Supplementary Information (SI) for Digital Discovery.

Supplementary information for:

Discovery of Hydrogen Storage Molecules using Large Language Models and Machine Learning

Hassan Harb,[a,] Magali S. Ferrandon,[b] Timothy A.Goetjen,[c] Seryeong Lee-,[b,c] Omar K. Farha,[c]*

*Massimiliano Delferro,[b] Rajeev Surendran Assary[a,]**

[a]Materials Science Division, Argonne National Laboratory, Lemont, IL 60439, United States

[b]Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, IL 60439, United States

[c]Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States

*Corresponding authors: HH: hharb@anl.gov; RSA, assary@anl.gov

Contents

Text S1: Overview of the G4MP2 method

G4MP2 is a high-level composite method used for calculating molecular energies with high accuracy. The method is designed to achieve near chemical accuracy by combining several different levels of theory and basis sets in a sequential approach. Below is a step-by-step outline of how G4MP2 is executed in Gaussian:

1. Initial geometry optimization and frequency calculation: The geometry of the molecule is first optimized using the B3LYP functional with a mid-sized basis set (6-31G(2df,p)). o A frequency calculation is performed at this level to confirm that the optimized geometry corresponds to a local minimum (no imaginary frequencies). The zero-point energy correction (EZPE) is obtained by scaling the B3LYP/6-31G(2df,p) frequencies by a factor of 0.9854.
2. Single Point Energy Calculations: Several single-point energy calculations are performed at the optimized geometry to refine the total energy. These calculations use increasingly sophisticated methods and basis sets to improve accuracy:
    1. MP2 Calculation: Second-order Møller-Plesset perturbation theory (MP2) is used with a larger basis set (GTMP2LargeXP) to provide a correlated electron description.
    2. CCSD(T) Calculation: Coupled-cluster with single, double, and perturbative triple excitations [CCSD(T)] is performed with a smaller basis set (GTBas1) to account for higher-order electron correlation effects.
3. High-Level Hartree-Fock Calculations: Additional Hartree-Fock calculations are conducted using larger basis sets (GFHFB3 and GFHFB4) to refine the Hartree-Fock energy component further. These steps help to improve the convergence and accuracy of the calculation.
4. Combining Results for Final Energy Estimation: The results from the MP2, CCSD(T), and Hartree-Fock calculations are combined according to the G4MP2 protocol to provide a final, high-accuracy estimate of the molecular energy. This combination effectively balances the computational cost with the need for high precision.

Label	SMILES	ΔH_{exp} kJ/mol H_2	Ref
1	N1N=CC=N1	22.3	[2]
2	CN1C=CN=C1	39.1	[2]
3	N1C=CN=C1	39.8	[2]
4	C1=C3C(=CC2=C1[N](C(=C2C)C)C)C(=C([N]3C)C)C	40.7	[2]
5	C1=CC=C2C(=C1)C3=CC=CC=C3N2(CC)	50.6	[3]
6	C1=CC=C2C(=C1)C3=CC=CC=C3N2	51.1	[3]
7	CN1C=CC2=CC=CC=C21	51.9	[4]
8	CC1=CC2=CC=CC=C2N1	55.2	[5]
9	C1=CNC=C1	56.1	[2]
10	C1=CC=C2C(=C1)C=CN2	56.6	[5]
11	C1=CN=CC=C1N	56.7	[6]
12	C1=CC=CC2=C1OC3=C2C=CC=C3	56.7	[2]
13	C1=CN=CC=N1	56.9	[7]
14	C1=CN=CN=C1	60.1	[2]
15	C1=CC=C(C=C1)O	61	[2]
16	C1=CC=C2C(=C1)N=C3C=CC=CC3=N2	61.3	[8]
17	C1=CC=C2C(=C1)C=CC=N2	61.9	[9]
18	C1=CC=NC=C1	62.3	[2]
19	CC1=CC=CC=C1CC2=CC=CC=C2	63.5	[9]
20	C1=CC=C(C=C1)N	64.0	[10,11]
21	CC1=CC2=CC=CC=C2C=C1	65.3	[9]
22	CC1=C(C=CC=C1CC2=CC=CC=C2)CC3=CC=CC=C3	65.4	[9]
23	C1=CC=C(C=C1)C2=CC=CC=C2	66.6	[2]

