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.

Flowchart of the exhaustive screening procedure

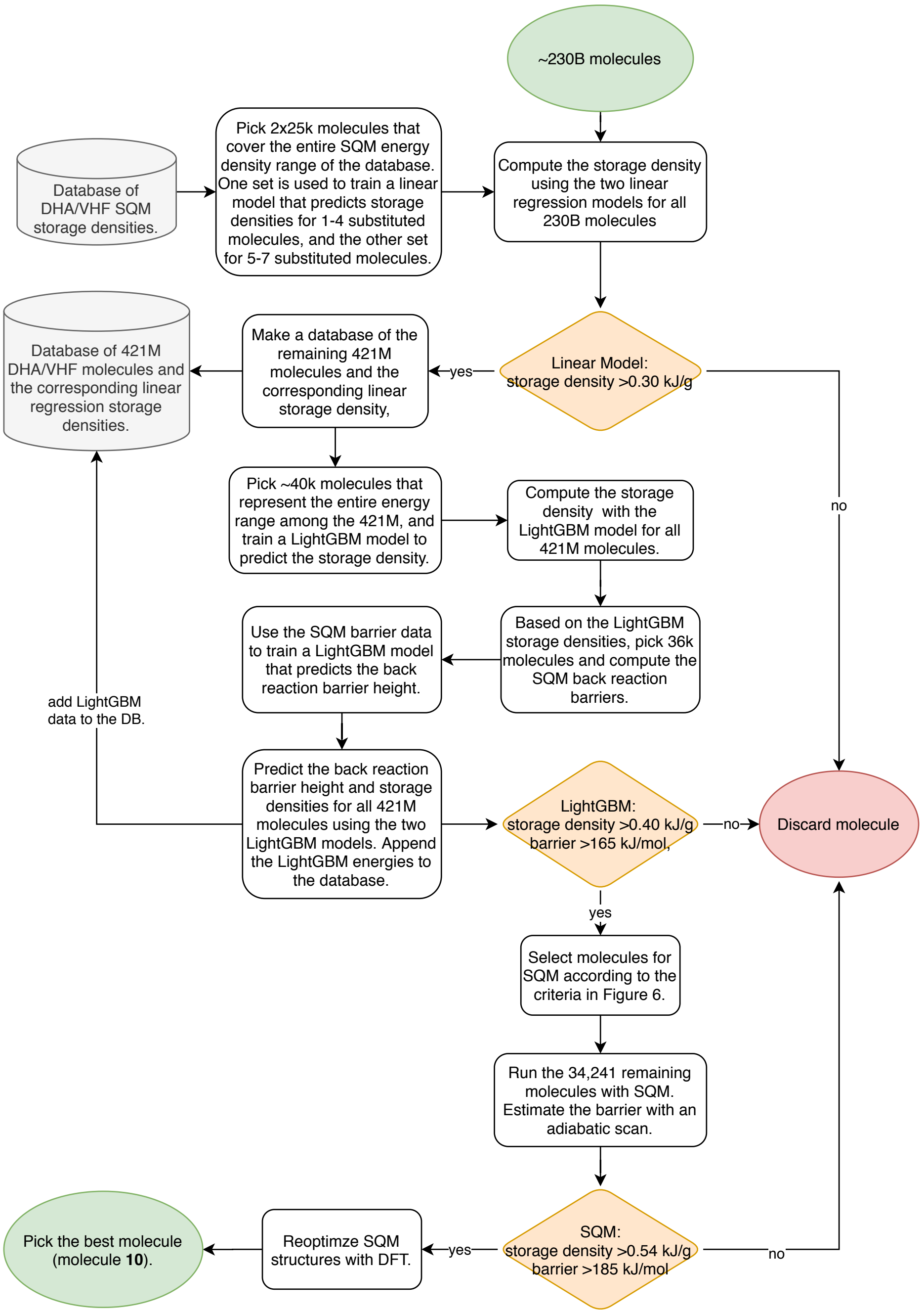
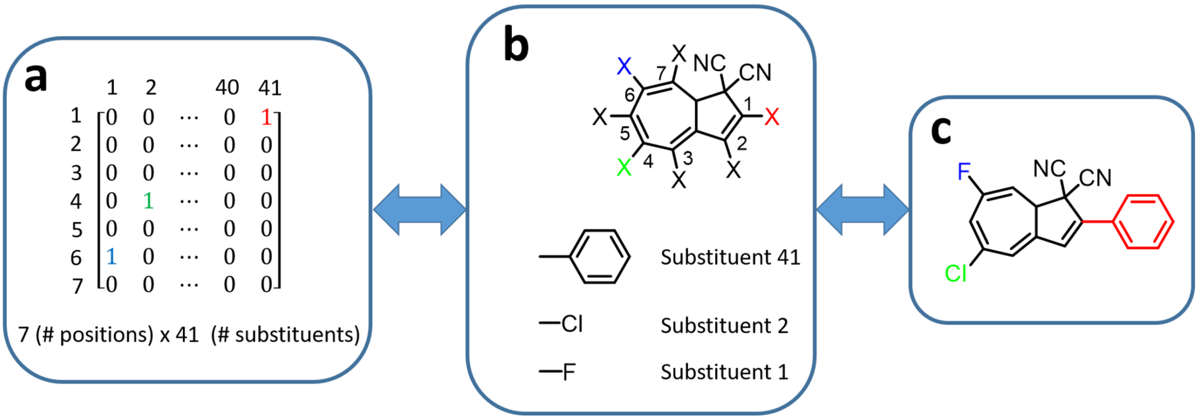

