Supporting Information: Data-Driven Exploration of Synthesizable Strained Hydrocarbons as High-Energy-Density Candidates for Sustainable Aviation Fuels

Sandip Giri,[†] Subhas Ghosal,[‡] and Anakuthil Anoop[*,¶,†]

†*Department of Chemistry, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India*

‡*Department of Chemistry, National Institute of Technology Durgapur, Durgapur 713209, West Bengal, India*

¶*School of Digital Sciences, Kerala University of Digital Science, Innovation, and Technology (Digital University Kerala), Technopark Phase IV, Pallipuram, Thiruvananthapuram, Kerala – 695317 INDIA*

E-mail: anoop.a@duk.ac.in

1 AIQM2 Thermochemical Protocol

All AIQM2 calculations were performed using the MLAtom software package[S1] interfaced with Gaussian 16.[S2] Geometry optimizations were conducted without symmetry constraints. Harmonic vibrational frequency analyses were carried out at the same level of theory to confirm the absence of imaginary frequencies and to obtain zero-point energy (ZPE) and thermal corrections.

The total AIQM2 energy is expressed as:

$$E_{\mathrm{AIQM2}} = E_{\mathrm{GFN2\text{-}xTB^*}} + E_{\mathrm{NN}} + E_{\mathrm{D4}(\omega\mathrm{B97X})},$$

where:

- $E_{\mathrm{GFN2\text{-}xTB^*}}$ is the semiempirical GFN2-xTB baseline energy without internal D4 dispersion,
- E_{NN} is the neural network correction term trained against CCSD(T)*/CBS reference data,
- $E_{\mathrm{D4}(\omega\mathrm{B97X})}$ is the explicit dispersion correction.

Standard thermochemical corrections were applied to obtain enthalpies at 298.15 K. Heat of combustion values were computed using balanced stoichiometric combustion reactions:

$$\mathrm{C}_x\mathrm{H}_y(g) + \left(x + \frac{y}{4}\right)\mathrm{O}_2(g) \rightarrow x\mathrm{CO}_2(g) + \frac{y}{2}\mathrm{H}_2\mathrm{O}(l) \tag{1}$$

The enthalpy of combustion is

$$\Delta H_{\mathrm{comb}} = \left[x \cdot H_{\mathrm{CO_2}} + \frac{y}{2} \cdot H_{\mathrm{H_2O}}\right] - \left[H_{\mathrm{fuel}} + \left(x + \frac{y}{4}\right) \cdot H_{\mathrm{O_2}}\right] \tag{2}$$

Where $H_{\mathrm{H_2O}}$ includes the water enthalpy corrections for the transition from the liquid to the gas phase. The enthalpy correction value (43.99 kJ/mol per water molecule) was obtained from standard experimental literature data. Molecules that exhibit a combination of high density, elevated net heat of combustion (NHOC), and a reduced melting point are identified as promising high-energy-density hydrocarbon (HEDH) fuel candidates.

2 Molecular Descriptor

Selected topological and shape descriptors were computed using RDKit.[S3]

2.1 Third-Order Kappa Shape Index (κ_3)

For molecules with an odd number of heavy atoms:

$$^3\kappa_\alpha = \frac{(A+\alpha-1)(A+\alpha-3)^2}{(P_3+\alpha)^2}$$

For molecules with an even number of heavy atoms:

$$^3\kappa_\alpha = \frac{(A+\alpha-3)(A+\alpha-2)^2}{(P_3+\alpha)^2}$$

where:

- A = number of heavy atoms
- P_3 = number of paths of length 3
- α = Hall–Kier correction factor

The Hall–Kier correction is defined as:

$$\alpha = \sum_{i=1}^{A} \left(\frac{r_i}{r_{C_{sp^3}}} - 1 \right)$$

where r_i is the covalent radius of atom i.

2.2 Balaban J Index

$$J = \frac{M}{\mu+1} \sum_{(i,j)\in E} (D_i D_j)^{-0.5}$$

where:

- M = number of bonds
- $\mu = M - A + 1$ (cyclomatic number)
- D_i = distance sum for atom i

2.3 Bertz Complexity Index (C_T)

$$C_T = C(\eta) + C(E)$$

Bond complexity:

$$C(\eta) = 2\eta \log_2 \eta - \sum_i \eta_i \log_2 \eta_i$$

Where η is the total number of connections and η_i is the number of equivalent connections of type i.

Along with these, we have also listed other descriptors in Table S1.

Table S1: List of all RDKit molecular descriptors used for predictions

Descriptor	**Description**
HeavyAtomMolWt	Molecular weight contributed only by heavy (non-hydrogen) atoms; reflects molecular size without hydrogen count bias.
HeavyAtomCount	Total number of heavy atoms in the molecule; a direct measure of molecular skeleton size.
NumValenceElectrons	Total number of valence electrons; correlates with molecular composition and electronic structure.
MolWt	Exact molecular weight including all atoms; strongly correlated with bulk properties such as density.
LabuteASA	Approximate solvent-accessible surface area based on atomic fragments; related to molecular packing and intermolecular interactions.
HallKierAlpha	Shape and branching descriptor capturing molecular flexibility and compactness.
BertzCT	Topological complexity index reflecting molecular branching and cyclicity.
Chi1	First-order molecular connectivity index; encodes atom connectivity and bonding patterns.
Chi1v	Valence-modified first-order connectivity index incorporating valence electron information.
MinEStateIndex	Minimum electrotopological state index; reflects local electronic environments within the molecule.
MolMR	Molecular refractivity derived from atomic contributions; related to molecular volume and polarizability.
SMR_VSA5	Fragment-based van der Waals surface area within a specific refractivity range; encodes size-polarizability coupling.
TPSA	Topological polar surface area; measures surface area contributed by polar atoms and heteroatoms.
VSA_EState8	Van der Waals surface area weighted by electrotopological states in a specific range.
fr_benzene	Binary/count descriptor indicating the presence of benzene rings.
NumAromaticCarbocycles	Number of aromatic carbocyclic rings; captures aromaticity-driven packing effects.
BalabanJ	Distance-based topological index sensitive to cyclicity and molecular branching.
Kappa3	Third-order molecular shape index; reflects overall molecular geometry and flexibility.

3 Data Availability

All scripts for molecular generation, data preprocessing, model training, and analysis were implemented in Python (version 3.10). The datasets generated during this study are available from Figshare via a private reviewer link: https://figshare.com/s/10fba241de0a1c9baf16 Complete datasets and trained model checkpoints will be available upon acceptance of the manuscript.

4 Workflow

In Fig. S1, we present a schematic pipeline for creating the ML model.

