table 2: M06-2X/6-31G(d) predicted storage densities and back reaction barrier heights for the three molecules with near 0.6 kJ/g storage density shown in Figures 4(a), based on the lowest free energy structures. The corresponding M06-2X/6-31G(d)-values for the parent molecule are 0.14 kJ/g and 119.1 kJ/mol, respectively

Ranking	Structure	Name	$\Delta H^{\circ}_{rxn}/M$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
X11		2-8:15-5:6	0.51	55.0
X12		2-6:15-5:6	0.51	58.1
X13		2-5:15-5:6	0.50	55.4

(Figure 5(d)) we chose 10 molecules and perform a systematic conformer search and use the lowest free energy structures to compute storage densities and back reaction barrier heights.

The resulting storage densities and free energy barrier heights are shown in Table 3. Seven of the molecules are predicted to have a 2.5-fold increase in storage density (0.34-0.36 kJ/g) compared to the parent system and three of these have predicted back reaction-barriers that are between 6.8 and 7.7 kJ/mol higher than the parent compound. Such an increase in barrier corresponds to a half-life of days rather than hours, which is sufficient for many practical applications.

Summary and outlook

We virtually screen 41 different substituents and 7 possible substituent positions (Figure 2) of the dihydroazulene/vinylheptafulvene (DHA/VHF) thermal heat battery (Figure 1) for molecules with high storage density (ΔH°_{rxn}/MW) and stability ($\Delta G^{\circ,\ddagger}$) The size of the chemical space is roughly 200 billion molecules. We start by screening all 35,588 singly and doubly substituted DHAs using semiempirical methods (SQM): GFN2-xTB for the storage density and PM3 for the barrier height of the back reaction. Compared to M06-2X/6-31G(d), PM3 tends to significantly overestimate the barrier relative to the reference compound, while GFN2-xTB tends somewhat underestimate the storage energy, but the methods are sufficiently accurate to identify promising molecules for further refinement (Figure 3).

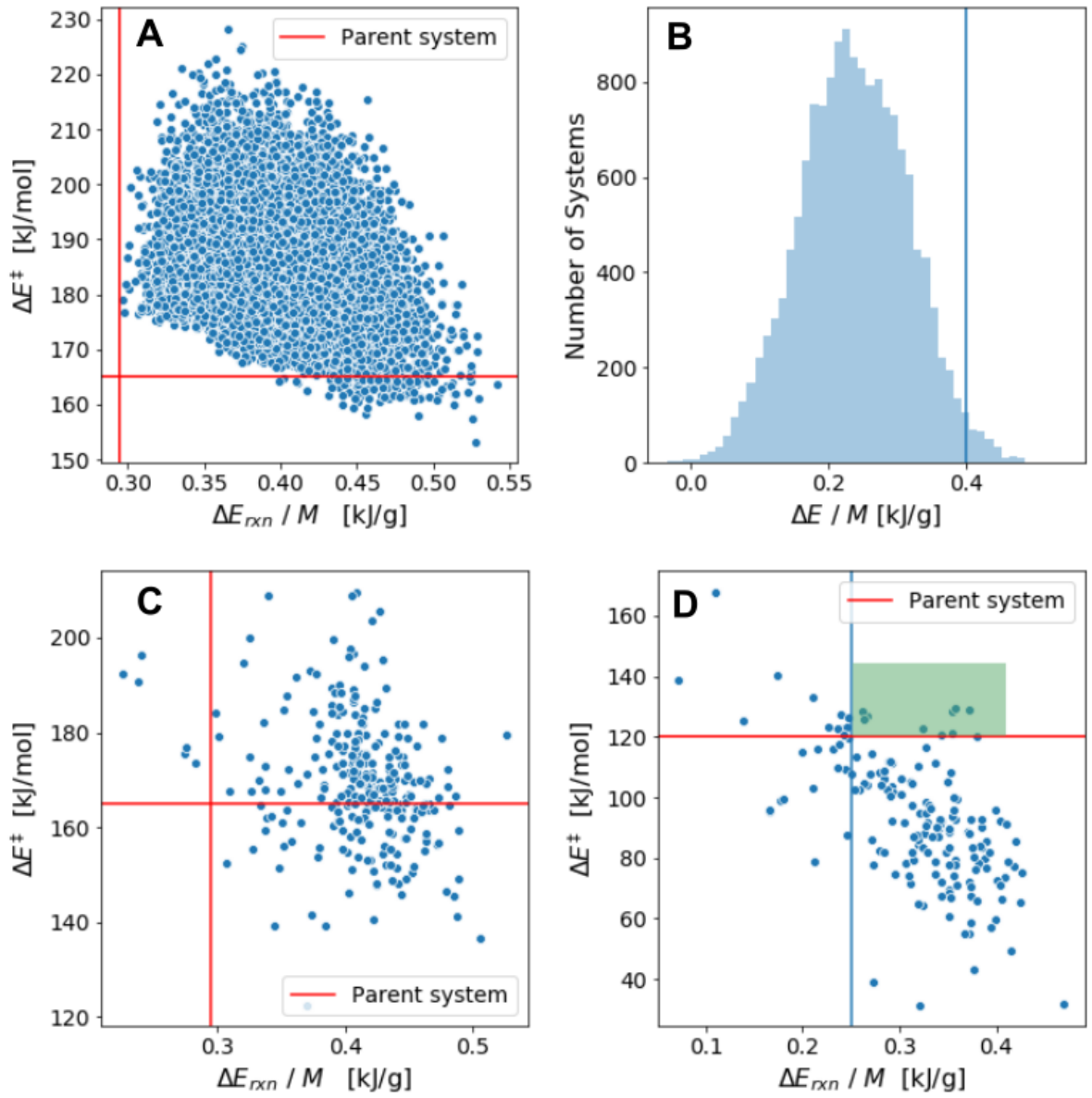
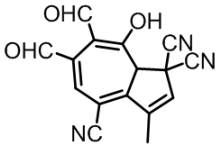
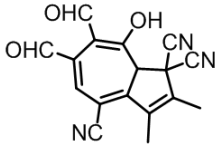
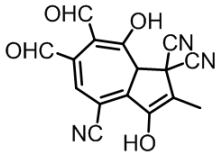
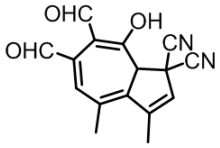
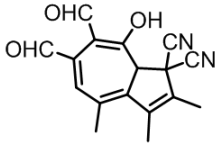
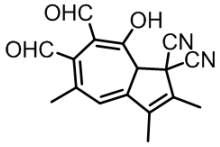
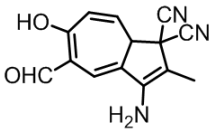
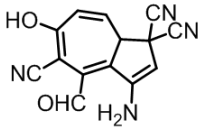
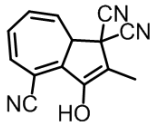
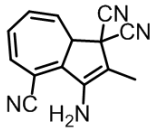