SMILES	ΔH (DFT) kJ/mol H_2
Cc1cc2ccccc2o1	69.6
c1ccc2nocc2c1	68.2
Cc1cnc2ccccc2c1	64.1
c1ccc2nc(C3CC3)ccc2c1	61.6
Cc1ccccc1Cc1ccncc1	68.5
Cc1ccccc1-c1cnccc1C	65.4
ON1C2=CC=CC=C2C2=C1C=CC=C2	61.6

Table S1: SMILES strings and the enthalpy of hydrogenation of the LOHC-7 seed set.[1]

24	C1=CC=C2C=CC=CC2=C1	66.7	12
25	CC1=CC=CC=C1	68.3	12
26	C1=CC=CC=C1	68.6	12
27	CC1=CC(=CC=C1)CC2=CC=CC=C2	71	9
28	C1C=CC=C1	107.5	2
29	C1CC=CC=C1	113.8	11
30	C1C=CCC=C1	113.8	11
31	C1CCC=CC1	118.83	11

Table S2: SMILES and experimental enthalpies of hydrogenation of the Expt-31 seed set

Table S3: Comparison of the G4MP2 calculated ΔH of the set of down selected molecules (Argo-1 to Argo-42 structures are shown in Figure 5) with the machine learning predicted values, alongside the absolute errors. Values in red show the molecules that were incorrectly identified as good LOHCs by our workflow. All values are given in units of kJ/mol H_2.

ID	ΔH G4MP2 (kJ/mol H_2)	ΔH ML (kJ/mol H_2)	Absolute Error (kJ/mol H_2)
Argo-01	51.79	50.28	1.50
Argo-02	57.44	61.09	3.66
Argo-03	55.11	60.76	5.65
Argo-04	58.62	59.05	0.43
Argo-05	56.39	50.52	5.87
Argo-06	54.16	59.42	5.26
Argo-07	57.06	52.79	4.27
Argo-08	57.76	60.48	2.72
Argo-09	63.41	59.87	3.54
Argo-10	57.97	59.59	1.61
Argo-11	56.70	59.84	3.14
Argo-12	52.97	60.10	7.13
Argo-13	58.39	55.77	2.62
Argo-14	63.60	60.75	2.85
Argo-15	67.49	60.50	6.99
Argo-16	69.25	64.91	4.35
Argo-17	54.23	61.19	6.96
Argo-18	61.06	64.47	3.41
Argo-19	52.81	62.41	9.60
Argo-20	56.08	63.94	7.86
Argo-21	71.78	59.83	11.95
Argo-22	55.84	59.97	4.14
Argo-23	60.64	61.36	0.72
Argo-24	51.95	62.21	10.27

Argo-25	50.22	49.91	0.32
Argo-26	55.12	51.10	4.02
Argo-27	54.00	61.81	7.81
Argo-28	55.81	50.06	5.75
Argo-29	51.19	60.20	9.01
Argo-30	55.57	48.10	7.47
Argo-31	53.99	50.56	3.43
Argo-32	54.19	61.18	6.99
Argo-33	57.88	53.54	4.34
Argo-34	51.57	60.66	9.09
Argo-35	57.12	60.48	3.36
Argo-36	73.47	60.46	13.02
Argo-37	59.85	59.75	0.10
Argo-38	75.48	63.29	12.19
Argo-39	67.92	59.38	8.54
Argo-40	60.77	58.37	2.40
Argo-41	56.19	61.33	5.14
Argo-42	57.22	61.24	4.02

Table S4: Detailed physicochemical properties of LOHCs, Argo molecules (1 to 42, 2D structures are shown in Figure 5) discovered in this study. Abbreviations used are MP for melting point (°C), BP for boiling point (°C), HL for hydrogen-lean state, HR for hydrogen-rich state, SA for synthetic accessibility, and PC for availability in PubChem database. Note that the MP and BP are predicted using the OPERA[159] model, SA scores are calculated using the method by Ertl and Schuffenhauer, and ΔH values are ML-predicted based on a random forest model trained on G4MP2 data. The "n/a" denotes instances where the models failed to predict MP/BP. Color scales highlight relative values within each column to visually emphasize trends, outliers, and property variation across the dataset.