~230B molecules
Database of DHA/VHF SQM storage densities.
Pick 2x25k molecules that cover the entire SQM energy density range of the database. One set is used to train a linear model that predicts storage densities for 1-4 substituted molecules, and the other set for 5-7 substituted molecules.
Compute the storage density using the two linear regression models for all 230B molecules
Database of 421M DHA/VHF molecules and the corresponding linear regression storage densities.
Make a database of the remaining 421M molecules and the corresponding linear storage density,
yes
Linear Model: storage density >0.30 kJ/g
no
Pick ~40k molecules that represent the entire energy range among the 421M, and train a LightGBM model to predict the storage density.
Compute the storage density with the LightGBM model for all 421M molecules.
Use the SQM barrier data to train a LightGBM model that predicts the back reaction barrier height.
Based on the LightGBM storage densities, pick 36k molecules and compute the SQM back reaction barriers.
add LightGBM data to the DB.
Predict the back reaction barrier height and storage densities for all 421M molecules using the two LightGBM models. Append the LightGBM energies to the database.
LightGBM: storage density >0.40 kJ/g barrier >165 kJ/mol,
no
Discard molecule
yes
Select molecules for SQM according to the criteria in Figure 6.
Run the 34,241 remaining molecules with SQM. Estimate the barrier with an adiabatic scan.
Pick the best molecule (molecule 10).
Reoptimze SQM structures with DFT.
yes
SQM: storage density >0.54 kJ/g barrier >185 kJ/mol
no

Figure S1: The flowchart illustrates the entire exhaustive screening procedure from 230B molecular candidates to a single molecule **10**.

Description of the representation used for machine learning



$$\begin{array}{c|ccccc} & 1 & 2 & & 40 & 41 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ 2 & 0 & 0 & \cdots & 0 & 0 \\ 3 & 0 & 0 & \cdots & 0 & 0 \\ 4 & 0 & 1 & \cdots & 0 & 0 \\ 5 & 0 & 0 & \cdots & 0 & 0 \\ 6 & 1 & 0 & \cdots & 0 & 0 \\ 7 & 0 & 0 & \cdots & 0 & 0 \end{array}$$

Figure S2: Illustration of how the ML representation is constructed. a) shows the 7 × 41 one-hot encoded matrix where each row corresponds to an open position in DHA scaffold and the columns a substituent. The example given i a) corresponds to the components in figure b) which is put together to form the molecule in c). With the representation you can build the molecule, and from the molecule you can build the matrix.