Figure 5: (a) Plot of ML-predicted barrier heights vs storage densities for 15,522 molecules with up to seven substituents found using 46,000 genetic algorithm searches. (b) Histogram of storage densities recomputed using GFN2-xTB. 300 molecules with storage densities higher than 0.4 kJ/g (vertical line) are selected for further study. The 177 molecules in the green box are selected for DFT calculations. (d) M06-2X/6-31G(d) barrier heights and storage densities for the 177 molecules. The 10 molecules in the purple box are selected for further study (Table 3). The vertical and horizontal lines in (a), (b), and (d) mark the storage energy and barrier height for the parent molecule shown in Figure 1.

The storage density and back reaction barrier of all 35,588 singly- and doubly-substituted DHA molecules are evaluated using SQM and used to identify 109 molecules for further study using M06-2X/6-31G(d) (Figure 4(a)). The energy densities and barrier heights

Table 3: M06-2X/6-31G(d) predicted storage densities and back reaction barrier heights for the 10 molecules highlighted in Figures 5(d), based on the lowest free energy structures. The corresponding M06-2X/6-31G(d)-values for the parent molecule are 0.14 kJ/g and 119.1 kJ/mol, respectively.

Ranking	Structure	Name	$\Delta H^{\circ}_{rxn}/M$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
1		GAsystem-48020	0.36	110.8
2		GAsystem-4113	0.36	125.9
3		GAsystem-119438	0.35	126.0
4		GAsystem-431344	0.35	116.4
5		GAsystem-133901	0.36	117.5
6		GAsystem-289328	0.34	112.4
7		GAsystem-362841	0.34	104.0
8		GAsystem-256082	0.30	109.5
9		GAsystem-1164	0.25	121.8
10		GAsystem-505	0.24	117.2

computed by reoptimizing the lowest energy SQM-conformations are then used to select 10 molecules for further study using a thorough conformational search (Figure 4(b) and Table 1). Five of the molecules are predicted to have an almost two-fold increase in storage density (0.24-0.25 kJ/g) compared to the parent system and all but one of these have predicted back reaction-barriers that are between 4.7 and 16.5 kJ/mol higher than the parent compound.

