High throughput virtual screening of 230 billion molecular solar heat battery candidates

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

[1] University of Copenhagen, Copenhagen, Danmark, Denmark

[2] University of Basel, Basel, Switzerland

Corresponding Author: Jan H Jensen
Email address: jhjensen@chem.ku.dk

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 lifetime of the high energy form (i.e. the free energy barrier to the back reaction) of the canonical parent compound must be increased significantly to be of practical use. We use semiempirical quantum chemical methods, machine learning, and density functional theory to virtually screen over 230 billion substituted DHA molecules to identify promising candidates. We identify a molecule with a predicted energy density of 0.38 kJ/g, which is significantly larger than the 0.14 kJ/g computed for the parent compound. The free energy barrier to the back reaction is 11 kJ/mol higher than the parent compound, which should correspond to a half-life of about 10 days - 4 months. This is considerably longer than the 3-39 hours (depending on solvent) observed for the parent compound and sufficiently long for many practical applications. Our paper makes two main important contributions: 1) a novel and generally applicable methodological approach that makes screening of huge libraries for properties involving chemical reactivity with modest computational resources, and 2) a clear demonstration that the storage capacity of the DHA/VHF thermocouple cannot be increased to >0.5 kJ/g by combining simple substituents.

High throughput virtual screening of 230 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

November 3, 2020

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 lifetime of the high energy form (i.e. the free energy barrier to the back reaction) of the canonical parent compound must be increased significantly to be of practical use. We use semiempirical quantum chemical methods, machine learning, and density functional theory to virtually screen over 230 billion substituted DHA molecules to identify promising candidates. We identify a molecule with a predicted energy density of 0.38 kJ/g, which is significantly larger than the 0.14 kJ/g computed for the parent compound. The free energy barrier to the back reaction is 11 kJ/mol higher than the parent compound, which should correspond to a half-life of about 10 days - 4 months. This is considerably longer than the 3-39 hours (depending on solvent) observed for the parent compound and sufficiently long for many practical applications. Our paper makes two main important contributions: 1) a novel and generally applicable methodological approach that makes screening of huge libraries for properties involving chemical reactivity with modest computational resources, and 2) a clear demonstration that the storage capacity of the DHA/VHF thermocouple cannot be increased to >0.5 kJ/g by combining simple substituents.

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[1, 2] as exemplified by the DHA/VHF thermocouple 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.

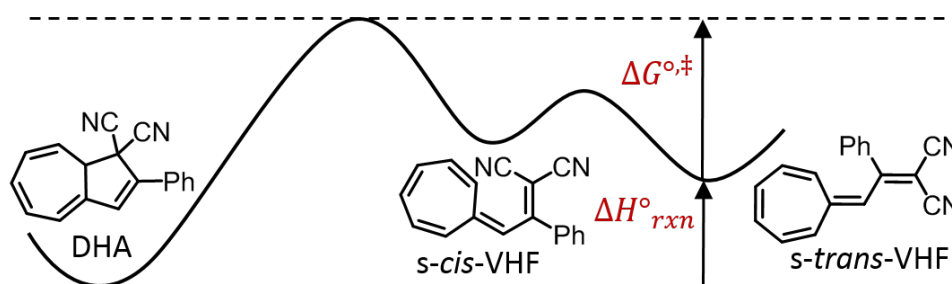



Figure 1: Schematic representation of the dihydroazulene/vinylheptafulvene (DHA/VHF) thermal heat battery. DHA is converted to VHF photochemically and the excess chemical energy (ΔH°_{rxn}) is released as heat when needed. The half life of VHF is determined by the free energy barrier of the back reaction to DHA. The molecule shown has been studied experimentally and is 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 (where ca 1 kJ/g is considered the practical maximum limit [3], but 0.3 kJ/g has been considered a reasonable target for some applications [4]), and 3) must be stable in the high-energy form for days or weeks. 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 0.14 kJ/g and the half-life of VHF is only 3-39 hours (depending on solvent).[5] Removal of one cyano group increases the storage density to 0.25 kJ/g and the half-life to years.[6] 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.[7] The goal of this study is to identify substituted DHAs with both higher energy density and longer half life through high throughput virtually screening.

We consider 42 common substituents (including H) at seven different positions, as shown in Figure 2, resulting in more than 230 billion molecules (slightly less than 42^7 due to permutational symmetry). To our knowledge this compound library is the largest library considered thus far. For example, it is three orders of magnitude larger than the "ultra-large" library of 170 million compounds used by Shoichet, Irwin, and coworkers.[8] While there has been a few virtual screening studies of thousands of reaction energies,[9] corresponding screening studies of barrier heights typically involve less than 100 molecules (e.g. [10]). Thus, screening barrier heights for billions or even thousands of molecules represents a fundamental challenge.

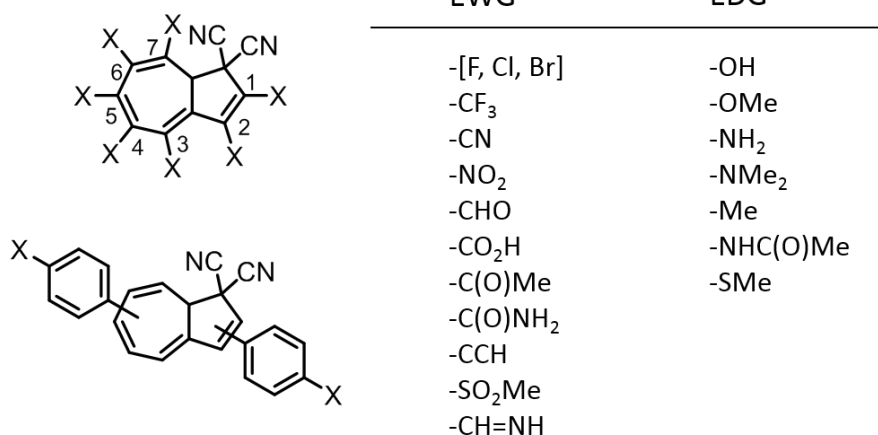



Figure 2: The substituents and positions considered in this study. Substitutions at position 8 is not considered, since removal of one cyano group has been shown to cause problems with the back reaction. The substituents are separated into electron withdrawing groups (EWG) and electron donating groups (EDG). There are a total of 42 different substituents counting hydrogen and phenyl resulting in more than 230 billion molecules (slightly less than 42^7 due to permutational symmetry).

This paper is organized as follows. First we demonstrate the utility of semiempirical quantum mechanics (SQM) by screening all 35,588 singly and doubly substituted DHAs. Then we use a simple linear regression model and a gradient boosting decision tree (GBDT) trained on SQM data, to screen all 230 billion molecules. Finally, the linear regression model is then used to demonstrate that the best molecules predicted by that model can be found efficiently using a genetic algorithm (GA), suggesting that a GA could be used to screen even larger chemical spaces, perhaps using SQM directly rather than machine learning.

Results and Discussion

The goal of this study is to identify molecules with an energy storage density that is as high as possible and a half-life of that is at least as long as the parent compound (shown in Figure 1) and preferably longer. We note that, according to transition state theory, even a modest 10 kJ/mol increase in the activation free energy corresponds to a 56 fold increase in the half life, which is 3-39 hours for the parent compound (depending on solvent). Even high level ab initio calculations can easily give an error of ±10 kJ/mol for barrier heights, so molecules with large storage densities but computed barrier heights similar to the parent compound are potentially worth testing experimentally.

Screening all singly and doubly substituted molecules

We start by screening all 35,588 singly and doubly substituted DHAs because they are most synthetically accessible and can be screened without machine learning models using semiempirical electronic structure methods (SQM). Figure 3(a) shows a plot of 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 Supplementary Information, PM3 tends to overestimate the barrier relative to the parent molecule (the horizontal red 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 red 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 (15 kJ/mol higher than the parent compound) and 0.33 kJ/g leads to 109 molecules for further study using DFT (green box in Figure 3(a)). For all 109 molecules we optimize the lowest energy GFN2-xTB DHA, VHF, and PM3 TS structures with M06-2X/6-31G(d) and the results are shown in Figure 3(b). As expected, the majority of the molecules have higher energy storage densities than the parent compound, but some have lower barriers with DFT. This indicates that by using PM3 energies, we are more likely to get false-positives than false-negatives. More false-positives will increase the number of redundant calculations; however, this is preferred over missing potentially excellent candidates.

From this set we choose ten molecules (the purple box in Figure 3(b), and Table S1) for more thorough examination by performing a systematic conformational search (see Computational Methodology section) and computing enthalpies and free energies for the lowest energy conformers using M06-2X/6-31G(d).