LOHC Label	MP (HL)	MP (HR)	BP (HL)	BP (HR)	SA (HL)	SA (HR)	PC (HL)	PC (HR)	ΔH	% wt. H_2
1	13.9	-2.3	277.1	257	1.45	2.75	TRUE	TRUE	50.3	6.67
2	6.4	-14	300.1	291	1.14	1.84	TRUE	TRUE	61.1	5.8
3	34.5	-10.7	241.2	192	1.31	3.03	TRUE	TRUE	60.8	6.62
4	-12.6	4.6	191.2	160	1.38	1.85	FALSE	TRUE	59.1	8.9
5	-17.7	-23.6	144.3	126	1.64	2.81	TRUE	TRUE	50.5	6.1
6	18.3	24.7	201.9	189	1.1	2.47	TRUE	TRUE	59.4	6.29
7	-54.9	-24.4	129.6	121	1.56	2.8	TRUE	TRUE	52.8	6.1
8	24.9	0.1	265.1	n/a	1.03	1.86	TRUE	TRUE	60.5	6.71
9	-25.6	-42.4	179.1	183	1.44	1.77	TRUE	TRUE	59.9	7.06
10	25.1	-28.4	290.6	281	1.21	2.03	TRUE	TRUE	59.6	5.75
11	39.9	-29.8	286.2	248	1.09	1.98	TRUE	TRUE	59.8	6.16
12	-11.3	-11.1	258.3	228	1.38	3.15	TRUE	TRUE	60.1	6.06
13	-23.4	0.2	233.5	179	1.58	1.9	FALSE	TRUE	55.8	7.92
14	28.1	9.7	306.6	271	1.27	1.76	TRUE	TRUE	60.7	7.26
15	36.2	-21.8	297.7	284	1.55	2.52	TRUE	TRUE	60.5	6.77
16	-30.1	-54	145.1	132	1.55	1.63	TRUE	TRUE	64.9	7.19
17	-8.8	29.1	216.9	197	1.37	2.51	TRUE	TRUE	61.2	5.67
18	-18.3	-1.6	174.1	163	1.8	3.06	TRUE	TRUE	64.5	6.39
19	20.9	-4.8	284.9	285	1.99	3.28	TRUE	TRUE	62.4	5.8
20	17.9	-10.5	265.2	253	1.15	2.74	TRUE	TRUE	63.9	6.71
21	38.2	-11.7	326.3	n/a	1.93	1.8	TRUE	TRUE	59.8	6.82
22	-11.3	-11.1	258.3	228	1.34	2.94	TRUE	TRUE	60	6.06
23	32.2	-21.8	280.6	284	1.79	2.52	TRUE	TRUE	61.4	5.8
24	14.9	-5.8	268.1	235	1.47	3.22	TRUE	TRUE	62.2	5.59
25	12.6	2.5	291.2	269	1.45	2.62	TRUE	TRUE	49.9	6.19
26	12.8	4	291.3	269	1.45	2.16	TRUE	TRUE	51.1	6.19
27	30.3	-6.4	265.5	226	1.49	3.42	TRUE	TRUE	61.8	6.06
28	14.6	3.4	317.9	277	1.5	2.63	TRUE	TRUE	50.1	5.78
29	14.9	-5.8	268.1	235	1.43	3.07	TRUE	TRUE	60.2	5.59
30	-2.5	-1.2	289	269	1.46	2.67	TRUE	TRUE	48.1	6.19
31	24.9	2.5	243.2	202	1.54	2.96	TRUE	TRUE	50.6	7.24
32	28.9	14.7	282.5	286	1.42	2.23	TRUE	TRUE	61.2	5.8
33	-16.2	-11.1	272.7	276	1.4	2.16	TRUE	TRUE	53.5	6.22
34	6.6	18.6	69.6	89.8	3.93	3.55	TRUE	TRUE	60.7	5.52
35	27.1	-35.8	258	243	1.06	1.99	TRUE	TRUE	60.5	6.63
36	-28.3	-85.6	176.8	157	1.63	1.66	TRUE	TRUE	60.5	6.39
37	-25.2	1.9	309.1	277	1.22	1.95	TRUE	TRUE	59.8	5.78
38	-15.4	-85.5	201.3	n/a	2.01	1.66	TRUE	TRUE	63.3	7.19
39	26.7	9.1	325.1	305	1.65	1.78	TRUE	TRUE	59.4	7.25
40	37.1	-14	297.4	291	1.59	1.84	TRUE	TRUE	58.4	6.77
41	-7.2	24.8	293.5	265	1.48	2.08	TRUE	TRUE	61.3	6.19
42	37	33.8	309.8	278	1.58	2.41	TRUE	TRUE	61.2	5.78