In order to screen the entire chemical sub-space we generate additional SQM-data for higher degrees of substitution and use it to fit linear regression models that reproduce the storage energies and barrier heights to within 4.4-12.1 and 5.5-17.6 kJ/mol depending on degree of substitution (Figure S3). These linear regression models are then used to search chemical space using a genetic algorithm (GA). 46,000 GA-searches result in 15,522 unique molecules (Figure 5(a)). From these, 177 molecules are selected based on SQM calculations (Figure 5(b)-(c)) for further study using M06-2X/6-31G(d) and 10 are selected for systematic conformational search (Figure 5(d)). Seven of these molecules are predicted to have a 2.5-fold increase in storage density (0.35-0.36 kJ/g) compared to the parent system and three of these have predicted back reaction-barriers that are between 6.8 and 7.7 kJ/mol higher than the parent compound.

Supplementary information

Additional figures referred to in the text can be found in SI.

Acknowledgments

This work was supported by a research grant (00022896) from VILLUM FONDEN. ASC acknowledges support by The National Centre of Competence in Research (NCCR) Materials Revolution: Computational Design and Discovery of Novel Materials (MARVEL) of the Swiss National Science Foundation (SNSF).

References

[1] Søren Lindbæk Broman, Sophie Lehn Brand, Christian Richard Parker, Michael Åxman Petersen, Christian Gregers Tortzen, Anders Kadziola, Kristine Kilså, and Mogens Brøndsted Nielsen. Optimized synthesis and detailed NMR spectroscopic characterization of the 1, 8a-dihydroazulene-1, 1-dicarbonitrile photoswitch. *ARKIVOC*, 2011.

[2] Nathan Brown, Marco Fiscato, Marwin H S Segler, and Alain C Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.*, 59(3):1096–1108, March 2019.

[3] Martina Cacciarini, Anders B Skov, Martyn Jevric, Anne S Hansen, Jonas Elm, Henrik G Kjaergaard, Kurt V Mikkelsen, and Mogens Brøndsted Nielsen. Towards

solar energy storage in the photochromic dihydroazulene-vinylheptafulvene system. *Chemistry*, 21(20):7454–7461, May 2015.

[4] Mjea Frisch, G W Trucks, H Bernhard Schlegel, Gustavo E Scuseria, Michael A Robb, James R Cheeseman, Giovanni Scalmani, Vincenzo Barone, Benedetta Mennucci, Gaea Petersson, and Others. Gaussian 09, revision d. 01, 2009.

[5] Stefan Grimme, Christoph Bannwarth, and Philip Shushkov. 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). *Journal of Chemical Theory and Computation*, 13(5):1989–2009, may 2017.

[6] Mia Harring Hansen, Jonas Elm, Stine T Olsen, Aske Nørskov Gejl, Freja E Storm, Benjamin N Frandsen, Anders B Skov, Mogens Brøndsted Nielsen, Henrik G Kjaergaard, and Kurt V Mikkelsen. Theoretical investigation of substituent effects on the Dihydroazulene/Vinylheptafulvene photoswitch: Increasing the energy storage capacity. *J. Phys. Chem. A*, 120(49):9782–9793, December 2016.

[7] W J Hehre, R Ditchfield, and J A Pople. Self—Consistent molecular orbital methods. XII. further extensions of Gaussian—Type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.*, 56(5):2257–2261, March 1972.

[8] Greg Landrum. Rdkit: Open-source cheminformatics. `http://www.rdkit.org`.

[9] Kasper Moth-Poulsen. Molecular systems for solar thermal energy storage and conversion. In Mogens Brøndsted Nielsen, editor, *Organic Synthesis and Molecular Engineering*, volume 103, pages 179–196. John Wiley & Sons, Inc., Hoboken, NJ, USA, November 2013.

[10] Frank Neese. Software update: the ORCA program system, version 4.0. *WIREs Comput Mol Sci*, 8(1):e1327, January 2018.

[11] Tanja Schulz-Gasch, Christin Schärfer, Wolfgang Guba, and Matthias Rarey. TFD: Torsion fingerprints as a new measure to compare small molecule conformations. *Journal of Chemical Information and Modeling*, 52(6):1499–1512, June 2012.

[12] Anders B Skov, Søren Lindbaek Broman, Anders S Gertsen, Jonas Elm, Martyn Jevric, Martina Cacciarini, Anders Kadziola, Kurt V Mikkelsen, and Mogens Brøndsted Nielsen. Aromaticity-Controlled energy storage capacity of the Dihydroazulene-Vinylheptafulvene photochromic system. *Chemistry*, 22(41):14567–14575, October 2016.