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

Figure 3(a) shows 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 as for the ten promising molecules. The results are summarized 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 amino group in position 2 have been shown to yield a large increase in storage energy, but also results in a significant decrease of the back reaction barrier [11]. 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

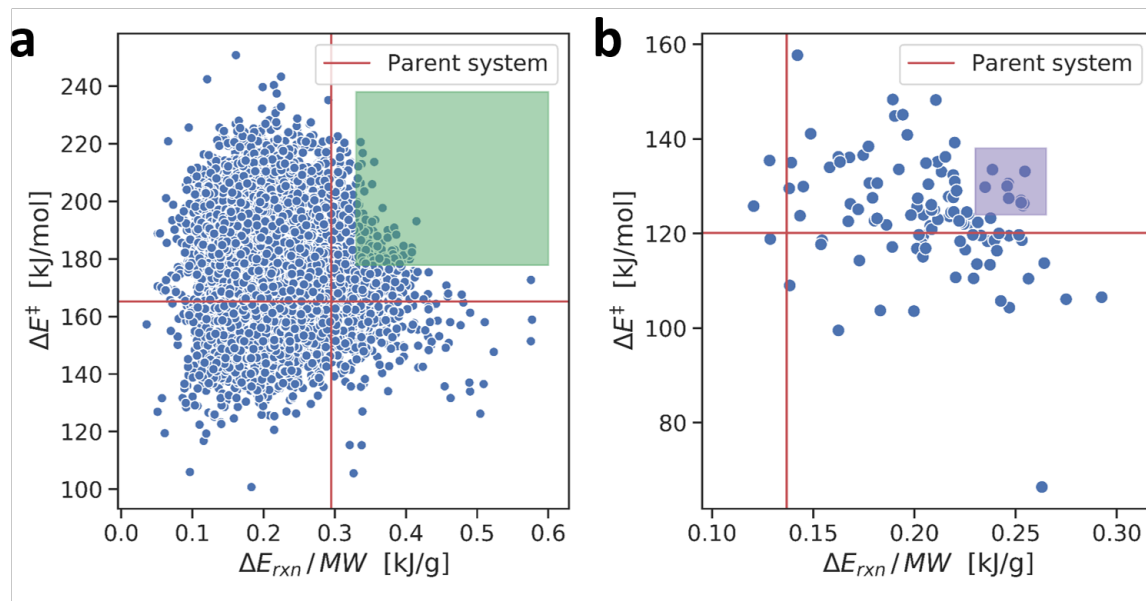
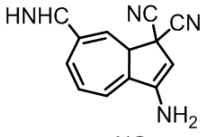
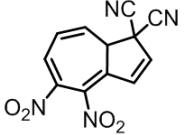
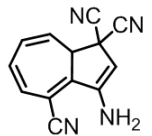
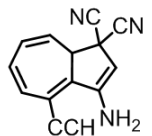
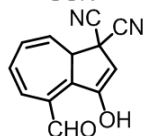
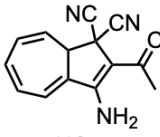
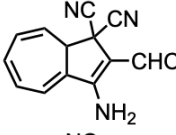
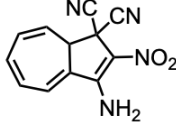



Figure 3: (a) Electronic barrier heights and storage energies computed for 32,623 singly- and doubly substituted molecules using SQM and "$5 + 5n_{rot}$" conformations as described in the Computational Methodology section. (b) Electronic barrier heights and storage energies computed for the 109 molecules 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 S1). The vertical and horizontal lines mark the storage energy and barrier height for the parent molecule illustrated in Figure 1.

very similar to the transition state structure, making the back reaction barrier very small.

To summarise, given the set of ligands shown in Figure 2 the largest energy density one can hope to achieve with a singly or doubly substituted DHA molecules is about 0.5 kJ/g. But the half-life of the high energy states are much to short (sub-second) to be of practical use. The largest possible energy density for molecules with half-lives on the order of hours or greater is around 0.25 kJ/g.

Screening all 230 billion molecules

To exhaustively screen the entire chemical space a fast estimation of the storage density and back-reaction barrier is needed. We represent each molecule as a 287 (7 x 41) bit vector, where each row (7) represents one of the seven open positions on the DHA motif, and each column (41) represents a non-hydrogen substituent. Using this representation, we fit two linear regression models to SQM storage densities. One model predicts storage densities for molecules with ≤4 substituents and the other model for molecules with more than 4 substituents. We found that two different models are needed to get an acceptably low error (MAEs of 0.02 and 0.04 kJ/g, respectively). Each model is trained on ∼25,000 molecules and tested on another ∼25,000 molecules that represent a wide range of storage densities (Figure S4). Figure S5 shows the number of times a ligand is found at a given

Table 1: M06-2X/6-31G(d) predicted storage densities and back reaction barrier heights for the five molecules with largest storage densities among the molecules highlighted in Figures 3(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.

Identifier	Structure	$\Delta H^{\circ}_{rxn}/MW$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
1		0.25	112.8
2		0.25	135.6
3		0.25	123.8
4		0.24	124.2
5		0.24	129.9

position in the training set for the linear regression models. Though the distribution is somewhat uneven, all ligands are represented at all positions at least 40 times, so the training set is reasonably representative of the entire chemical space The RMSD values of the 1-4 substituent model and the 5-7 substituent model are 0.025 and 0.050 kJ/g, respectively (Figure S6).

The simplicity of the linear ML-models means that they can be used to perform an exhaustive search of all 230 billion molecules in only 12 hours using a single CPU. During the exhaustive search, all molecules with predicted storage densities smaller than 0.30 kJ/g are discarded (Figure 4(a)). With the linear regression model, we reduce the chemical space of interest from 230 billion to roughly 421 million molecules. As seen from Table S2, molecules with 5-7 substituents greatly outnumber molecules with 1-4 substituents.

The 421 million molecules are still far too many to screen using SQM methods, and the results from the linear model of 5-7 molecules indicate some degree of non-linearity (which are the group that the majority of the remaining molecules belong to). Thus, to more accurately predict the storage densities of the remaining molecules, we train a new model using the gradient boosted tree method, LightGBM [12]. The LightGBM model is trained on new SQM data chosen among the remaining 421 million molecules. The

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 3(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

Identifier	Structure	$\Delta H^{\circ}_{rxn}/MW$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
6		0.51	55.0
7		0.51	58.1
8		0.50	55.4

new SQM data is chosen, such that the storage densities are evenly distributed between the minimum and maximum storage densities (0.30 kJ/g - 0.57 kJ/g) as predicted by the linear regression model (Figure S7). The LightGBM model is trained on 39,057 molecules and validated on 6,510 molecules, respectively. Finally, the new model is tested on 6,510 molecules and shown to have an RMSD value of 0.036 kJ/g (Figure 5a), lower than the 0.050 kJ/g RMSD for the linear regression model for molecules with 5-7 substituents. The LightGBM model is then used to predict storage densities of all 421 million molecules in approximately 10 hours.

Based on the storage densities predicted using the LightGBM model, we select 36,000 molecules and compute PM3 back reaction barrier heights to train and test an new Light-GBM model. There is a rough inverse correlation between storage densities and barriers to the back reaction, so the molecules are selected in such a way that all storage density ranges are represented (Figure S7). It proved challenging to converge all 36,000 transition state calculations, therefore the barrier heights are estimated using adiabatic scans along with the breaking bond. A LightGBM model was trained, validated, and tested using a 83/7/10 split of the data and yielded a RMSD of 21.21 kJ/mol (Figure 5b). The model is used to predict the back reaction barriers of all the 421 million remaining molecules.

The LightGBM energies are then used to select molecules with a storage density of $\geq$0.40 kJ/g and a barrier height of $\geq$165 kJ/mol (the PM3 barrier for the parent system), of which there are 957,587. The highest storage density found in this set is 0.58 kJ/g so the main conclusion is that there are no molecules among the 230 billion molecules considered here with a storage density approaching 1 kJ/g. From these roughly 1 million molecules we need to select at most 50,000 molecules for SQM refinement. Figure 6 shows an overview of the number of molecules with various ranges of barrier heights and