Text S2: LLM Performance

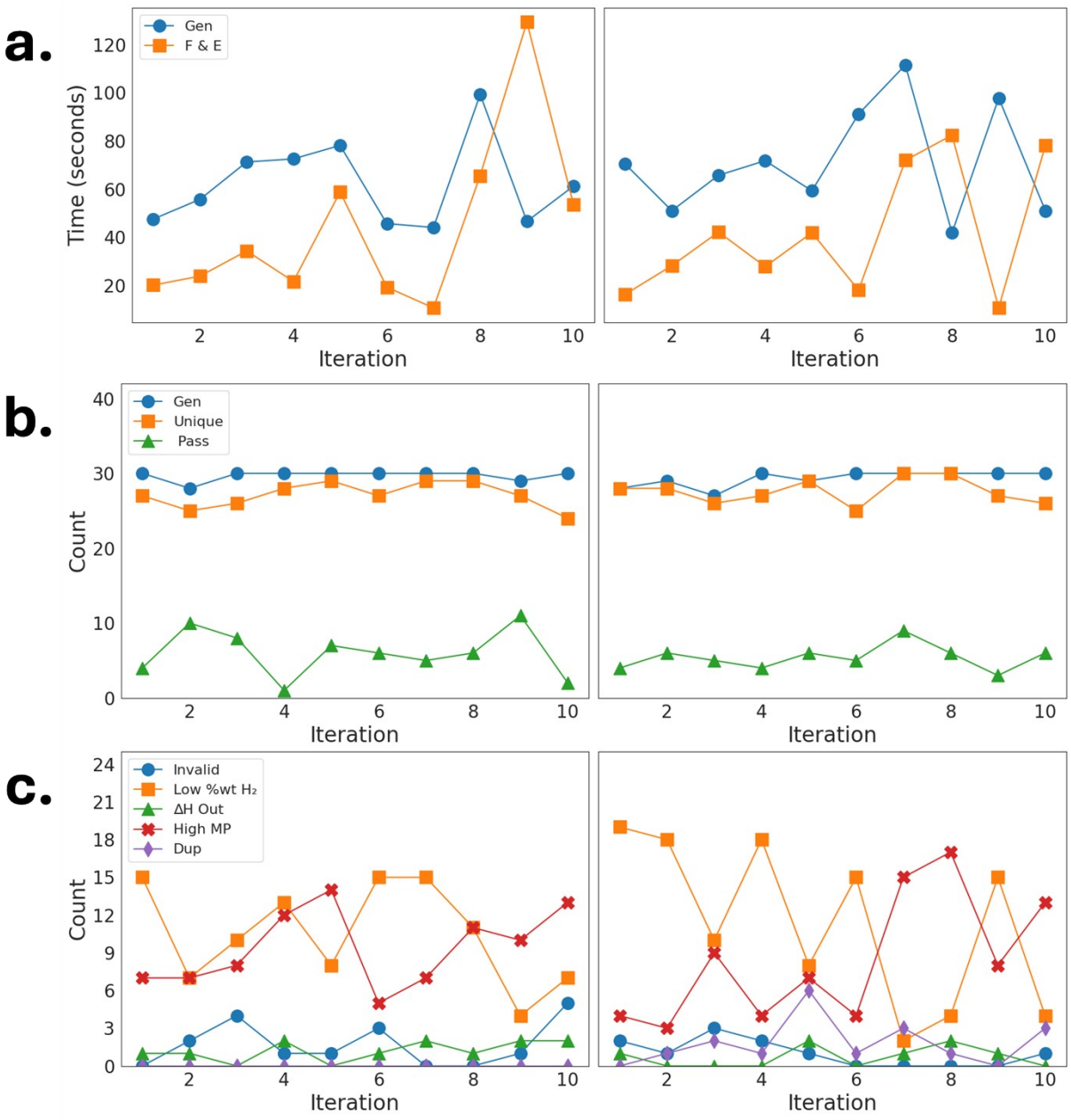


Figure S1: Performance analysis of the LLM-guided molecular generation process. Plots on the left side represent runs associated with Argo-DD-39 and plots on the right side represent runs associated with Argo-Exp-35. (a): Computation time per iteration, including LLM generation (*Gen*) and filtering and evaluation (*F & E*). (b): The number of generated SMILES (*Gen*), unique SMILES (*Unique*), and molecules that passed filtering criteria (*Pass*) across iterations. (c):

Breakdown of rejected molecules based on filtering criteria, including invalid chemical structures (*Invalid*), low hydrogen storage capacity (*Low %wt. H_2*), out-of-range reaction enthalpy (ΔH *Out*), high melting point (*High MP*), and duplicates (*Dup*).