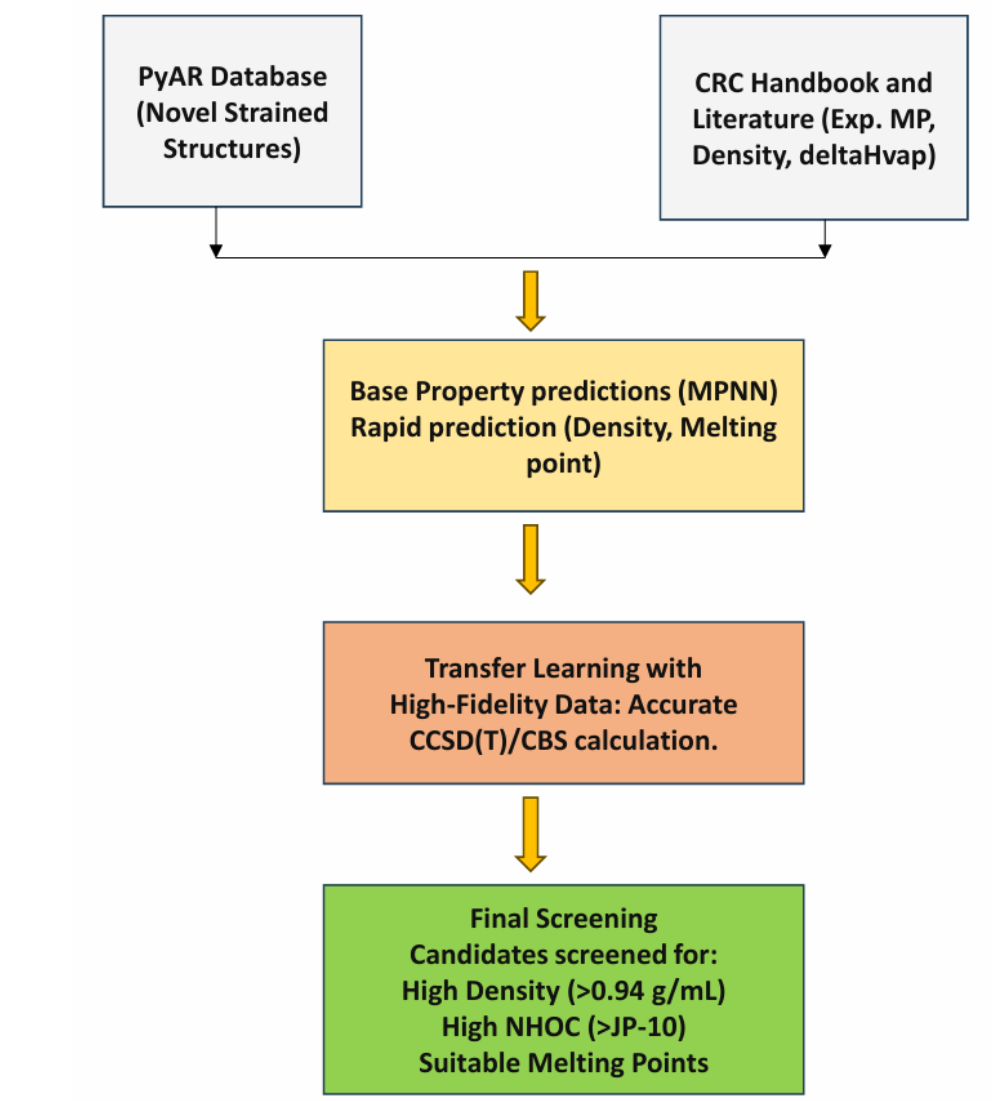


Figure S1: Schematic workflow outlining the computational procedure for ML model preparations and subsequent screening.

5 D-MPNN Training

In Fig. S2, the Directed Message Passing Neural Network (D-MPNN) operates by iteratively updating the atom and bond messages with some selected RDKit descriptors as additional features to the message passing protocol. Finally, a readout function aggregates these features to predict the target physicochemical properties, capturing the complex local chemical environments that are essential for accurate fuel characterization.

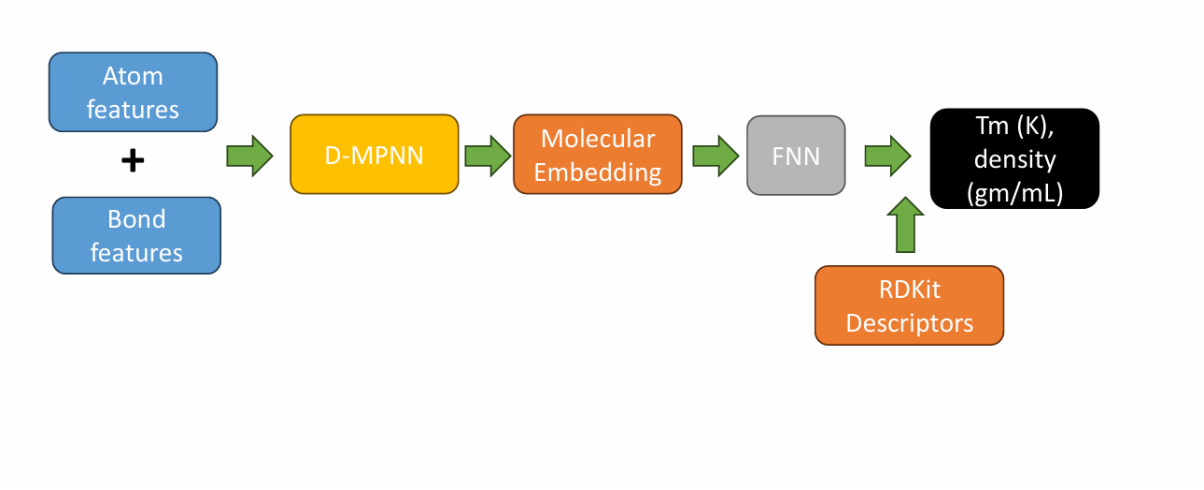


Figure S2: Schematic representation of the ChemProp architecture utilized for property prediction with additional descriptors

5.1 Heat of Combinations (HOC)

We have used the D-MPNN framework on the fine-tuned dataset of Heat of combustion. The model was trained on 10,686 data points. Here, we have utilized 5 5-fold cross-validation to prevent overfitting. Hyperparameter optimization yielded a message hidden dimensionality of 300, a depth of 10, and a 2-layer feed-forward network (FFN) with a hidden dimensionality of 1000 using ReLU activation. The learning rate was set to 1.1×10^{-5} at the initial stage, 1.1×10^{-4} at the maximum, and 1.1×10^{-6} at the final stage. Along with the learned embedding, we also used the following RDKit descriptors: BalabanJ, Kappa3, BertzCT, Chi1, Chi1v, HallKierAlpha, HeavyAtomMolWt, LabuteASA, MinEStateIndex, MolMR, MolWt, SMR_VSA5, TPSA, VSA_EState8, fr_benzene, and NumAromaticCarbocycles.

The final fine-tuned model contains approximately 1.5 million trainable parameters (comprising 227 K in the message-passing module and 1.3 M in the feed-forward network). To prevent the model from memorizing structural patterns within the 658-point experimental dataset, we employed a dropout rate of 0.2, limited training to 100 epochs, and utilized 5-fold cross-validation.

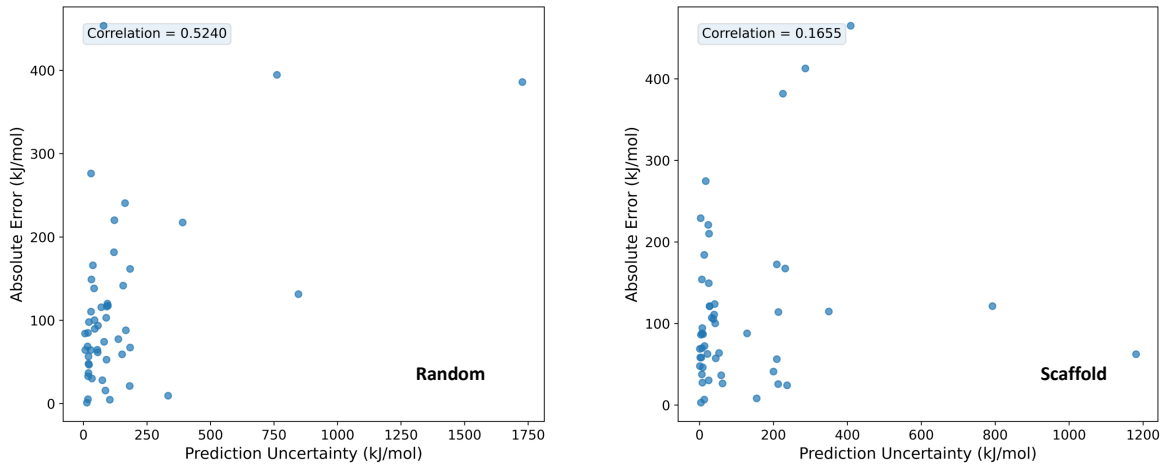


Figure S3: Uncertainty quantification for the heat of combustion model on test set (a) Random and (b) Scaffold split.