[13] James J. P. Stewart. Optimization of parameters for semiempirical methods i. method. *Journal of Computational Chemistry*, 10(2):209–220, mar 1989.

[14] Yan Zhao and Donald G Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theor. Chem. Acc.*, 120(1):215–241, May 2008.

Supporting Information

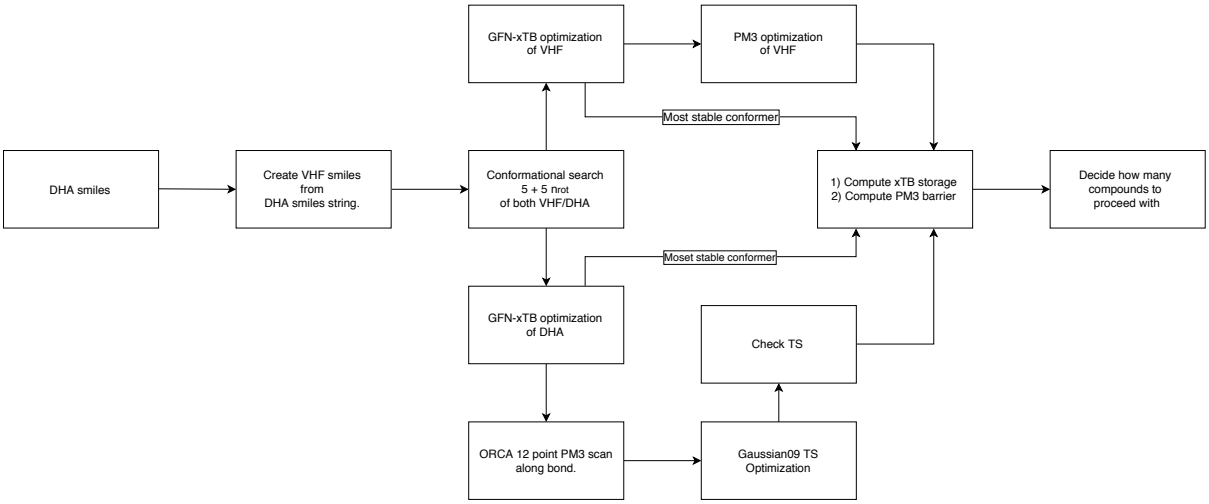


Figure S1: Flowchart illustrating the high-throughput screening procedure.

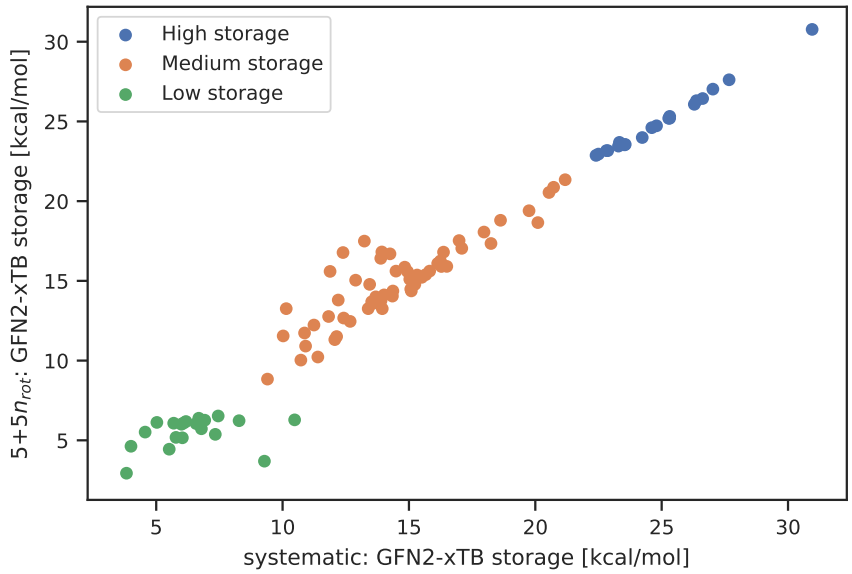
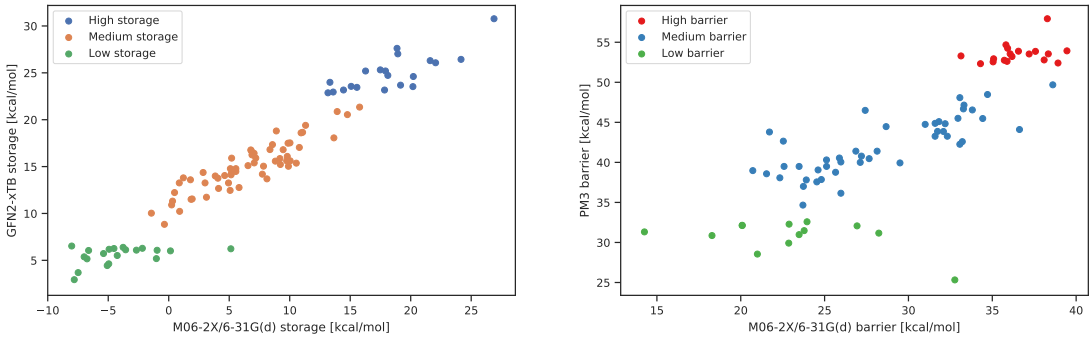