The LLM performance plots shown in Figure S1 provide a detailed comparison of the molecular generation workflow across multiple iterations for Argo-DD-39 and Argo-Exp-35, listing the computation time, molecular processing rates, and rejection reasons. These insights help assess the efficiency and consistency of LLM-driven molecular generation, highlighting the effects of iterative refinement. The top row of Figure S1 illustrates the variation in generation time (s) across iterations. Some runs complete in as little as 60 seconds, while others take over 175 seconds. Notably, LLM generation time remains relatively stable between 50–75 seconds, whereas filtering and evaluation time fluctuates significantly between 25–125 seconds. This behavior is observed in both Argo-DD-39 and Argo-Exp-35 runs, indicating that post-generation processing contributes most to variability in iteration time, rather than the LLM itself.

The middle row of Figure S1 shows that approximately 30 SMILES strings are generated per iteration in both cases. Note that the number of molecules that successfully pass screening is significantly lower, emphasizing the importance of post-generation filtration. At each step, the workflow retains only 0–10 molecules, reinforcing that stringent filtering criteria are crucial in identifying viable LOHC candidates. The bottom row of Figure 6 provides a breakdown of rejected SMILES at each iteration. The most common rejection reasons are high MP and low hydrogen storage capacity, each accounting for up to 20 rejections (~66% rejections) per iteration. In contrast, invalid structures, duplicates, and ΔH values out of range are much less frequent, generally falling within 0–5 rejections per iteration. These trends indicate that while the LLM effectively generates valid molecular structures, many candidates fail to meet thermodynamic and phase behavior constraints, further underscoring the necessity of rigorous screening in the LOHC discovery process.

Overall, these results demonstrate that an iterative LLM-driven molecular discovery approach can efficiently generate, evaluate, and refine LOHC candidates within a reasonable computational time frame. While the generative model does not yield identical outputs for the same prompt, leading to fluctuations in processing and rejection rates, this variability enhances the exploration of chemical space by preventing the generation of redundant structures. The iterative nature of the process ensures that new candidates are continually proposed, evaluated, and refined, ultimately leading to the discovery of viable LOHC molecules within a limited number of cycles. This adaptability, combined with rapid screening and filtering, highlights the potential of integrating LLMs and ML models for accelerating molecular discovery in energy storage applications.

Dataset	Seed	Avg LLM Gen Time (s)	Avg Molecules Passed	Avg Molecules Rejected
Argo-Exp-35	Expt-31	71.1	5.4	22.2
Argo-DD-39	LOHC-7	62.2	6.0	21.1

Table S5: Summary of LLM-guided LOHC molecular generation performance for the Argo-Exp-35 and Argo-DD-39 datasets. The table reports the average LLM generation time per iteration, the number of molecules that passed filtering, and the number of rejected molecules. Argo-Exp-35 was seeded with experimentally known LOHCs from Expt-31, while Argo-DD-39 was seeded with data-driven candidates from LOHC-7.

Table S5 shows the summary of LLM-guided LOHC molecular generation performance for the Argo-Exp-35 and Argo-DD-39 datasets. Here, we list metrics such as average LLM generation time, the number of molecules passing filtering, and rejection rates. A critical difference between the two runs is the size of the seed set used: Argo-Exp-35 was initiated with 31 seed SMILES, while Argo-DD-39 started with only 7 seed SMILES. Despite this significant difference in initial molecular diversity, both runs converged to a similar total number of discovered molecules over approximately 10 iterations, suggesting that a well-curated, smaller, and chemically meaningful seed set may be of good value to prompt the LLM model.

One notable trend in the computational performance is that Argo-Exp-35 exhibits a higher average LLM generation time (71.1 s) compared to Argo-DD-39 (62.2 s). This difference could be attributed to the smaller and more focused seed set in Argo-Exp-35, which may have constrained the LLM's chemical space exploration, requiring more iterative refinement and adjustments. In contrast, the smaller seed set in Argo-DD-39 provided a concise chemical foundation, potentially leading to more direct molecular generation, hence reducing iteration complexity. However, the overall variation remains moderate, indicating that the prompt size does not drastically impact generation time, making the approach efficient even with smaller seed sets. In terms of molecular filtering, both runs exhibit a comparable number of molecules passing selection per iteration, with Argo-Exp-35 averaging 5.4 molecules and Argo-DD-39 averaging 6.0 molecules. The slight difference further suggests that the smaller but more focused seed set in Argo-DD-39 may have led to slightly higher molecular quality in generation. Similarly, rejection rates remain close, with Argo-Exp-35 rejecting 22.2 molecules per iteration compared to 21.1 in Argo-DD-39. These results indicate that the model explores a broad yet relevant molecular space in both cases, reinforcing the robustness of the generation process.