Linear regression model: distribution of data

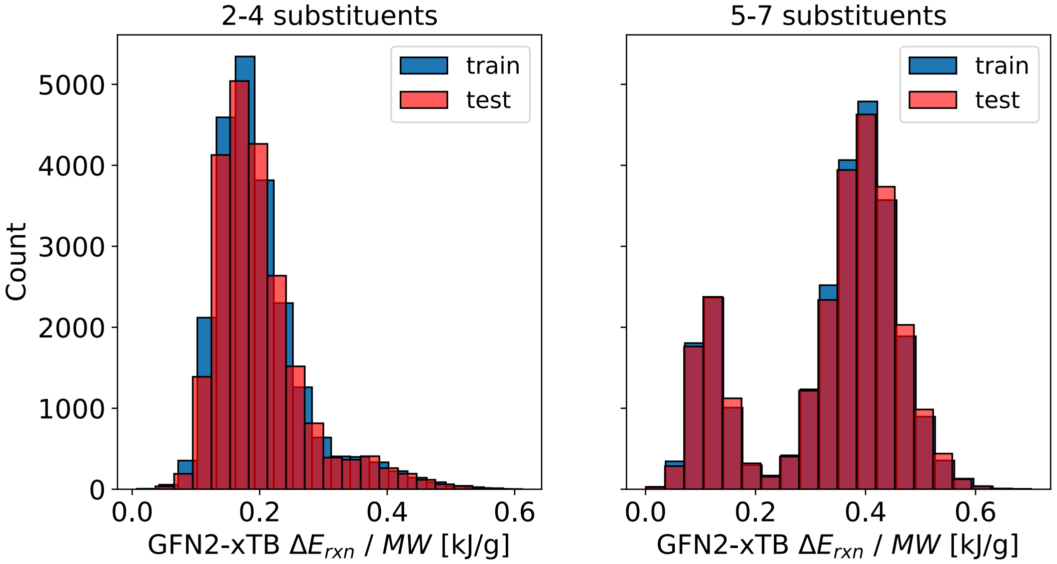


Figure S3: Distribution of training and test data for the linear regression model.