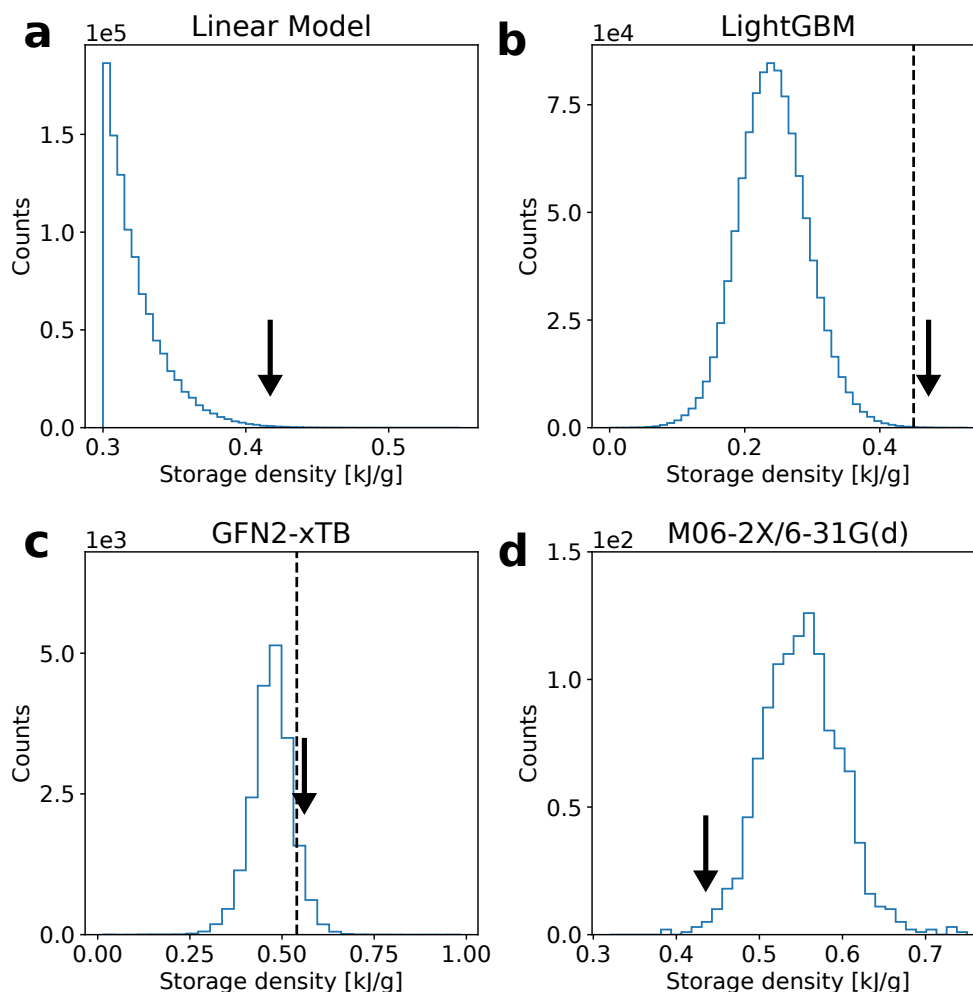



Figure 4: Distribution of storage densities for a) the linear regression model, b) the LightGBM regression model, c) GFN2-xTB, and d) DFT M06-2X/6-31G(d). The black arrow indicates the storage density for **9** and the dashed line the cut-off used to select molecules for further study. The distribution of a) and b) does not show all 421 million molecules, but only a subset of 1 million randomly chosen molecules.

storage densities. After much deliberation we chose two sets for SQM refinement. The first set is the 1969 molecules with barriers and storage densities of >165 kJ/mol and >0.50 kJ/g. Given the tendency of PM3 (and, hence, the LightGBM model) to overestimate the barrier height, these are almost certainly false positives but, given their high storage densities, we include them "just in case" The second set is the 33,085 molecules with barriers and storage densities of >185 kJ/mol and >0.45 kJ/g (Figure 4(b)), which, given the higher barriers, are more likely to include true positives.

The combined sets of molecules contain 34,241 unique molecules, and out of these it was possible to compute the SQM storage density for 24,388 molecules. The remaining molecules did either not converge, or the connectivity changed during the optimization, and are therefore unlikely to be good candidates. For the remaining 22,258 molecules with storage densities above 0.4 kJ/g, we performed an adiabatic scan with PM3 to rapidly

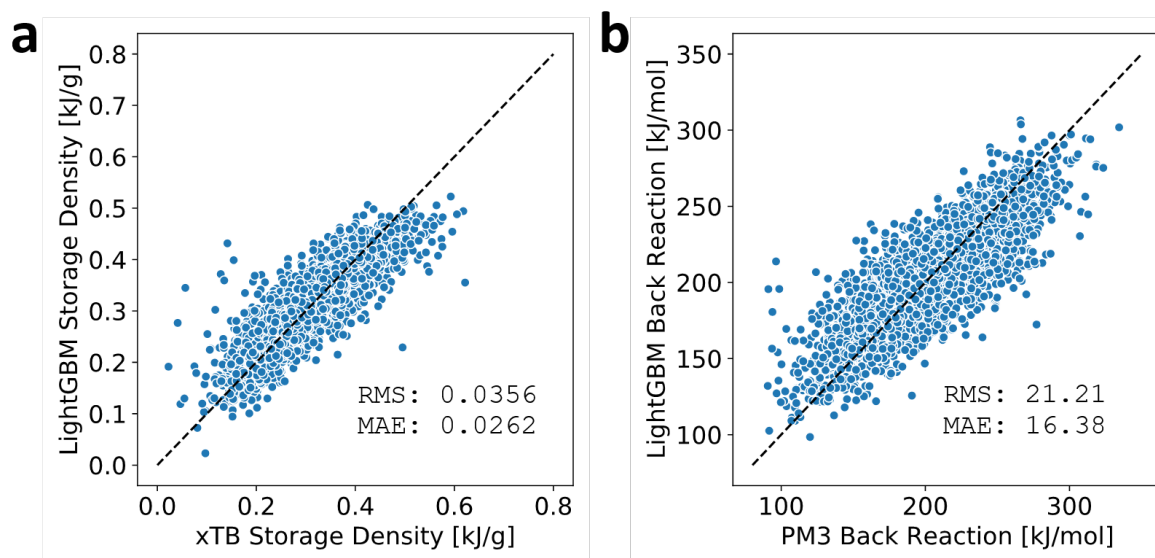



Figure 5: Illustration of the performance of the LightGBM models. a) shows the actual storage density computed using GFN2-xTB with $5+5n_{rot}$ conformers compared to the predicted storage density using the LightGBM. b) compares the back barrier energy found using an PM3 adiabatic scan along the breaking bond to the predicted LightGBM model.

Barrier (scan) [kJ/mol] / Density [kJ/g]

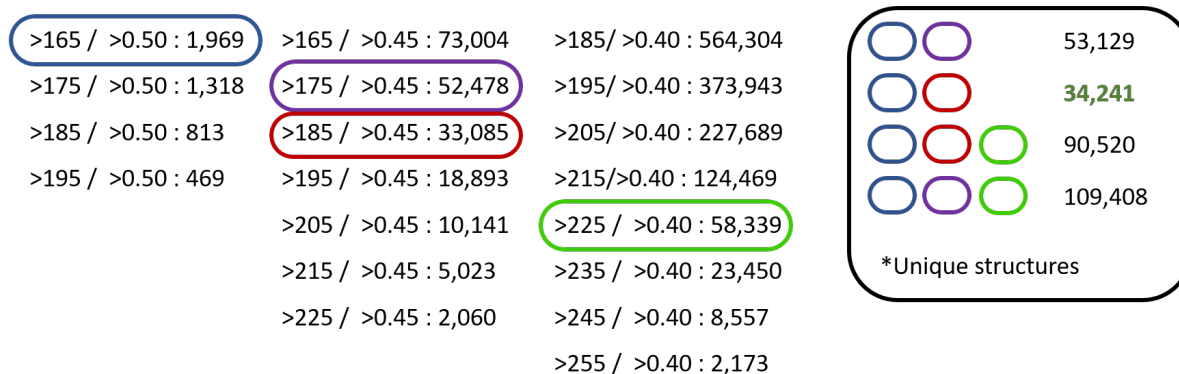
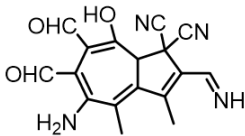



Figure 6: Number of molecules with barriers and energy densities above certain thresholds. The inset on the right shows the number of unique molecules for various combinations of subsets.

estimate the back reaction barrier height.

The next step is to select around 1500 molecules for DFT calculations. Through experimenting with cut-offs, we pick 1177 molecules with storage densities ≥ 0.54 kJ/g (Figure 4(c)) and barriers ≥ 185 kJ/mol. The ground state SQM structures and transition state guess for the chosen molecules are then reoptimized with DFT. The optimizations succeed for 954 molecules, while the remaining 223 molecules fail due to the location

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

Identifier	Structure	$\Delta H^{\circ}_{rxn}/MW$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
9		0.38	130.0

of a wrong transition state (202 molecules) or because the connectivity of the product molecule changed during the optimization (21 molecules). We select random 20 molecules with wrong transition states and the 21 molecules with wrong products and perform a 15 point adiabatic scan along with the breaking bond in steps of 0.1 Å, staring from DHA. None of the 41 molecules showed any indication of a high-energy transition state by visual inspection of the scans, which indicates that the back reaction barrier is very low. Thus, the 223 molecules are not investigated any further.