5.2 Melting point

Here, we have used a similar D-MPNN-based framework for melting point prediction. The model was trained on a comprehensive dataset totaling 7,783 data points, comprising 605 points from dataset A and 7,178 points from dataset B. We utilized a random split and 5-fold cross-validation over 100 epochs. Hyperparameter optimization yielded a message hidden dimensionality of 300, a depth of 10, and a 2-layer feed-forward network (FFN) with a hidden dimensionality of 1000 using ReLU activation. The learning rate was set to 1.1×10^{-5} at the initial stage, 1.1×10^{-4} at the maximum, and 1.1×10^{-6} at the final stage. Also, we have used 2-D normalized RDKit descriptors to improve the metrics. This D-MPNN configuration achieved an R^2 of 0.769, an MAE of 20.988 K, and an RMSE of

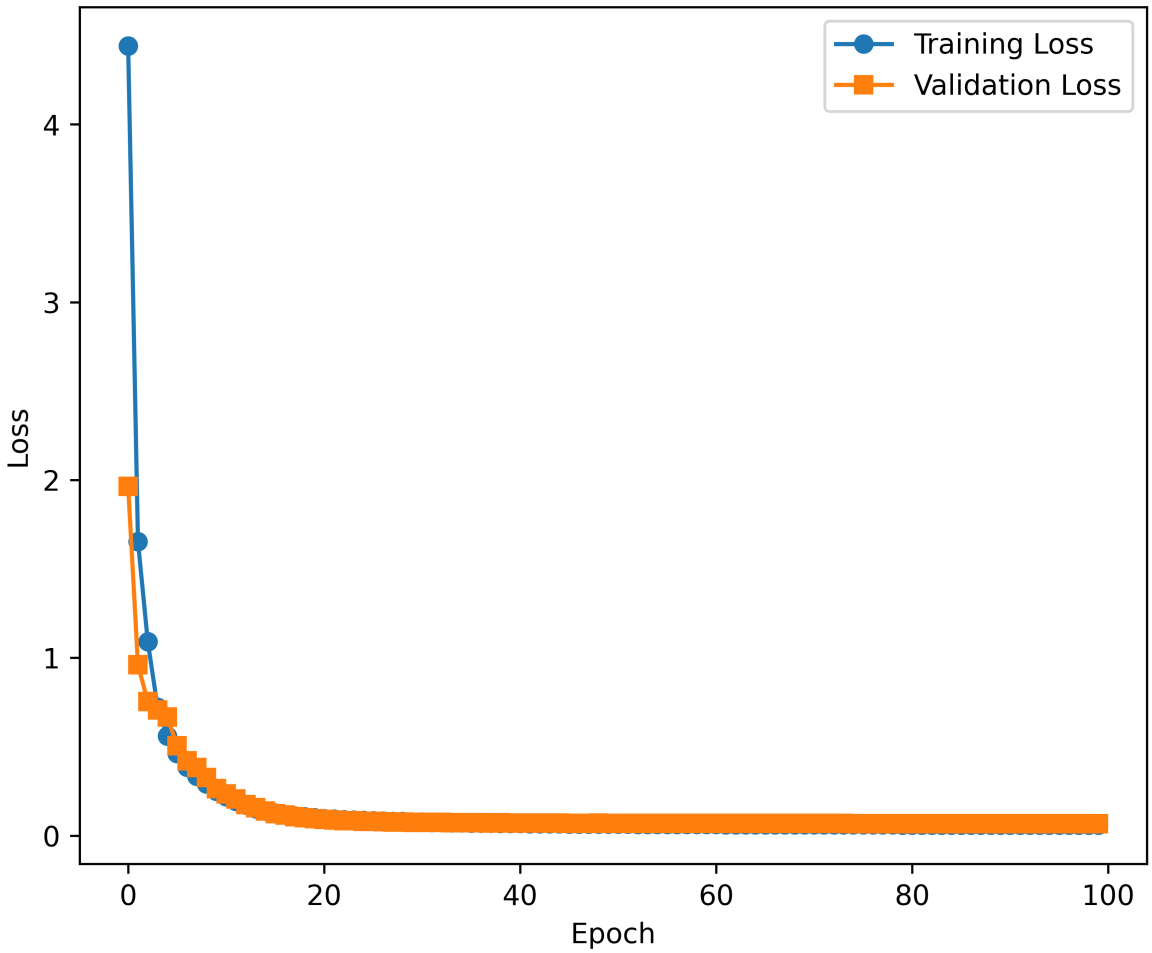


Figure S4: Training and validation loss curves for the fine-tuning of the heat of combustion process using a random data split.

35.797 K. The parity between the predicted and actual values is presented in Figure S5.

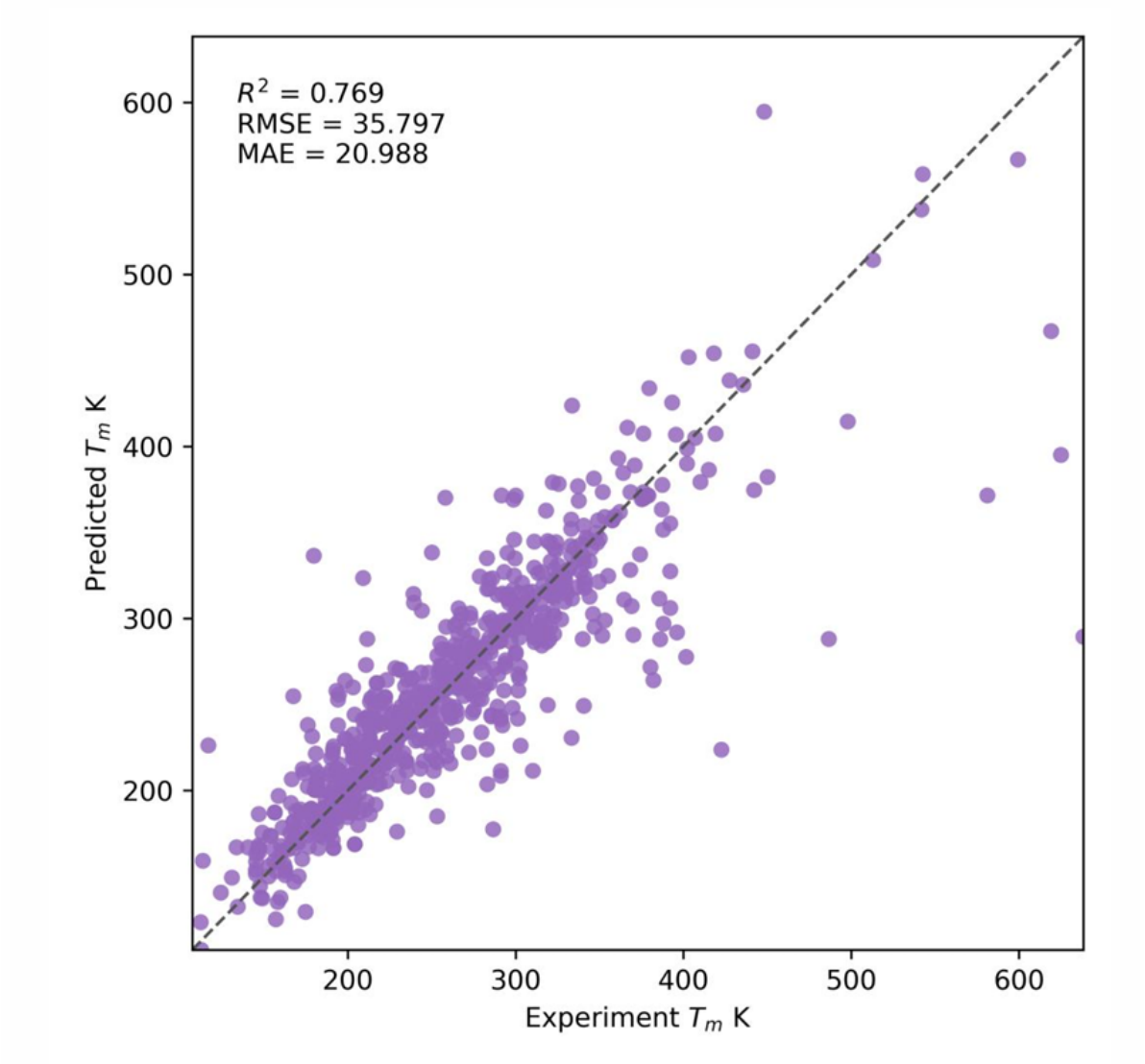


Figure S5: Parity plot evaluating the accuracy of ChemProp-predicted melting points against experimental data.