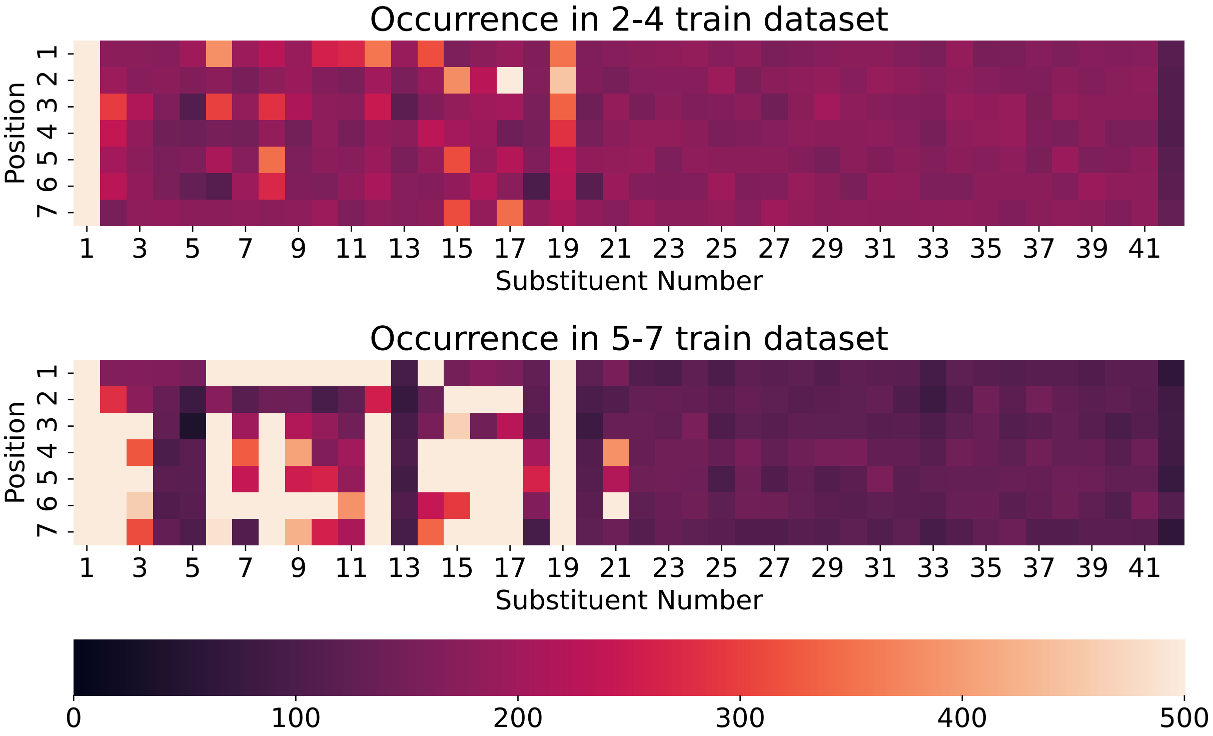


Figure S4: Illustration of how many times a ligand is found at a given position in the linear regression training data. All ligands are represented at least 40 times.

Linear regression: performance

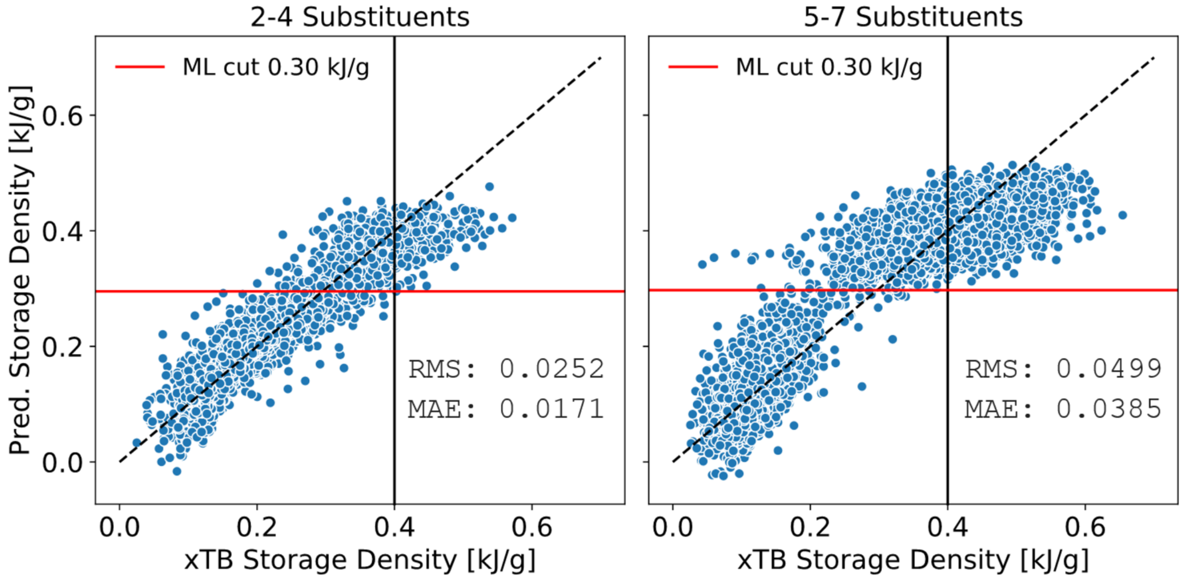


Figure S5: Illustration of the performance of the linear regression models. The actual storage density computed using GFN2-xTB with 5+5n_{rot} conformers is compared to the predicted storage density using the linear regression model for the test set. The red line corresponds to the cut-off used during the exhaustive search.