Of the 954 molecules the majority have back reaction barriers that are smaller than the parent molecule. However, we do find one molecule, shown in Table 3, with a relatively high storage density and barrier height (Figures 4(d) and S9). The molecule is subjected to a more thorough DFT investigation, as described in the supporting information, and the resulting reaction enthalpy and Gibbs free activation energy is shown in Table 3. The best thermal heat battery candidate is therefore **9** with a predicted energy density of 0.38 kJ/g and a barrier height of 130.0 kJ/mol. For comparison the energy density of the parent system (0.14 kJ/g) is almost three times lower, while the barrier height (119.1 kJ/mol) is 11 kJ/mol lower. According to TS theory, an 11 kJ/mol increase in barrier height increases the half life by a factor of 85, i.e. from 3-39 hours (depending on solvent) to about 10 days - 4 months. The computed absorption spectrum of **9** is shown in Figure 7, together with that of the parent system. Compound **9** retains an absorption peak abound 350 nm, but the intensity is reduced by about a factor of three.

Test of genetic algorithm search

While it is possible to exhaustively screen 42 substituents at seven positions, it will not be possible for future screening efforts that involve all nine positions and/or more than 42 substituents since the number of molecules will number in the trillions or more. We therefore test whether a genetic algorithm (GA) can find the best candidate in such a large data set efficiently, using the linear regression model since we know what the correct answer is. To screen for both high storage density and long half life, we train a linear regression model to predict the back reaction barrier height using same molecules we used to train the storage density model (Figure S8). Since the model is only used to test the efficiency of GA searches, we trained one model (instead of two as is done for the storage

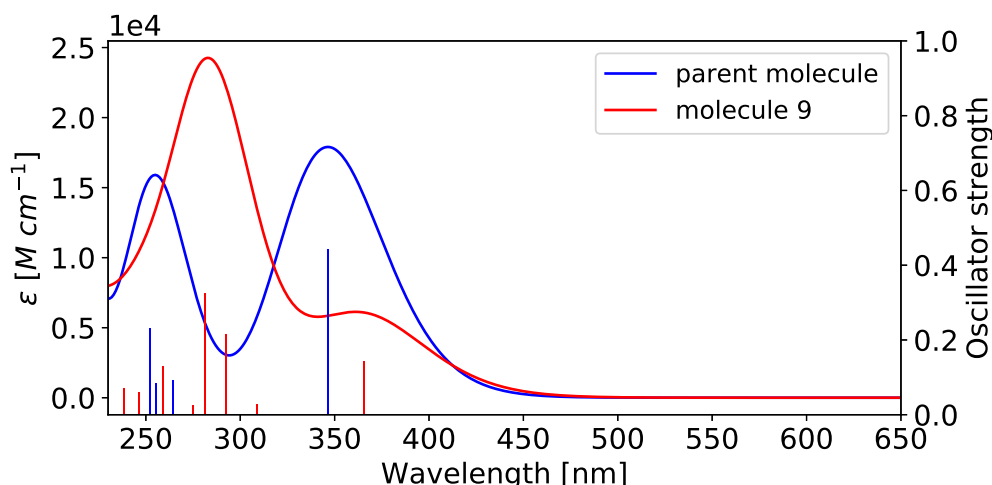
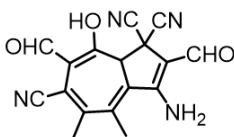



Figure 7: Computed (CAM-B3LYP/6-311+G(d)//M06-2X/6-31G(d)) absorption spectra of compound **9** and the parent compound.

density).

The best candidate in the search space is defined as the molecule with the largest storage density and a barrier height $\geq$180 kJ/mol. From the exhaustive search we know that this molecule is **10** shown in Table 4 with a storage density of 0.531 kJ/g.

Table 4: Structure, predicted storage density, and back reaction barrier height for the molecule with largest storage density and barrier heights above 180.0 kJ/mol as predicted by the linear regression models. The last columns gives the rediscovery percentage during 1,000 GA runs.

Identifier	Structure	$\Delta E_{rxn}/MW$ [kJ/g]	$\Delta\Delta H_f^{\ddagger}$ [kJ/mol]	Rediscovery %
10	![structure of molecule 10]	0.531	190.37	74.0

For the genetic algorithm each gene is defined as a vector with seven numbers (bases) ranging from 0 to 41, representing the seven possible positions and 42 possible substituents. The score for each gene is computed as

$$score = \min(\Delta E_{\mathrm{rxn}}, t_{\mathrm{rxn}})/t_{\mathrm{rxn}} + \min(\Delta E^{\ddagger}, t_{\ddagger})/t_{\ddagger} \qquad (1)$$

where $t_{\ddagger}$ and t_{rxn} are set to 180 kJ/mol and 1.0 kJ/g, respectively.[13] The GA search starts with an initial population of 100 randomly generated genes of singly substituted molecules, for which the scores are computed using the linear regression models. Pairs of genes are then selected with a probability proportional to their score (roulette selection)

and mated, by choosing a random cut point between bases for two parent genes and recombining. After each mating a mutation is performed 20% of the time by randomly changing one of the bases. This process is repeated 100 times and the 100 best scoring genes before and after mating is selected as the new population and the process is repeated for 100 generations.

We perform 1,000 separate GA searches and record the highest scoring molecule. **10** is found 74% of the time. Thus, one should be able to find the best molecule with 99% certainty by running only four GA searches, i.e. by testing at most 40,000 different molecules. For comparison, we calculated storage densities and back-reaction barriers for >100,000 and >70,000 molecules as part of training and testing the machine learning models and their predictions. Thus, it may be computationally more efficient to search the chemical space using GA combined with SQM calculations rather than developing ML models for the properties of interest.

Computational Methodology

Semiempirical Calculations

ΔH_{rxn}° is approximated as the difference in electronic energy (ΔE_{rxn}) computed using GFN2-xTB[14] while $\Delta G^{\circ,\ddagger}$ is approximated as the difference in heat of formation ($\Delta \Delta H_f^{\ddagger}$) computed using PM3[15] (collectively referred to as "SQM"). GFN2-xTB is chosen due to its computational efficiency, while PM3 is chosen because it is available in both ORCA 4.0[16] and Gaussian09[17] (see below). For each DHA structure the VHF structure is automatically generated using RDKit. $5 + 5n_{rot}$ random conformations (where n_{rot} is the number of rotateable bonds in the molecule) are generated using RDKit and optimized using GFN2-xTB. Optimizations that result in discrepancies between the input and output connectivity are discarded. The lowest energy conformers of DHA and VHF are used to compute ΔE_{rxn}. In some cases the conformational search is done systematically (see next subsection) at the SQM as described below.

To compute the energy barrier an 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. When screening all 230 billion molecules the highest energy structure and energy is used as an estimate for the transition state. However, for the single and double substituted case the highest energy structure is used as starting point for a transition state (TS) search using PM3 in Gaussian09 while computing the Hessian in each step. For the molecules where the TS search converged it is verified that the normal mode associated with the imaginary frequency lies along the reacting CC-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 is that 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).

 Chemistry Journals
Analytical, Inorganic, Organic, Physical, Materials Science

Figure S3 shows a comparison of barriers computed by locating the TS to barriers estimated by scans. There is a good correlation between the two barriers for medium sized barriers, but the scan tends to overestimate high barriers and underestimate low barriers. The use of scans to estimate barriers is this likely to lead to false positives, which are subsequently eliminated by DFT calculations, but is unlikely to lead to false negatives (i.e. we are unlikely to miss any promising candidates by using estimated barriers).

DFT refinement

Select structures are investigated further at the M06-2X[18]/6-31G(d)[19] level of theory(using Gaussian09). This level of theory was chosen as a good compromise between computational efficiency and accuracy as judged by comparison to DLPNO-CCSD(T) and CCSD(T)-F12a calculations[20]. DFT is used in one of three ways. The first quick approach estimates DFT energies by reoptimizing the structure with the lowest SQM energy for DHA, VHF, and the TS and the electronic energy is used to compute the storage density and barrier. A benchmark investigation has been carried out by [11].
The second approach is a more thorough DFT searches, used to obtain the final DFT energies. In the second approach a systematic conformer search is performed using SQM where each rotateable bond is rotated by $\pm 120°$ 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[21] 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

We use both linear regression and a Gradient Boosted tree method (LightGBM) in this study. For the linear regression, a molecule is represented by positional seven-hot encoding, i.e. vector with 287 (41×7) binary elements, where each chunk of 41 is one-hot encoded representation of a non-H substituent at a particular position. This representation was used to train three different machine learning (ML)-models using SQM data. Linear regression and kernel ridge regression as implemented in Scikit-learn and a gradient boosted tree method as implemented in LightGBM. We found very little difference in performance between regression and kernel ridge regression and use the former, simpler, model in this study.