5.3 Density

To facilitate high-throughput screening, liquid density predictions were obtained using a support vector machine regression (SVM) model.[S4] Given the limited size of the experimental dataset (650 data points), the SVM approach demonstrated superior predictive performance compared to ChemProp. We have used the following RDKit descriptors: HeavyAtomMolWt, HeavyAtomCount, NumValenceElectrons, MolWt, LabuteASA, HallKierAlpha, BertzCT, Chi1, Chi1v, MinEStateIndex, MolMR, SMR_VSA5, TPSA, VSA_EState8,

fr_benzene, CalcNumAromaticCarbocycles, BalabanJ, and Kappa3. The model was trained using 5-fold cross-validation, and hyperparameters were systematically tuned across a defined grid space ($C \in \{5, 8, 10, 11, 13, 14, 15, 20\}$ and $\epsilon \in \{0.01, 0.1, 0.2, 0.5\}$). Optimal predictive performance was achieved with a linear kernel, $C = 8$, and $\epsilon = 0.01$, yielding an R^2 of 0.962, a mean absolute error (MAE) of 0.016 g/mL, and a root-mean-square error (RMSE) of 0.020 g/mL. The parity between the experimental values and the predicted liquid densities derived from this optimized SVM model is illustrated in Figure S6.

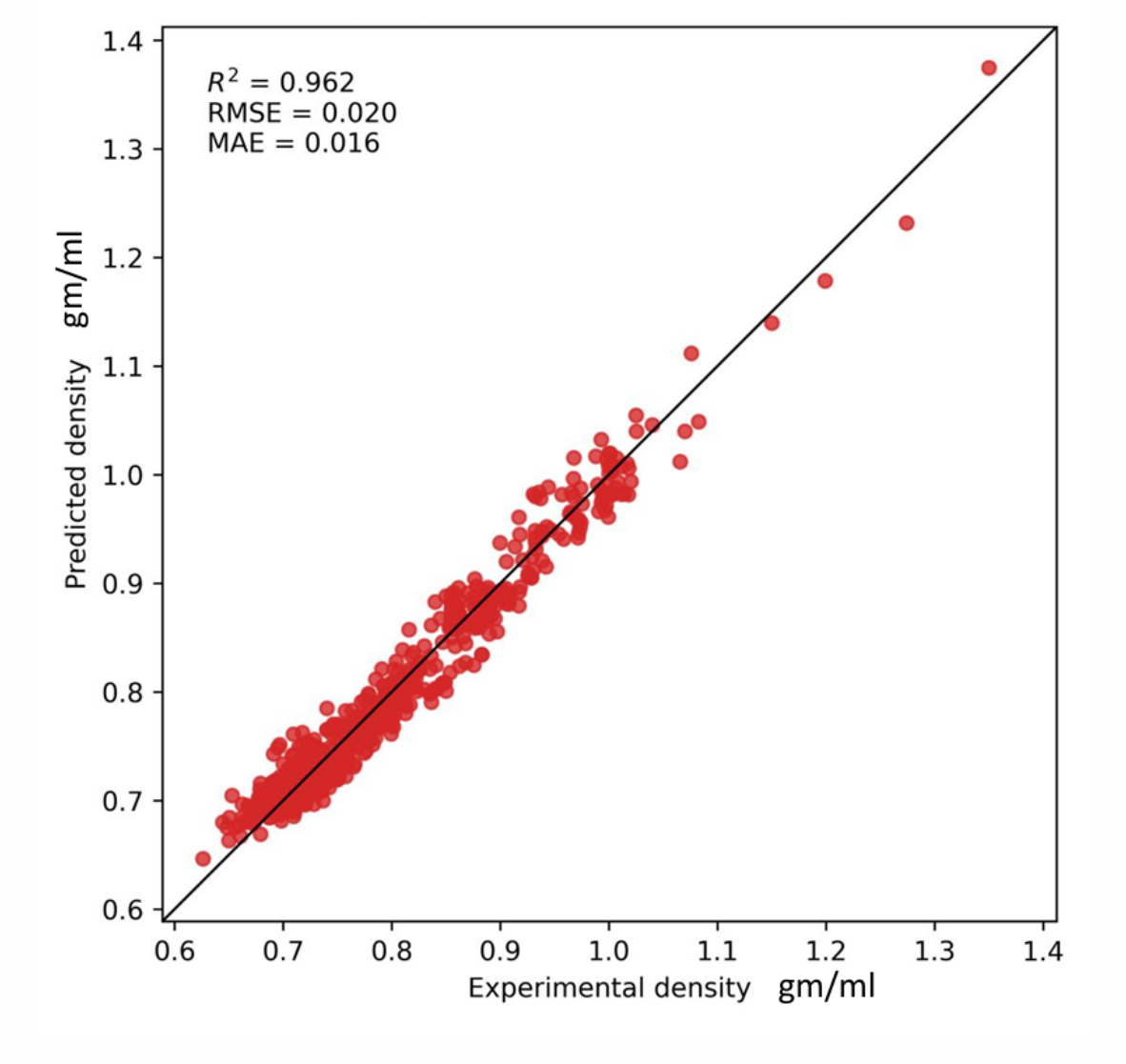


Figure S6: Parity plot demonstrating the correlation between experimental liquid densities and values predicted by the optimized Support Vector Machine (SVM) regression model.

6 2D-Fuel structures

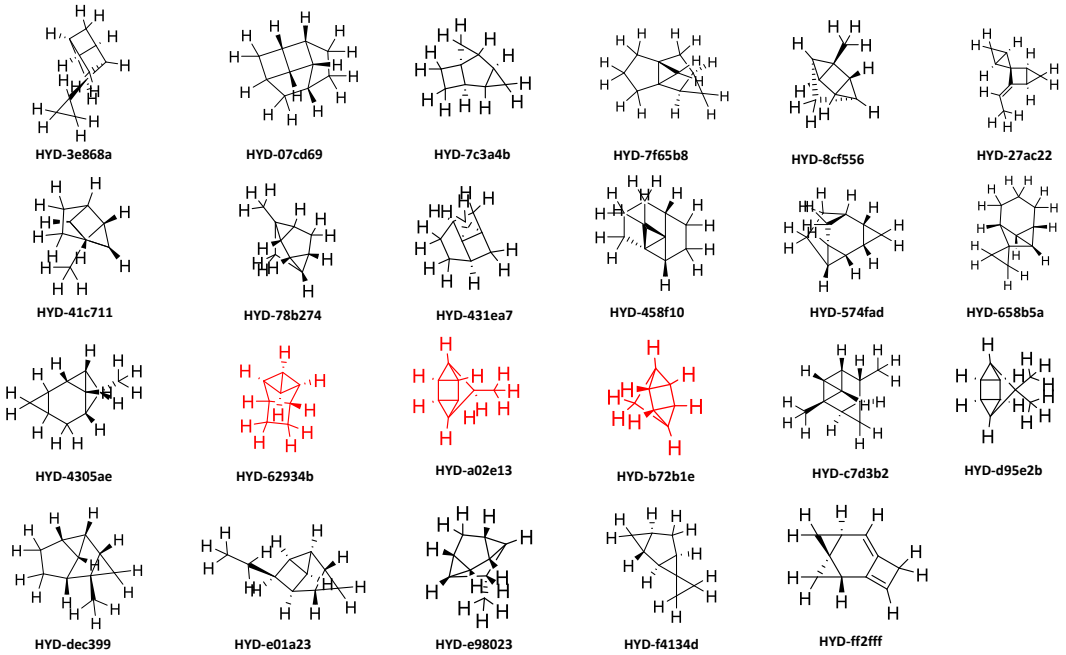


Figure S7: 2D ChemDraw illustrations of the 23 optimal high-energy-density hydrocarbon (HEDH) fuel candidates identified through the screening protocol.