LightGBM: distribution of data

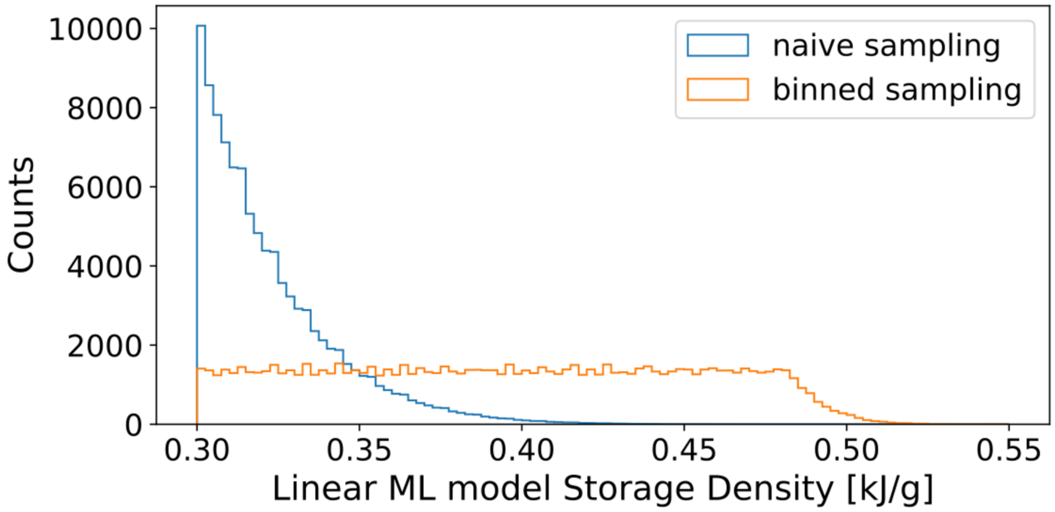


Figure S6: Illustration of binned sampling used to construct the LightGBM training, validation, and test sets.

Exhaustive screening: DFT storage density vs. barrier

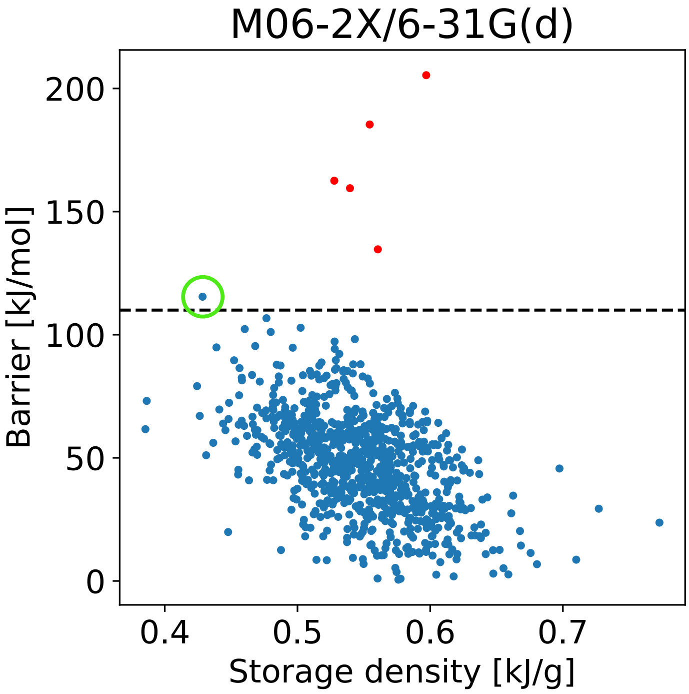


Figure S7: The calculated M06-2X/6-31G(d) storage density and back reaction barrier heights for the 954 molecules included in the first DFT investigation. Molecule **9** is marked with a green circle, the only with back reaction barrier above the cut-off illustrated by the black dashed line (110 kJ/mol). 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.