The combined use of positional binary encoding ($\mathbf{X}$) and linear regression amounts to an additive model

$$\Delta E = \mathbf{w} \cdot \mathbf{X} + b = \sum_{i=1}^{7} w_{ij} + b \qquad (2)$$

where the regression coefficient w_{ij} is the effect of placing substituent j on position i on either the reaction energy or barrier height and b is the corresponding value for the

unsubstituted molecule.

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 230 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 S1).

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 3(a)). The energy densities and barrier heights computed by reoptimising the lowest energy SQM-conformations are then used to select 10 molecules for further study using a thorough conformational search (Figure 3(b) and Table S1). 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 space we generate additional SQM-data for higher degrees of substitution and use it to fit linear regression models that reproduce the storage energies to within 0.017 and 0.038 kJ/g depending on degree of substitution (Figure S6). These models are then used to estimate the storage density of all 230 billion molecules and the 421 million molecules with storage densities higher than 0.30 kJ/g are selected for further study (Figure 4(a)). Gradient boosted tree method (LightGBM) models for the storage density and back reaction barrier are trained on new SQM data chosen among the these 421 million molecules, with respective MAEs of 0.026 kJ/g and 16.4 kJ/mol. These models are used to predict the storage densities and barrier heights for all 421 molecules, and 34,241 molecules with storage densities larger than 0.45 kJ/g and high barrier heights are chosen for further study (Figures 4(b) and 6). The highest storage density found in this set is 0.58 kJ/g so it is already clear at this point that there are no molecules among the 230 billion molecules considered here with a storage density approaching 1 kJ/g. The storage densities of the 34,241 molecules are computed using SQM and the barrier heights for the 22,258 molecules with SQM storage densities above 0.4 kJ/g are estimates using PM3.

Next, 1177 molecules with storage densities $\geq$0.54 kJ/g (Figure 4)) and barriers $\geq$185 kJ/mol are selected for further refinement using DFT. The lowest energy geometry of

DHA and VHF is reoptimized using M06-2X/6-31G(d), which also used to estimate the barrier height using adiabatic scans. Of the 954 molecules for which the DFT calculations succeed, the majority have back reaction barriers that are smaller than the parent molecule. However, one molecule, shown in Table 3, has a relatively high storage density and barrier height (Figures 4(d) and S9) and is subjected to a more thorough DFT investigation.

Our conclusion is that the best thermal heat battery candidate among the 230 billion is **9** (3) with a predicted energy density of 0.38 kJ/g and a barrier height of 130.0 kJ/mol. For comparison the energy density of the parent system (0.14 kJ/g) is almost three times lower, while the barrier height (119.1 kJ/mol) is 11 kJ/mol lower. According to TS theory, an 11 kJ/mol increase in barrier height increases the half life by a factor of 85, i.e. from 3-39 hours (depending on solvent) to about 10 days - 4 months. The computed absorption spectrum of **9** is shown in Figure 7, together with that of the parent system. Compound **9** retains an absorption peak abound 350 nm, but the intensity is reduced by about a factor of three.

The main conclusion of our work is that it is unlikely that the storage density of DHA can be increased to a value above 0.5 kJ/g by substitution at positions 1-7 (at least without decreasing the barrier to back reaction). Yet, storage densities above 0.3 kJ/g of both molecules **9** and **10** may be sufficient for some applications when considering also their long storage times. There are, however, other drawbacks with these molecules. For example, their absorption profiles are not optimum. And maybe more problematic, these molecules have rather complicated substitution patterns. There are today synthetic protocols available for functionalizing selectively at positions 1, 2, 3, and 7 (and in part position 6) of DHA,[2] but so far these protocols have only been used to introduce substituents efficiently at maximum three positions (see SI for further synthetic considerations). So new synthetic protocols need to be developed for introducing substituents around the entire DHA core, and some suitable protecting groups need to be installed, for example to allow both amino and aldehyde functionalities in the same molecule. To avoid intermolecular reactions between such groups, it could be attractive to simply make **9** and **10** part of the monomeric repeat units of a polymer scaffold via the amine functionality. Organization of DHA units along a polymer may in fact possibly also enhance the energy density further as observed for some azobenzene-based materials.[22] A Other design strategies will also be pursued in future work, such as replacement of the cyano groups at position 1 or the introduction of non-carbon atoms in the DHA scaffold (heterocyclic structures).

To our knowledge this compound library is the largest library considered thus far and the first to include barrier heights as a screening parameter. Notably, we show that the substituent effects on barrier heights can be estimated using ML and a very simple representation with sufficient accuracy to be useful. Despite the huge library size the screen is carried out using comparatively modest computational resources by using SQM as an intermediate step between ML and DFT calculations. SQM allows for large data sets for ML models to be constructed in a matter of days and the pre-screening tens of thousands of candidates so that DFT calculations are only performed on the most promising candi-

dates. This is important because the prediction of reliable reaction energies and barrier heights requires thorough conformational searches that require significant computational resources at the DFT level. The overall methodological approach outlined in this paper is generally applicable to lead optimisation of properties involving chemical reactivity.

Supplementary information

Additional figures referred to in the text can be found in SI. The code and data resulting from this study can be found here https://github.com/jensengroup/dha_htvs and https://sid.erda.dk/sharelink/EwaEr2JMrb, respectively.

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). We thank Maria Harris Rasmussen for helpful discussions.

References

[1] 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.

[2] Mogens Brøndsted Nielsen, Nicolai Ree, Kurt V Mikkelsen, and Martina Cacciarini. Tuning the dihydroazulene–vinylheptafulvene couple for storage of solar energy. *Russian Chemical Reviews*, 89(5):573, 2020.

[3] Timothy J Kucharski, Yancong Tian, Sergey Akbulatov, and Roman Boulatov. Chemical solutions for the closed-cycle storage of solar energy. *Energy Environ. Sci.*, 4(11):4449–4472, October 2011.

[4] V A Bren, Alexander D Dubonosov, Vladimir I Minkin, and V A Chernoivanov. Norbornadiene–quadricyclane — an effective molecular system for the storage of solar energy. *Russian Chemical Reviews*, 60(5):451–469, May 1991.

[5] 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*, xi:51–67, 2011.

[6] 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.

[7] 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.

[8] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A Tolmachev, Brian K Shoichet, Bryan L Roth, and John J Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019.

[9] Adrian Jinich, Benjamin Sanchez-Lengeling, Haniu Ren, Rebecca Harman, and Alán Aspuru-Guzik. A mixed quantum Chemistry/Machine learning approach for the fast and accurate prediction of biochemical redox potentials and its Large-Scale application to 315 000 redox reactions. *ACS Cent Sci*, 5(7):1199–1210, July 2019.

[10] Terry Z H Gani and Heather J Kulik. Understanding and breaking scaling relations in Single-Site catalysis: Methane to methanol conversion by FeIV=O. *ACS Catal.*, 8(2):975–986, February 2018.

[11] 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.

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems 30*, pages 3146–3154, 2017.

[13] 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.

[14] 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.

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

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

Peer J Chemistry Journals

Analytical, Inorganic, Organic, Physical, Materials Science

[17] 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.

[18] 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.

[19] 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.

[20] Mads Koerstz, Jonas Elm, and Kurt V Mikkelsen. Benchmark study of the structural and thermochemical properties of a Dihydroazulene/Vinylheptafulvene photoswitch. *J. Phys. Chem. A*, 121(16):3148–3154, April 2017.

[21] 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.

[22] Si Wu and Hans-Jürgen Butt. Solar-thermal energy conversion and storage using photoresponsive azobenzene-containing polymers. *Macromolecular Rapid Communications*, 41(1):1900413, November 2019.

[23] Virginia Mazzanti, Martina Cacciarini, Søren L Broman, Christian R Parker, Magnus Schau-Magnussen, Andrew D Bond, and Mogens B Nielsen. On the bromination of the dihydroazulene/vinylheptafulvene photo-/thermoswitch. *Beilstein Journal of Organic Chemistry*, 8:958–966, June 2012.

[24] Josefine Mogensen, Oliver Christensen, Martin Drøhse Kilde, Martin Abildgaard, Lotte Metz, Anders Kadziola, Martyn Jevric, Kurt V. Mikkelsen, and Mogens Brøndsted Nielsen. Molecular solar thermal energy storage systems with long discharge times based on the dihydroazulene/vinylheptafulvene couple. *European Journal of Organic Chemistry*, 2019(10):1986–1993, January 2019.

[25] Nickie C. M. Lubrin, Alexandru Vlasceanu, Benjamin N. Frandsen, Anders B. Skov, Martin Drøhse Kilde, Kurt V. Mikkelsen, and Mogens Brøndsted Nielsen. Dialkylated dihydroazulene and vinylheptafulvene derivatives - synthesis and switching properties. *European Journal of Organic Chemistry*, 2017(20):2932–2939, May 2017.

Supporting Information

The code and data resulting from this study can be found here https://github.com/jensengroup/RMSD_ and https://sid.erda.dk/sharelink/EPvv68fOTp, respectively.