Despite differences in seed set size, both runs ultimately yield a similar number of total discovered molecules, underscoring a key insight: the quality of the seed set, guided by chemical intuition, is likely more important than its size. This observation has significant implications for resource efficiency. If a smaller, well-curated set of molecules can guide the LLM to generate high-quality candidates just as effectively as a larger set, then we can optimize the molecular discovery process while minimizing computational overhead. This means that prompting the model with a smaller but chemically meaningful seed set not only preserves performance but also saves computational resources, making this approach highly scalable for future applications in molecular discovery.

We note here, in contrast to our previous virtual screening study, which required 8,192 cores on the Argonne Leadership Computing Facility (ALCF) Theta supercomputer to screen ~160 billion molecules at a rate of 3 million molecules per second, the present LLM+ML workflow completes each generation-screening cycle in approximately 60–70 seconds on a standard laptop (Apple M1 Pro, 16 GB RAM), yielding 5–6 viable LOHC candidates per iteration. This demonstrates a dramatic reduction in computational cost while producing a comparable number of final candidates.

Text S3: Seed Molecule-based Assessment

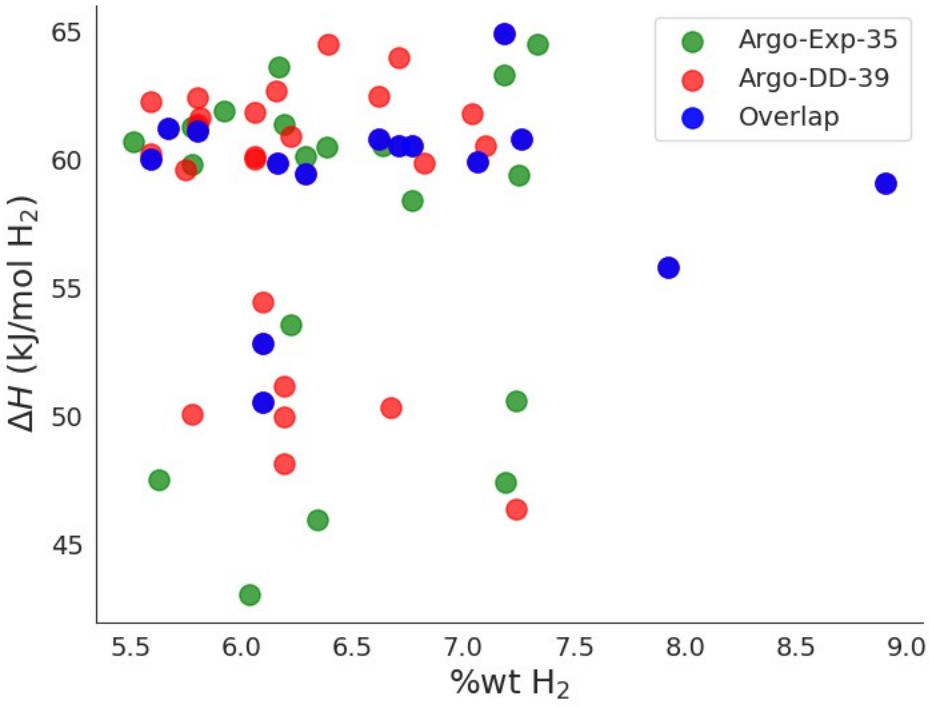


Figure S2: Scatter plot of the molecular distribution discovered in this study, showing enthalpy of hydrogenation (y-axis, ΔH) in units of kJ/mol per H_2 and the hydrogen storage capacity (x-axis, %wt. H_2). Green and red points represent molecules from the Argo-Exp-35 and Argo-DD-39 datasets (see methods section), respectively. Blue points indicate molecules present in both sets (labeled as overlap).