7 3D-Fuel structures

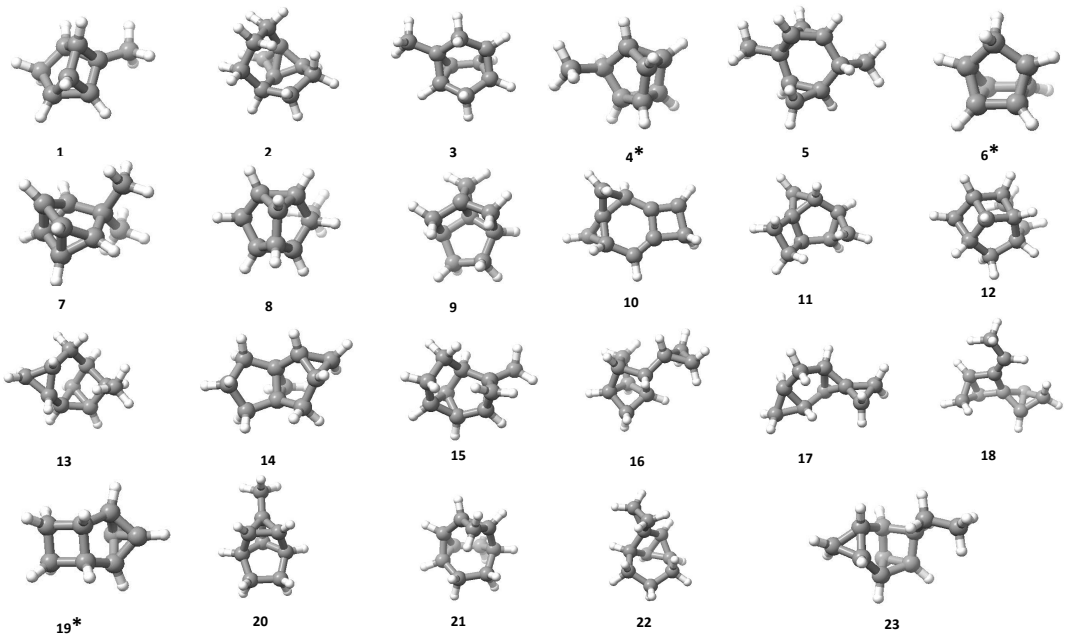


Figure S8: 3D molecular visualizations of those 23 optimal high-energy-density hydrocarbon (HEDH) fuel candidates.