Benchmarking SQM methods

To verify that semiempirical methods can reproduce a trend that is similar to the traditional computational scheme, we make two data sets of 100 random double substituted DHA derivatives each. One data set is selected to test the back reaction barrier, and the other is chosen to examine the storage energy. The 100 samples in each test set are chosen, such that they represent a wide range of storage energies and back reaction barriers. The data sets are separated into three categories: high barriers/storage energies, medium, and low based on the SQM energies. The data is distribution such that there are 20 samples in the high energy category, 60 in the medium, and 20 in the low.

The energies in the back reaction barrier data set is computed with PM3 and M06-2X/6-31G(d), while the storage energies are calculated using GFN2-xTB and M06-2X/6-31G(d). Figure S1 illustrates the DFT energies compared to the corresponding semiempirical energies for (a) the storage energy and (b) the back reaction barrier.

Figure S1(a) shows a clear correlation between the storage energies computed with GFN2-xTB and corresponding DFT energies. This means, that molecules that have large storage densities with GFN2-xTB can be expected to have large DFT storage densities. The correlation between DFT and PM3 back reaction barriers in Figure S1(b) is less clear, since the groups are overlapping. However, molecules that have high barriers with DFT also have high barriers with PM3, which is sufficient for screening purposes. The the vertical and horizontal lines indicate the DFT and PM3 barrier heights for the parent system (119 and 165 kJ/mol) and show that PM3 tends to overestimate the number of molecules with barriers higher than that of the parent compound. Thus, using the a PM3 barrier cutoff of 165 kcal/mol should retain all molecules with a DFT barrier height that is higher than that of the parent compound, including many false positives. Alternatively, the cutoff can be raised up to 180 kJ/mol (dotted horizontal line) to reduce the number false positives without loosing too many true positives.

Benchmarking Conformational Search

To test how the stochastic conformational search compares to the systematic search, we use the same data set used to compare the storage densities. The data set contains 100 different double substituted DHA derivatives, which is chosen such that they represents a wide range of energies. Again, the energies are separated into three categories: 20 molecules with high energy storage, 60 with medium, 20 with low. The energy difference for all 100 molecules is computed with GFN2-xTB using a systematic conformational search rotating ($\theta = 120$ degrees) and the stochastic conformer search. Figure S2 com-

Peer J Chemistry Journals

Analytical, Inorganic, Organic, Physical, Materials Science

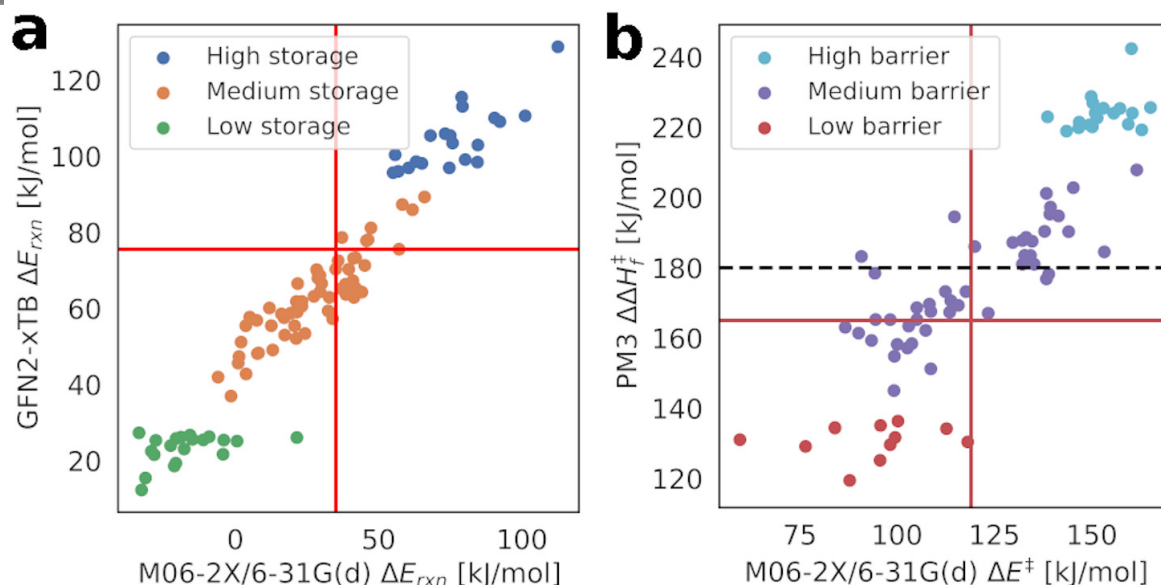



Figure S1: Comparison of GFN2-xTB (a) and PM3 (b) ability to predict the storage energy and barrier compared to DFT.

pares the energy difference between DHA and VHF computed with the systematic search to the stochastic search when (a) 5 and (b) $5+5n_{rot}$ conformers are generated at random, where n_{rot} is the number of rotatable bonds in the molecule.

At first glance, the two graphs seem quite similar, which illustrates that the energy difference is relatively insensitive to the number of random conformers generated. However, at closer inspection, it is apparent that only generating five conformers leads to more outliers compared to creating $5+5n_{rot}$ conformers. As a result, two molecules with high storage energy are misclassified (circled in black in figure S2(a)). The "$5 + 5n_{rot}$-approach in Figure S2(b) follows the trend from the systematic conformer search well. The molecules that deviates from the trend, if anything, overestimates the storage energy, which in the worst case, results in false positives. False positives are not a major concern, because they can be caught at a later stage in the screening procedure.

 Chemistry Journals
Analytical, Inorganic, Organic, Physical, Materials Science

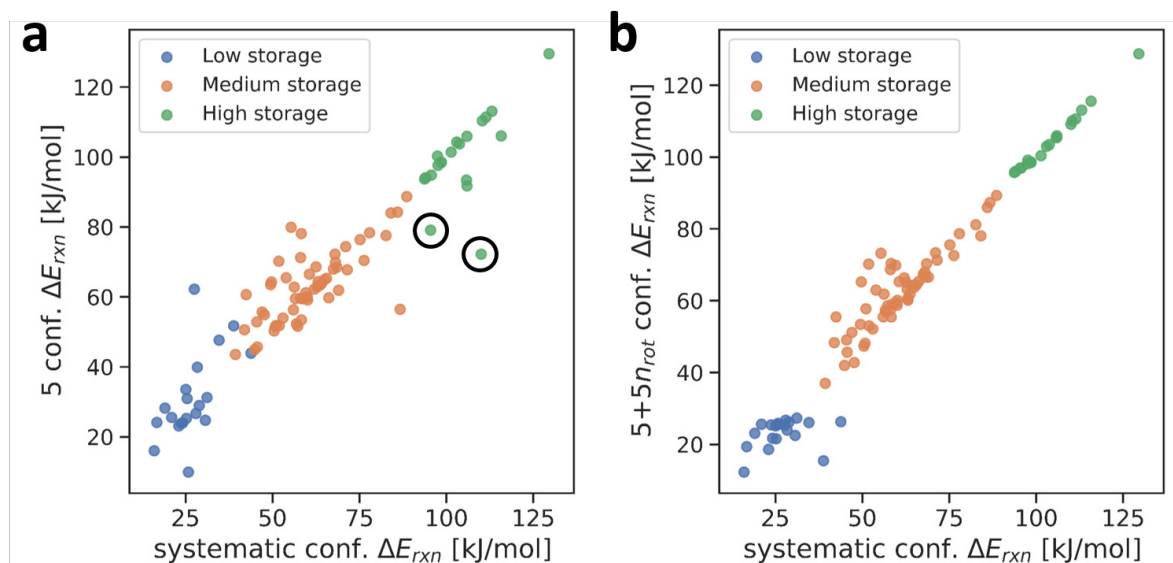



Figure S2: A plot comparing the electronic energy difference between DHA and VHF found using a systematic conformational search and a) 5 or b) $5+5n_{rot}$ randomly generated conformers. The energies are computed using the GFN2-xTB semi-empirical method