At the end of the iterative generation and refinement process, we obtained two distinct datasets, Argo-Exp-35 and Argo-DD-39, are based on the seed molecules used for prompting the LLM, shown in Figure S2. The dataset Argo-Exp-35 consists of 35 molecules generated using experimental LOHC molecules from Expt-31 as prompts. The Argo-DD-39 dataset contains 39 molecules from recent molecular discovery publication,[51] generated from LOHC-7 as prompts. The scatter plot in Figure S2 shows the distribution of hydrogen storage capacity (%wt. H_2) versus ΔH of the LOHC candidates discovered in both runs. Molecules from Argo-Exp-35 are shown in green, while those from Argo-DD-39 are marked in red, with overlapping candidates (molecules found in Argo-DD-39 and Argo-Exp-35) between the two sets highlighted in blue. The spread of data points indicates that both runs yielded diverse molecular candidates spanning a broad range of ΔH values (45–65 kJ/mol H_2), with many molecules concentrated around 60 kJ/mol H_2. In terms of hydrogen storage capacity, most candidates fall within the 5.5–7.5 %wt. H_2 range, aligning with practical LOHC benchmarks. Only one molecule exceeds 8 %wt. H_2, indicating that while higher-capacity candidates can emerge, they are less common in the generated sets. The presence of overlapping molecules between the two generative approaches suggests that, despite starting from different seed sets, the LLM explores a complementary and overlapping chemical space.

Category	Condition	Count
Datasets	All Argo-DD-39 molecules	39
	All Argo-Exp-35 molecules	35
Dataset Overlap	Molecules found only in Argo-DD-39	24
	Molecules found only in Argo-Exp-35	20
	Molecules found in both Argo-DD-39 and Argo-Exp-35	15
Experimental Overlap	Molecules in Argo-DD-39 that were also in Exp-31	6
Data-Driven Overlap	Molecules in Argo-Exp-35 that were also in LOHC-7	0

Table S6: Summary of the discovered molecules in Argo-DD-39 and Argo-Exp-35 datasets, highlighting dataset composition, overlap between the two sets, and connections to prior experimental (Exp-31) and data-driven (LOHC-7) studies.

Table S6 provides a breakdown of the molecular sets identified in the Argo-DD-39 and Argo-Exp-35 runs, along with their overlap. The two runs resulted in 39 and 35 molecules, respectively, with a significant portion of the generated candidates being unique to each set. Specifically, 24 molecules were exclusive to Argo-DD-39, while 20 were unique to Argo-Exp-35. Meanwhile, 15

molecules were identified in both runs, demonstrating that although the two approaches explored somewhat distinct molecular spaces, they still converged on a shared subset of viable LOHC candidates.