Linear Regression: Barrier Performance

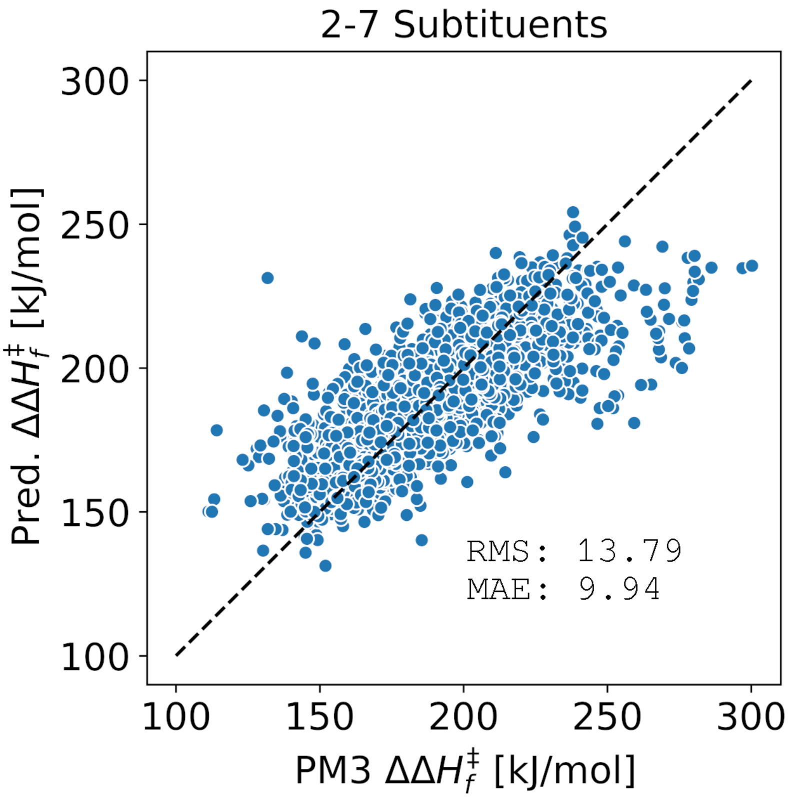


Figure S8: Comparison of the Linear Model ability to predict PM3 heat of formation to the actual PM3 heat of formation for the test set.

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 S9 compares 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 S9(a)). The "5 + 5n_{rot}-approach in Figure S9(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.

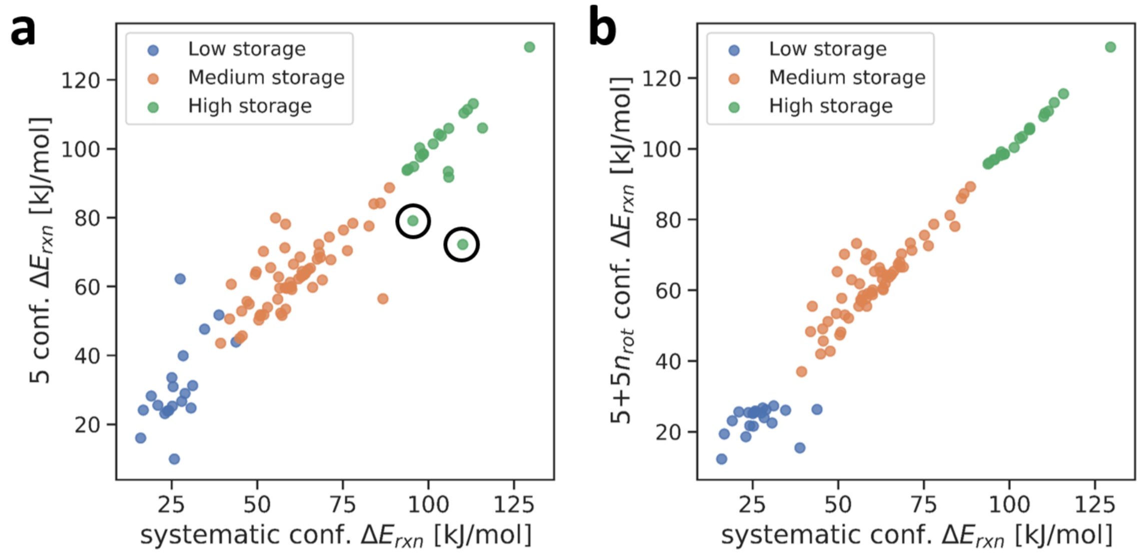


Figure S9: A plot comparing the electronic energy difference between DHA and VHF found using a systematic conformational search and a) 5 or b) 5+5n_{rot} randomly generated conformers. The energies are computed using the GFN2-xTB semi-empirical method. The two circled points indicate two molecules with high storage energy that are misclassified.

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 S10 illustrates the DFT energies compared to the corresponding semiempirical energies for (a) the storage energy and (b) the back reaction barrier.

Figure S10(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 S10(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.