Benchmarking barrier hight using a barrier scan

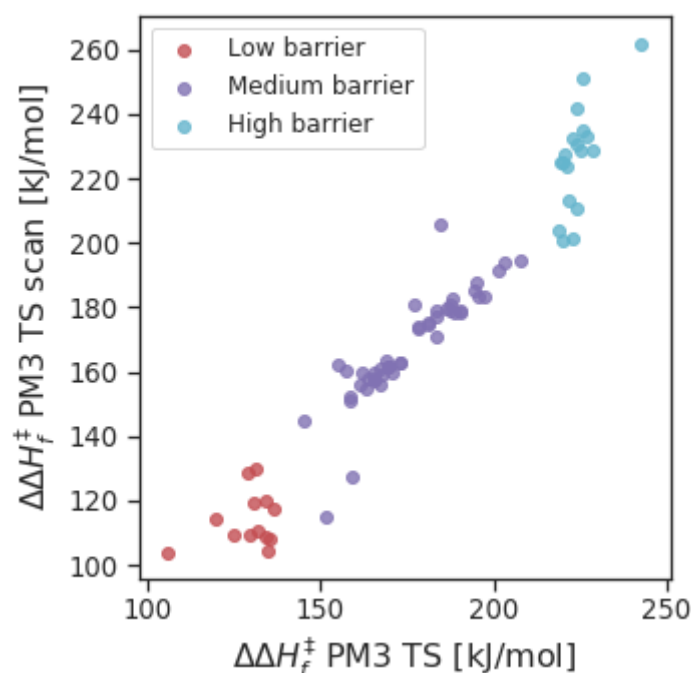
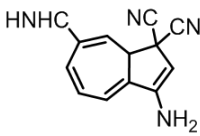
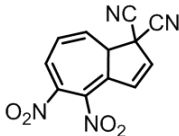
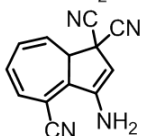
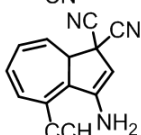
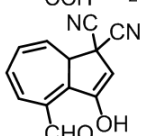
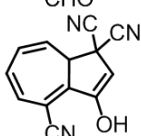
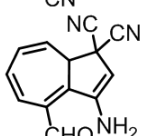
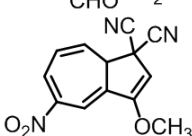
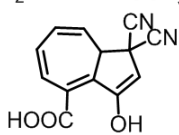
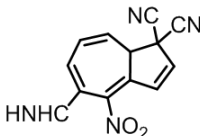



Figure S3: A plot comparing the barrier height computed using the real transition state (PM3 TS) and the energy found during the scan along the breaking bond (PM3 TS scan).

Synthetic considerations

It is clear that some of the found molecules may be challenging to synthesize or isolate. For example, the imine unit of molecule 9 can be subject to hydrolysis. Moreover, the presence of both nucleophilic amino and electrophilic aldehyde functionalities in molecules **9** and **10** may cause these to undergo oligomerization reactions. Yet, instead of protecting the amino group, it could be attractive to simply make the structure part of a monomeric repeat unit of a polymer scaffold via the amine functionality, allowing organization of DHA units along a polymer and thereby not only separating the units from intermolecular reactions, but possibly also enhancing the energy density as observed for some azobenzene-based materials.[22] It is possible to incorporate a bromo substituent at position 3 of DHA,[23] and this compound may serve as a precursor for introducing the amino functionality of **10** by a metal-catalyzed amination. If the aldehyde functionalities are already installed at this stage, they should be protected as for example acetals. A synthetic protocol for introducing a cyano group at position 7 was recently reported.[24] Synthetic protocols for installing aldehyde groups need to be developed (relevant for both **9** and **10**), while a method for introducing acetyl groups at positions 6 or 7 has been reported from an ethynyl-substituted precursor.[24] Functionalization at position 2 of DHA is usually accomplished early in the synthesis via an acetophenone as the key substrate.[5] For the target molecules suggested above, acetophenone would thus have to be replaced by other substrates. We reckon that introduction of alkyl groups in the seven-membered ring can also be done early in the synthesis; maybe from an alkylated tropone as precursor. In this regard, a convenient synthesis of 5,7-dimethyl-DHA has been reported using 2,7-dimethyltropone as starting material.[25] In all, development of functionalization methods of DHA is a very active area, and the target molecules suggested in this work are important for guiding synthetic chemists towards developing new, relevant protocols to be used in combination with existing methods.

Other supplementary figures and tables

Table S1: M06-2X/6-31G(d) predicted storage densities and back reaction barrier heights for the 10 molecules highlighted in Figures 3(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.

Identifier	Structure	$\Delta H^{\circ}_{rxn}/MW$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
1	(HNHC, NC, CN, NH$_2$)	0.25	112.8
2	(NC, CN, O$_2$N, NO$_2$)	0.25	135.6
3	(NC, CN, CN, NH$_2$)	0.25	123.8
4	(NC, CN, CCH, NH$_2$)	0.24	124.2
5	(NC, CN, CHO, OH)	0.24	129.9
S6	(NC, CN, CN, OH)	0.22	123.8
S7	(NC, CN, CHO, NH$_2$)	0.22	131.6
S8	(NC, CN, O$_2$N, OCH$_3$)	0.21	131.7
S9	(CN, CN, HOOC, OH)	0.20	138.8
S10	(NC, CN, HNHC, NO$_2$)	0.17	146.3

Figure S4: Distribution of training and test data for the linear regression model.

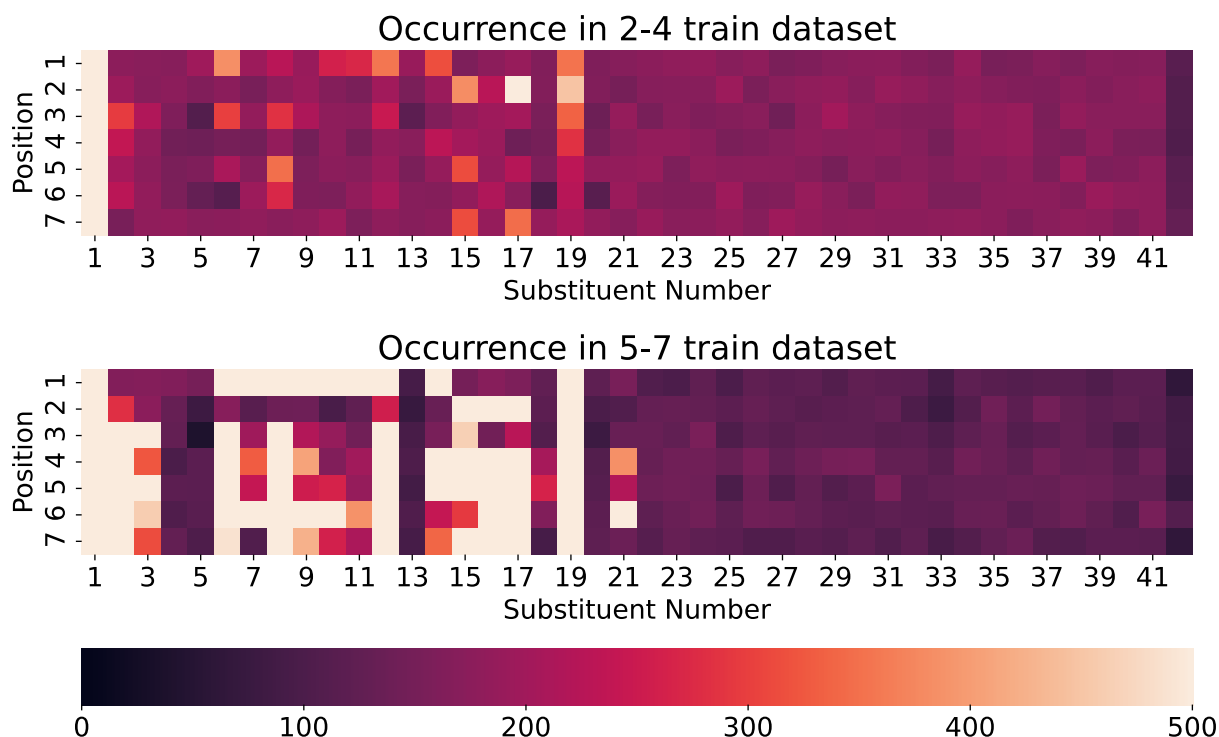



Figure S5: Illustration of how many times a ligand is found at a given position in the linear regression training data. All ligands are represented at least 40 times.