8 Retrosynthesis pathway

HYD-3e868a

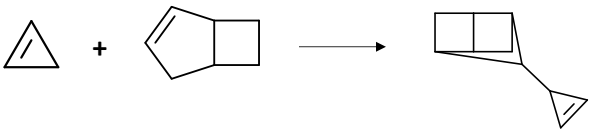
Step 1

Step 2

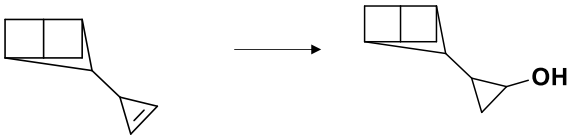
Step 3

Step 4

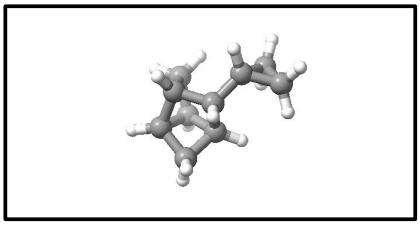
Step 5



Step 6

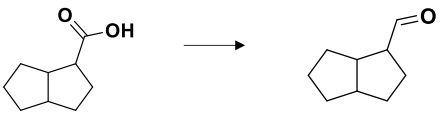


Step 7

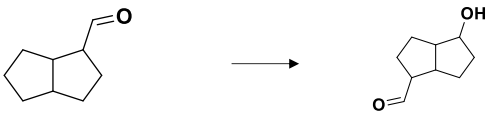


HYD-07cd69

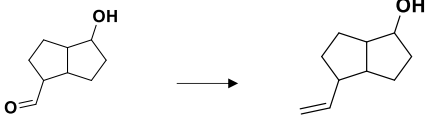
Step 1



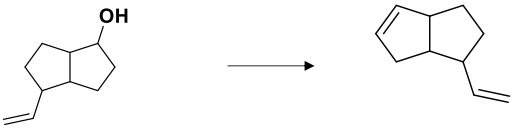
Step 2



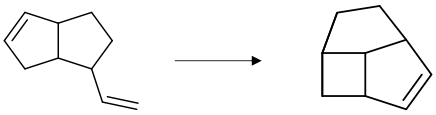
Step 3



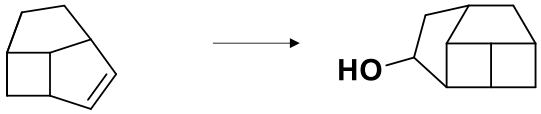
Step 4



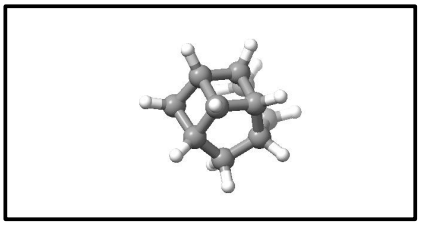
Step 5



Step 6

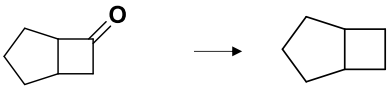


Step 7



HYD-7c3a4b

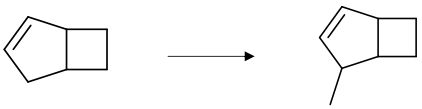
Step 1



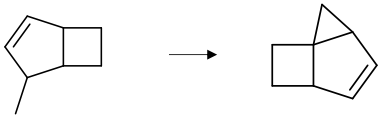
Step 2

Step 3

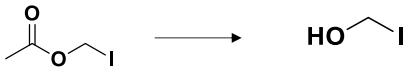
Step 4



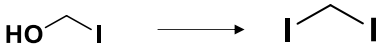
Step 5



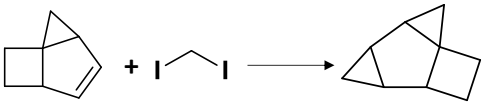
Step 6

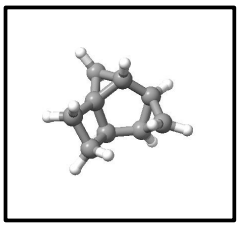


Step 7



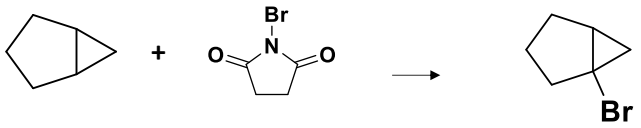
Step 8



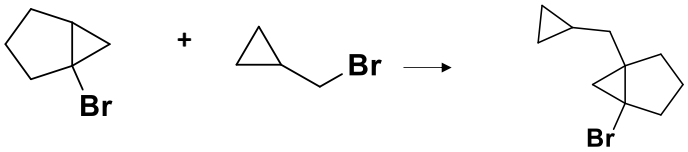


HYD-7f65b8

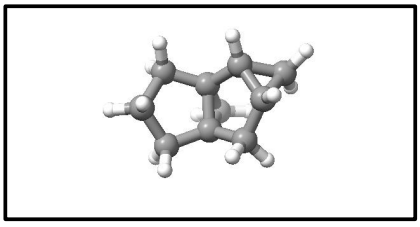
Step 1



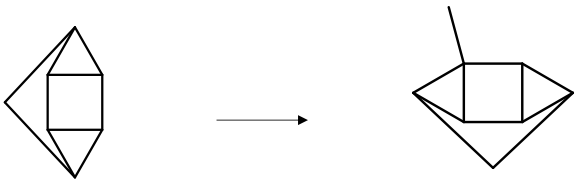
Step 2

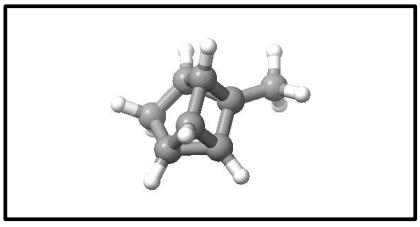


Step 3



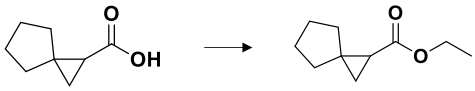
HYD-8cf556





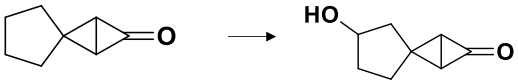
HYD-27ac22

Step 1

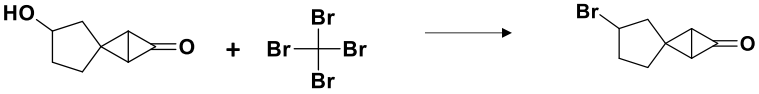


Step 2

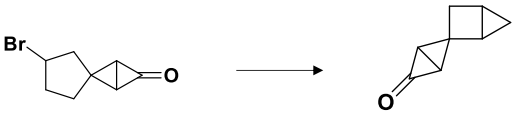
Step 3



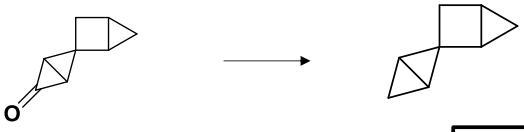
Step 4



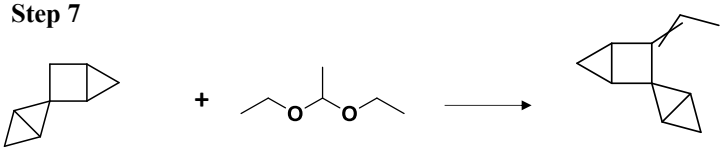
Step 5

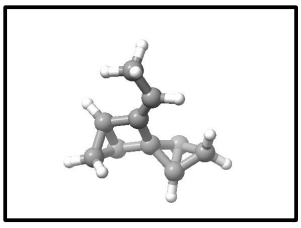


Step 6



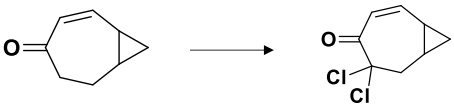
Step 7



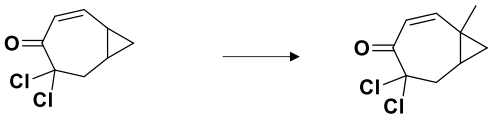


HYD-41c711

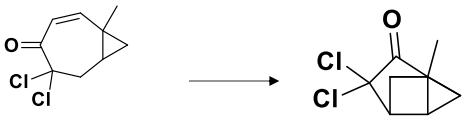
Step 1



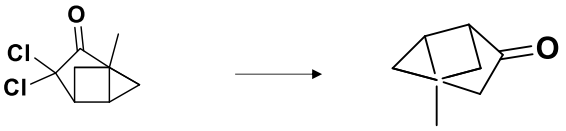
Step 2



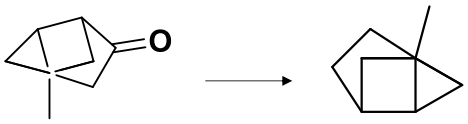
Step 3

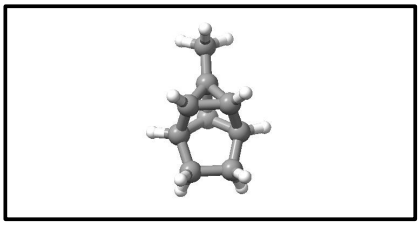


Step 4



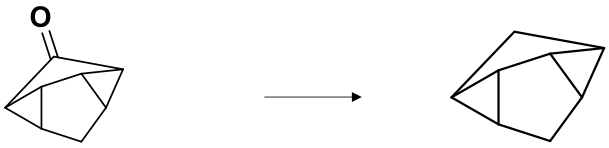
Step 5