Figure S2: Comparison between systematic and $5 + 5n_{rot}$ conformational search. The MSE is 0.78 kcal/mol



(a) Comparison of storage energy

(b) Comparison of the barrier energy

Figure S3: Comparison of GFN2-xTB (a) and PM3 (b) ability to predict the storage energy and barrier compared to DFT.

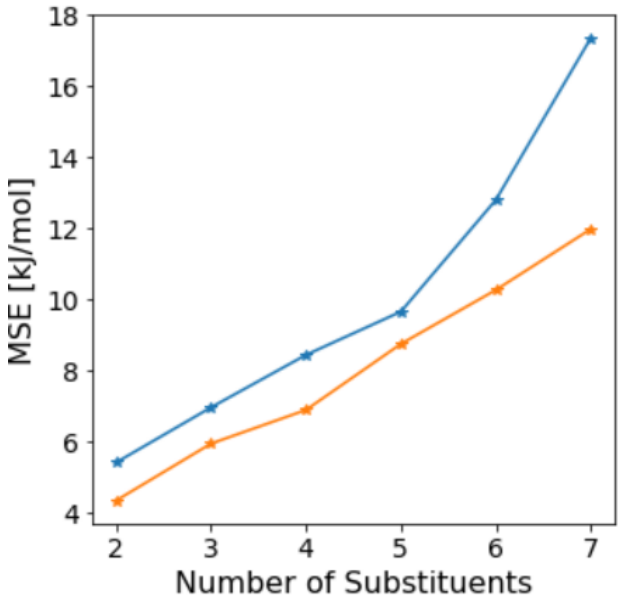


Figure S4: MSE error for the regression models of storage energy (orange line) and barrier height (blue line) with increasing number of substituents. The repsective MSE for the storage and barrier heights range from 4.4 to 12.1 kJ/mol, and from 5.5 to 17.6 kJ/mol.

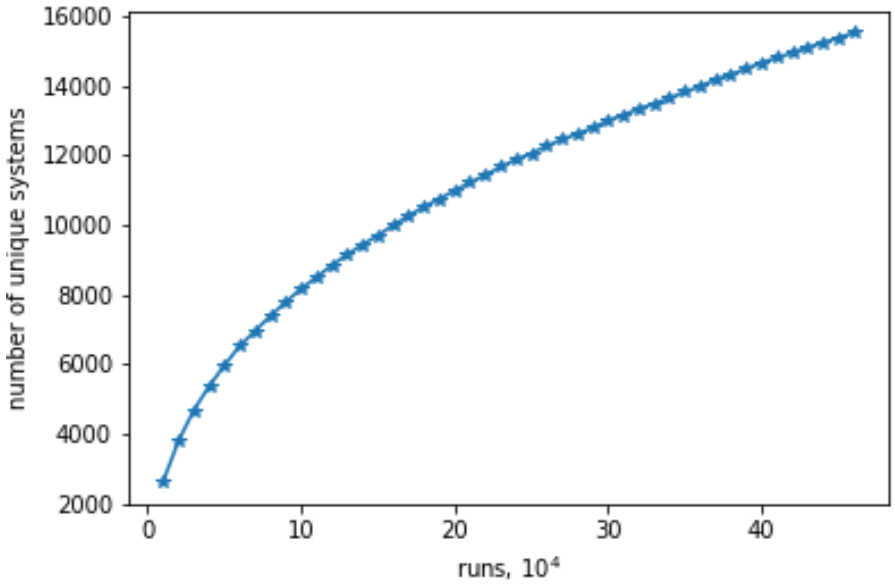


Figure S5: Total number of unique molecules after 10.000, 20.000, ...,46.000 runs .