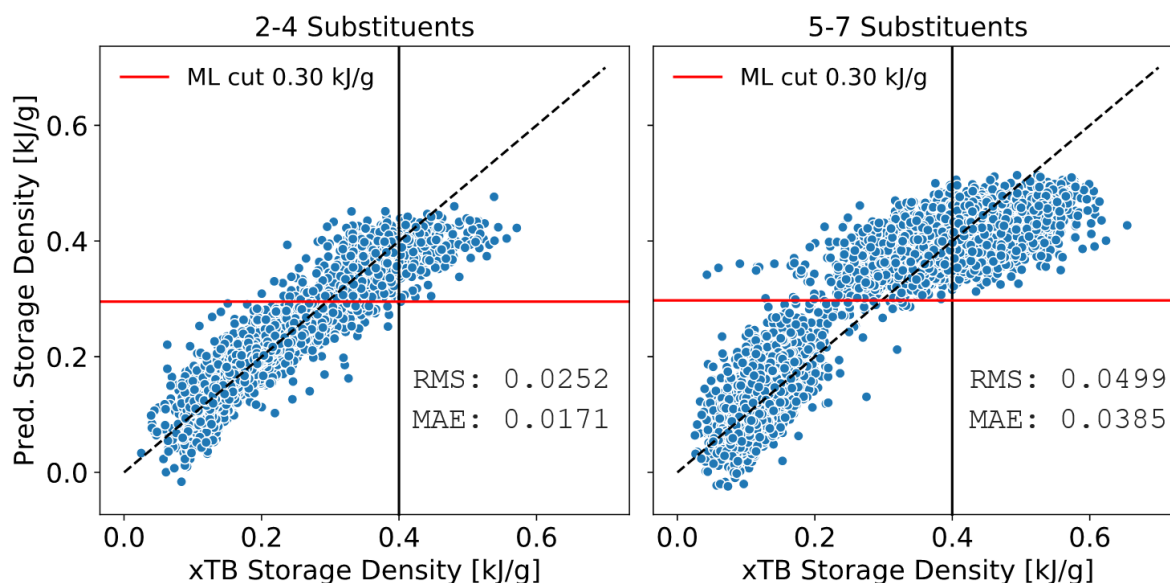



Figure S6: Illustration of the performance of the linear regression models. The actual storage density computed using GFN2-xTB with $5+5n_{rot}$ conformers is compared to the predicted storage density using the linear regression model. The red line corresponds to the cut-off used during the exhaustive search.

Table S2: Amount of molecules for each group of substituted molecules.

Num. Sub.	Molecule count
1	8
2	896
3	47,970
4	1,450,528
5	20,091,801
6	118,620,802
7	281,356,379
Total:	421,568,384

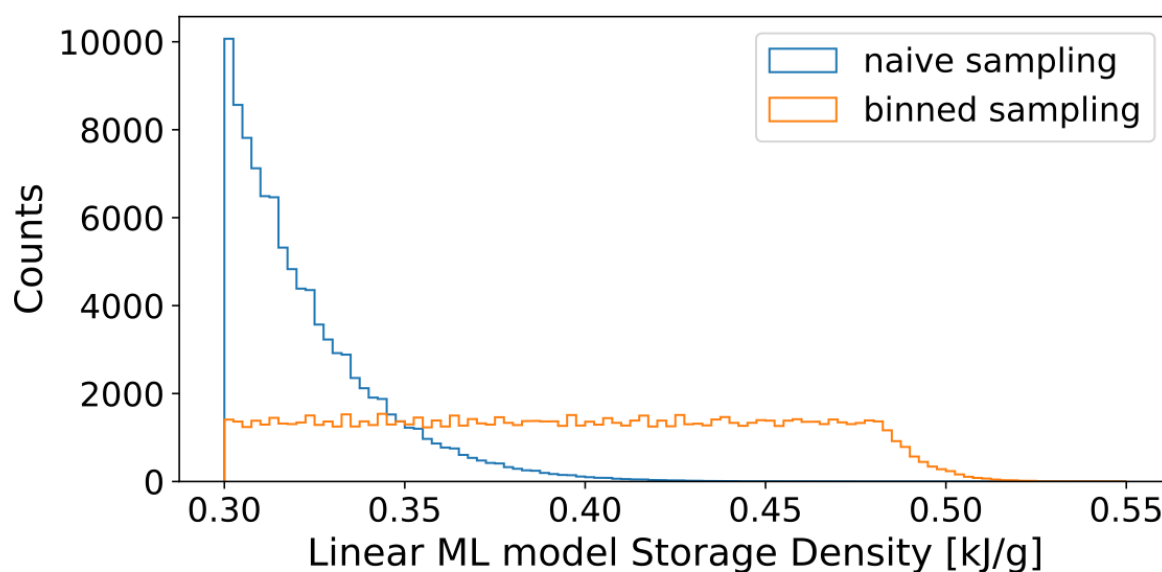



Figure S7: Illustration of binned sampling used to construct the LightGBM training, validation, and test sets.

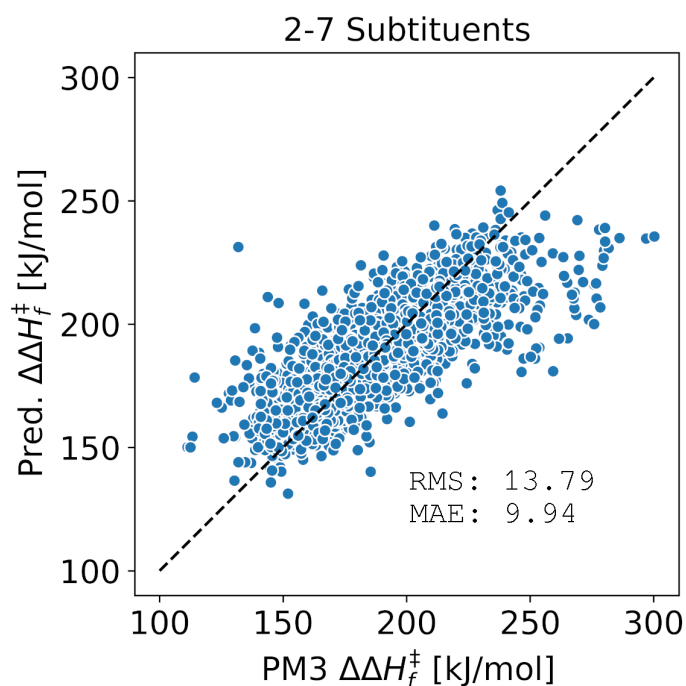



Figure S8: Comparison of the Linear Model ability to predict PM3 heat of formation to the actual PM3 heat of formation.

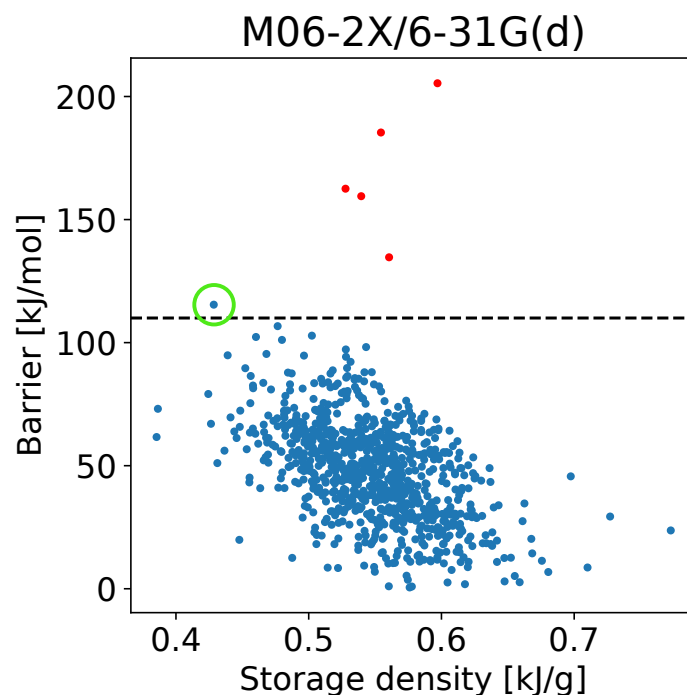



Figure S9: The calculated M06-2X/6-31G(d) storage density and back reaction barrier heights for the 954 molecules included in the first DFT investigation. Molecule **9** is marked with a green circle, the only with back reaction barrier above the cut-off illustrated by the black dashed line (110 kJ/mol). The red molecules were initially labeled as correct by the automatic TS check. However, by visual inspection, the labeling was changed to an incorrect TS.

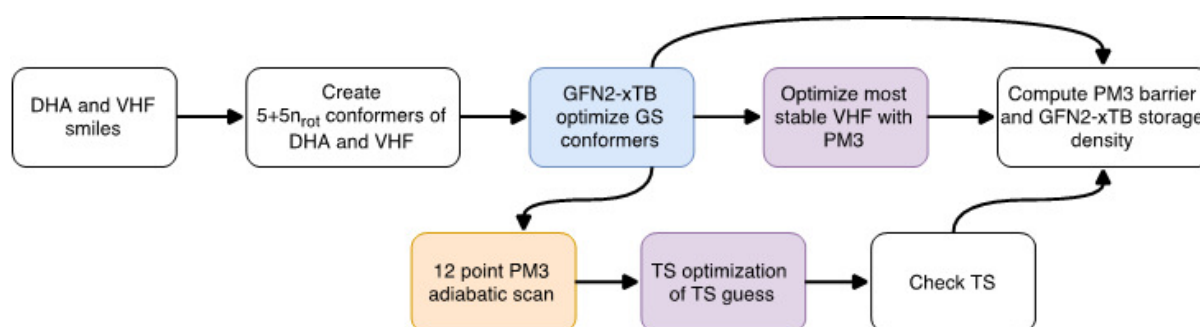



Figure S10: Flowchart illustrating the automatic SQM high-throughput screening procedure.