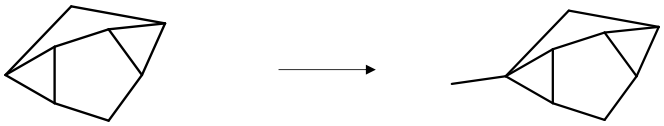


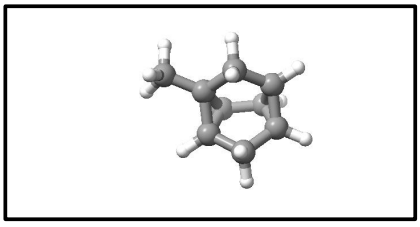
HYD-78b274

Step 1



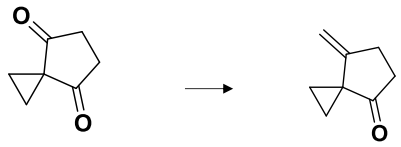
Step 2



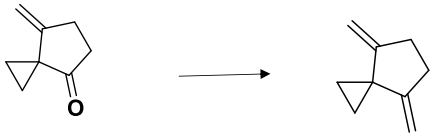


HYD-431ea7

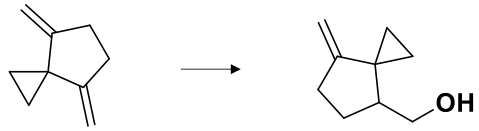
Step 1



Step 2

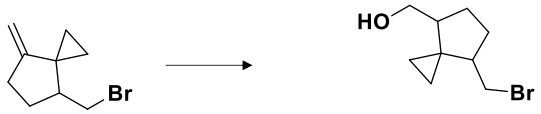


Step 3

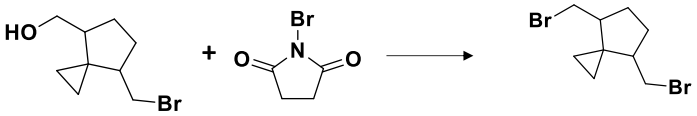


Step 4

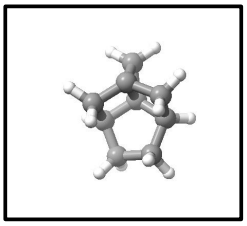
Step 5



Step 6

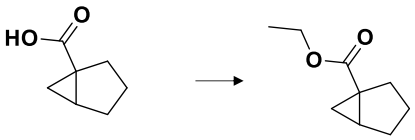


Step 7

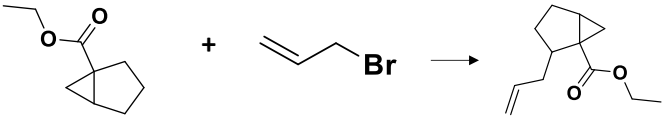


HYD-458f10

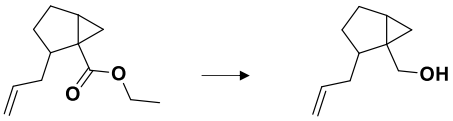
Step 1



Step 2

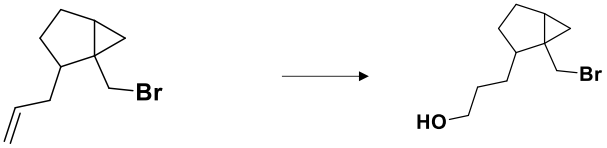


Step 3



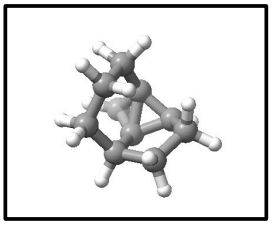
Step 4

Step 5



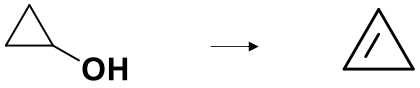
Step 6

Step 7

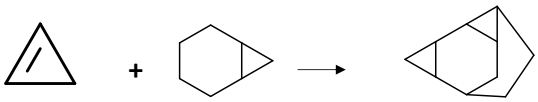


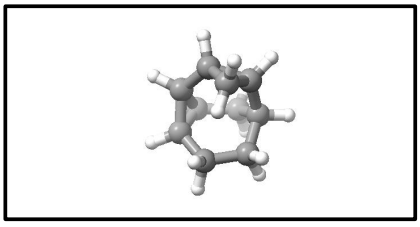
HYD-574fad

Step 1



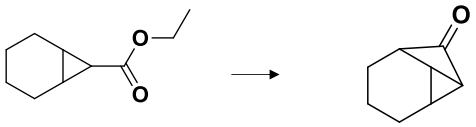
Step 2



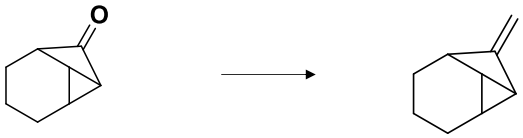


HYD-658b5a

Step 1



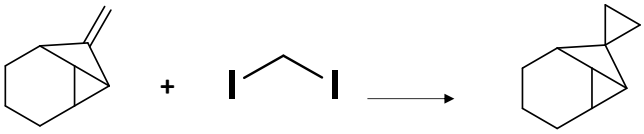
Step 2

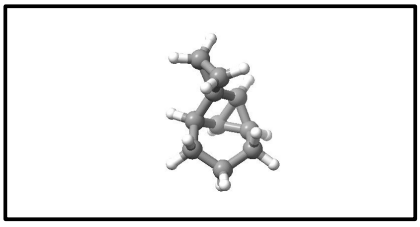


Step 3

Step 4

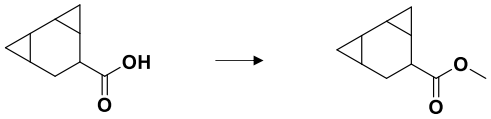
Step 5





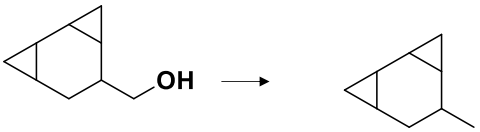
HYD-4305ae

Step 1

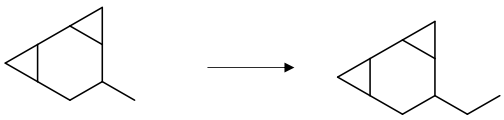


Step 2

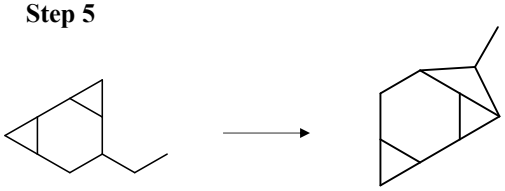
Step 3

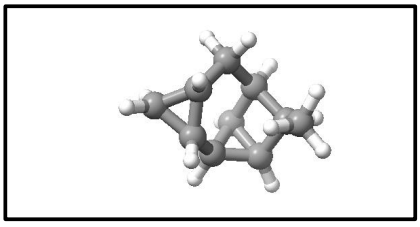


Step 4



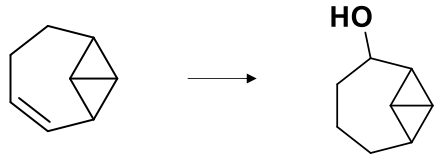
Step 5



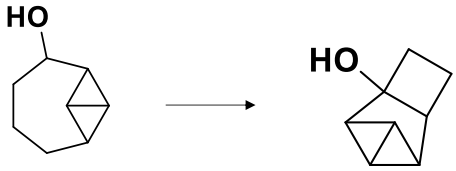


HYD-62934b

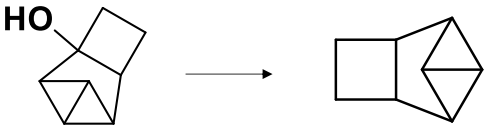
Step 1

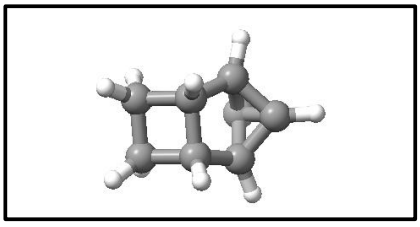


Step 2

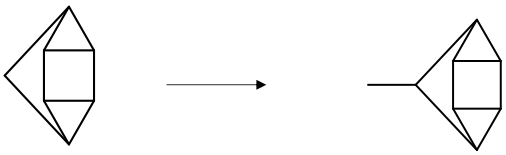


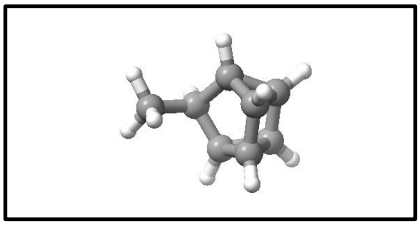
Step 3





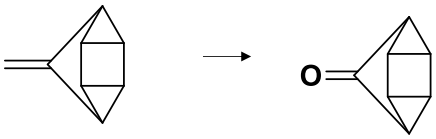
HYD-a02e13



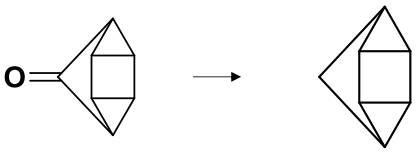


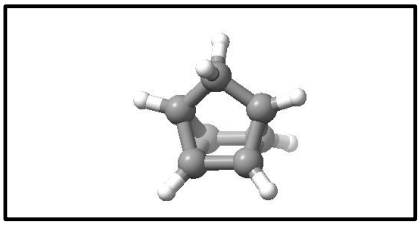
HYD-b72b1e

Step 1

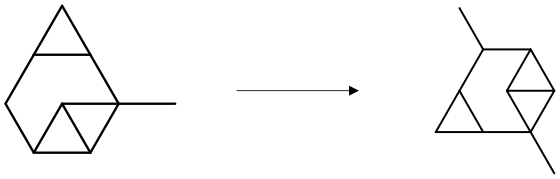


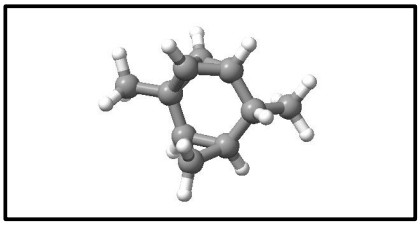
Step 2



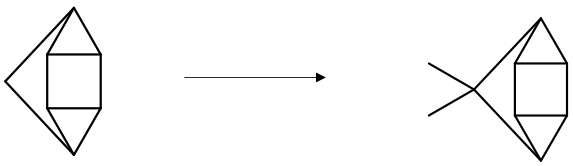