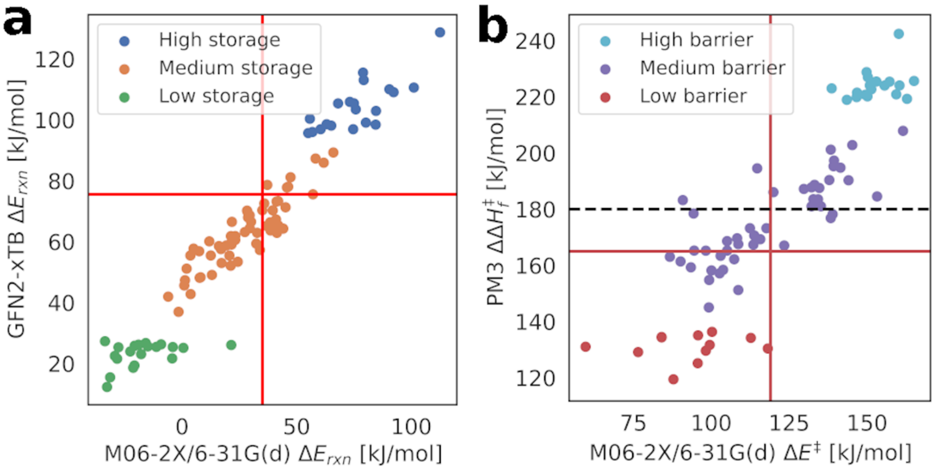


Figure S10: Comparison of GFN2-xTB (a) and PM3 (b) ability to predict the storage energy and barrier compared to DFT.

Benchmarking barrier hight using a barrier scan

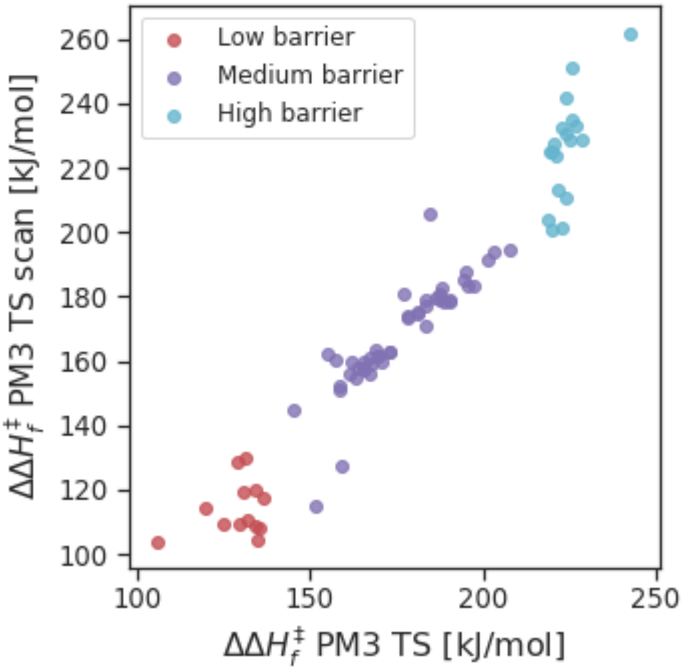


Figure S11: A plot comparing the barrier height computed using the real transition state (PM3 TS) and the energy found during the scan along the breaking bond (PM3 TS scan).

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.

SI: 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.

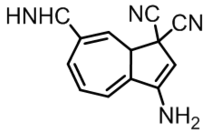
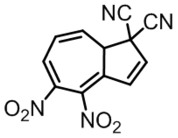
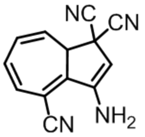
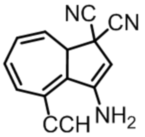
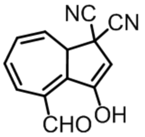
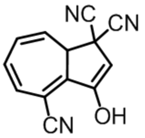
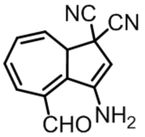
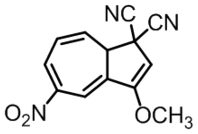
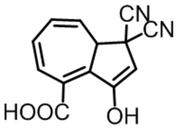
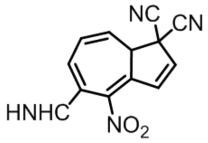
Identifier	Structure	$\Delta H^{\circ}_{rxn}/MW$ [kJ/g]	$\Delta G^{\circ,\ddagger}$ [kJ/mol]
1		0.25	112.8
2		0.25	135.6
3		0.25	123.8
4		0.24	124.2
5		0.24	129.9
S6		0.22	123.8
S7		0.22	131.6
S8		0.21	131.7
S9		0.20	138.8
S10		0.17	146.3

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