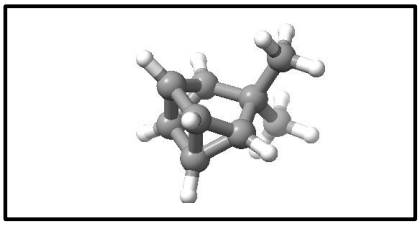
HYD-c7d3b2





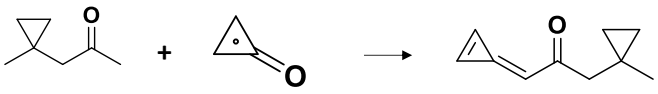
HYD-d95e2b





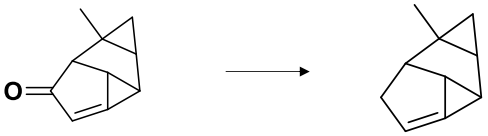
HYD-dec399

Step 1



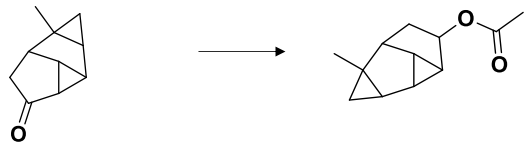
Step 2

Step 3

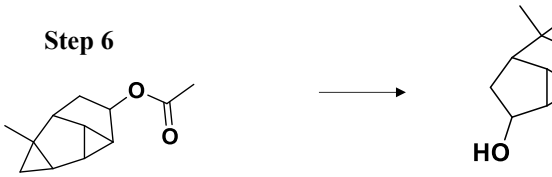


Step 4

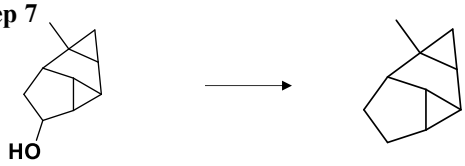
Step 5

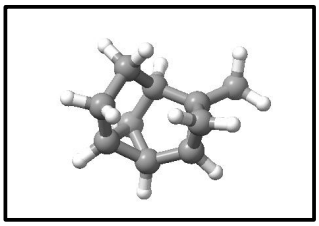


Step 6



Step 7

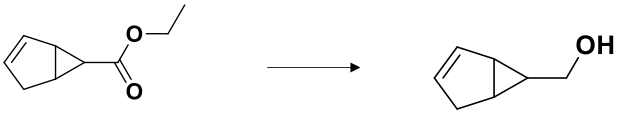




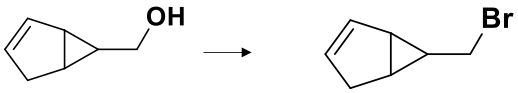
HYD-e01a23

Step 1

Step 2

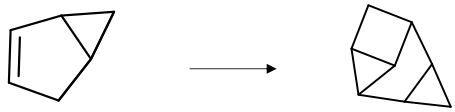


Step 3



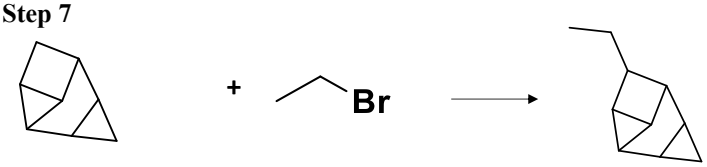
Step 4

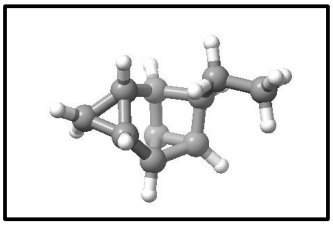
Step 5



Step 6

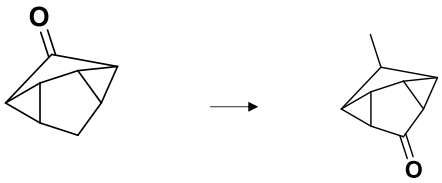
Step 7



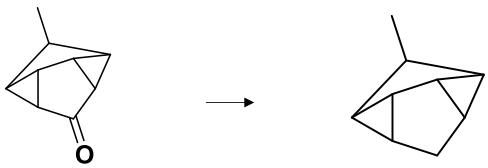


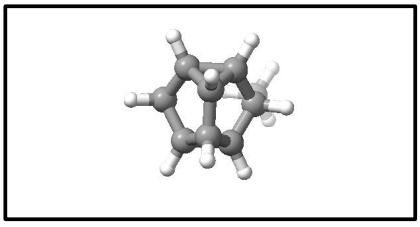
HYD-e98023

Step 1



Step 2

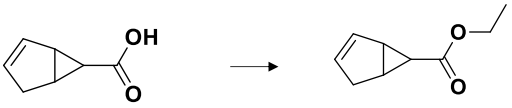




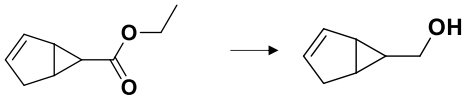
HYD-f4134d

Step 1

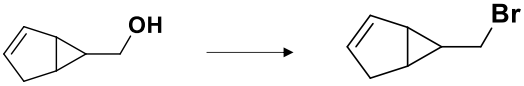
Step 2



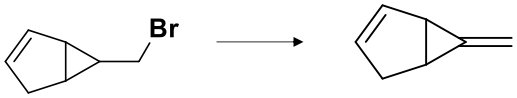
Step 3



Step 4



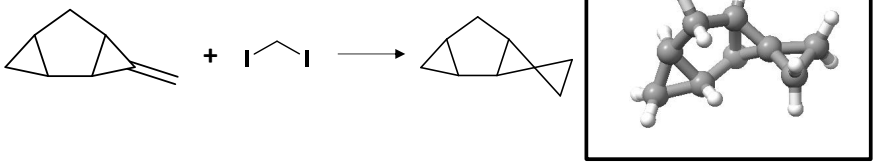
Step 5



Step 6

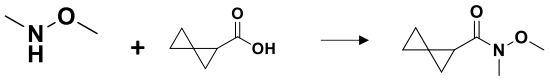
Step 7

Step 8



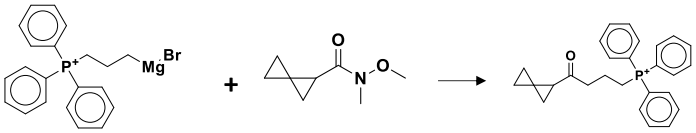
HYD-ff2fff

Step 1

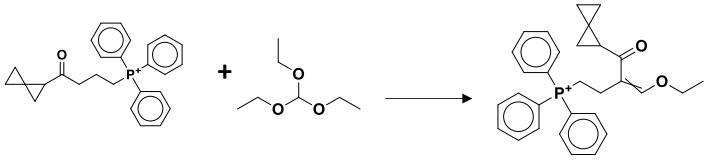


Step 2

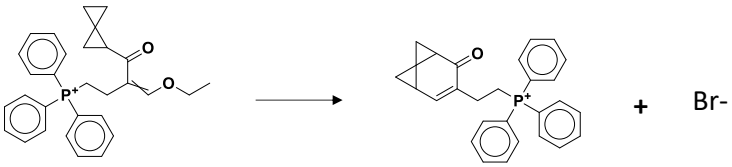
Step 3



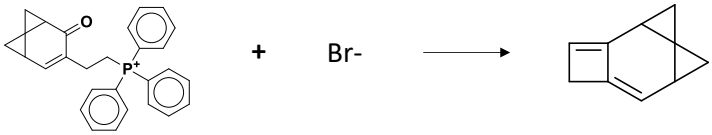
Step 4

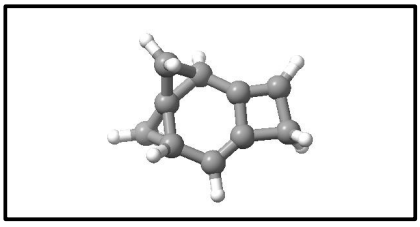


Step 5



Step 6





References

(S1) Dral, P. O. MLatom: A program package for quantum chemical research assisted by machine learning. *Journal of computational chemistry* **2019**, *40*, 2339–2347.

(S2) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(S3) Landrum, G.; Tosco, P.; Kelley, B.; Rodriguez, R.; Cosgrove, D.; Vianello, R.; Gedeck, P.; Jones, G.; Kawashima, E.; Nealschneider, D.; others rdkit/rdkit: 2025_03_1 (Q1 2025) Release. *Zenodo* **2025**,

(S4) Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.