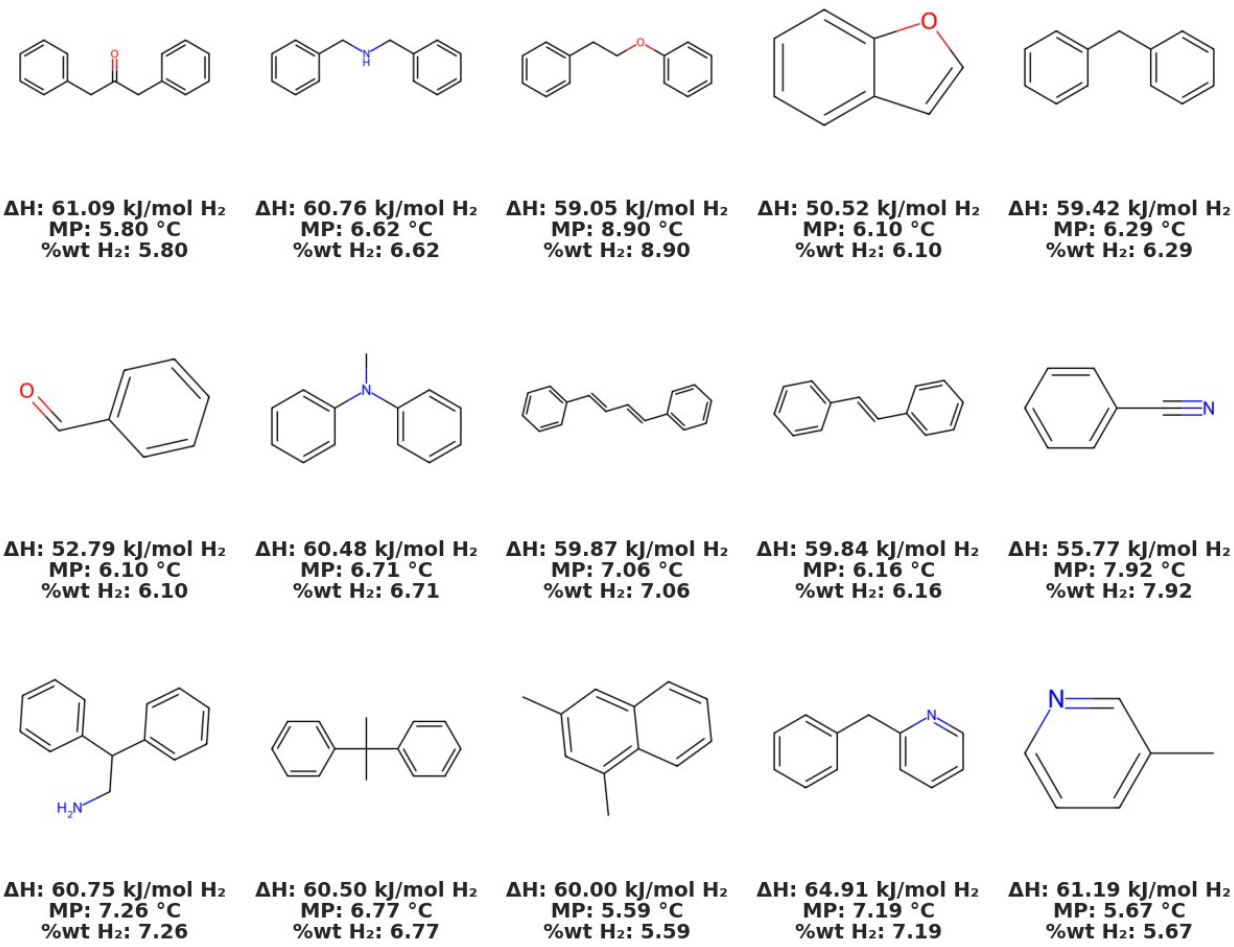


Figure S3: Structures and properties of the subset of shared molecules from Argo-Exp-35 and Argo-DD-39

The overlap analysis reveals some insights. Among the molecules generated in Argo-DD-39, six were also present in Exp-31, indicating that the generative pipeline was able to "rediscover" some previously known experimental candidates (Entry: Experimental Overlap, Table S6). This suggests that the LLM recognizes patterns aligned with experimentally studied LOHCs, which is expected given its extensive pretraining on scientific literature and molecular structures. On the other hand, no molecules from Argo-Exp-35 were found in LOHC-7, suggesting that this run explored a distinct region of chemical space compared to the previous high-throughput screening approach.[51] While this highlights the ability of our earlier data-driven discovery method to navigate large molecular spaces effectively, it also reinforces the idea that LLM-guided generation

does not simply reproduce prior computational findings but instead can follow a different molecular exploration path. The presence of a shared subset (15 molecules) between the two runs suggests that, despite these differences, both approaches can still converge on chemically relevant candidates (Figure S3 in supplementary information).

Text S4: ML Verification

Random Forest Regression was chosen for three practical reasons: (1) our training dataset (~10k reactions) is relatively small, where Random Forest performs comparably to deep learning methods; (2) it provides fast inference, which is essential for rapid screening in an iterative LLM-guided workflow; and (3) it offers interpretability through feature importance rankings, giving chemical insight into what drives ΔH.

To verify the accuracy of the ML predicted reaction enthalpies, we performed G4MP2 calculations on the 42 molecules. The results are summarized in Figure S4 alongside the errors comparing the ML-predicted hydrogenation enthalpies (ΔH ML) with their G4MP2-calculated values (ΔH G4MP2). Enthalpies and errors are also presented in the supporting information (Table S3). The ML model demonstrated strong predictive accuracy, yielding MAE of 5.32 kJ/mol H_2 and RMSD of 6.24 kJ/mol H_2. This agreement between ML and G4MP2 values reinforces the reliability of the predictive component of our workflow.

Out of the 42 molecules evaluated, 39 passed the G4MP2 validation, with hydrogenation enthalpy values remaining within the desired range of 40–70 kJ/mol H_2. However, three molecules, Argo-21, Argo-36, and Argo-38, (shown in red in Table S3) exhibited ΔH values exceeding this threshold. Notably, all three contain styrene-like structural motifs, suggesting that this functional group may contribute to elevated hydrogenation enthalpies. This then allows us to eliminate the three outliers and end up with a set of 39 new LOHC molecules.

The discrepancy between ML and G4MP2 could indicate that the training set lacked sufficient representation of vinyl structures. This can potentially limit the model's ability to capture the π-conjugation and alkene stabilization. While the ML model successfully captured broad thermodynamic trends, these outliers highlight the need for further refinement. This can be done through employing more advanced ML architectures, such as graph neural networks, to better capture structural dependencies.

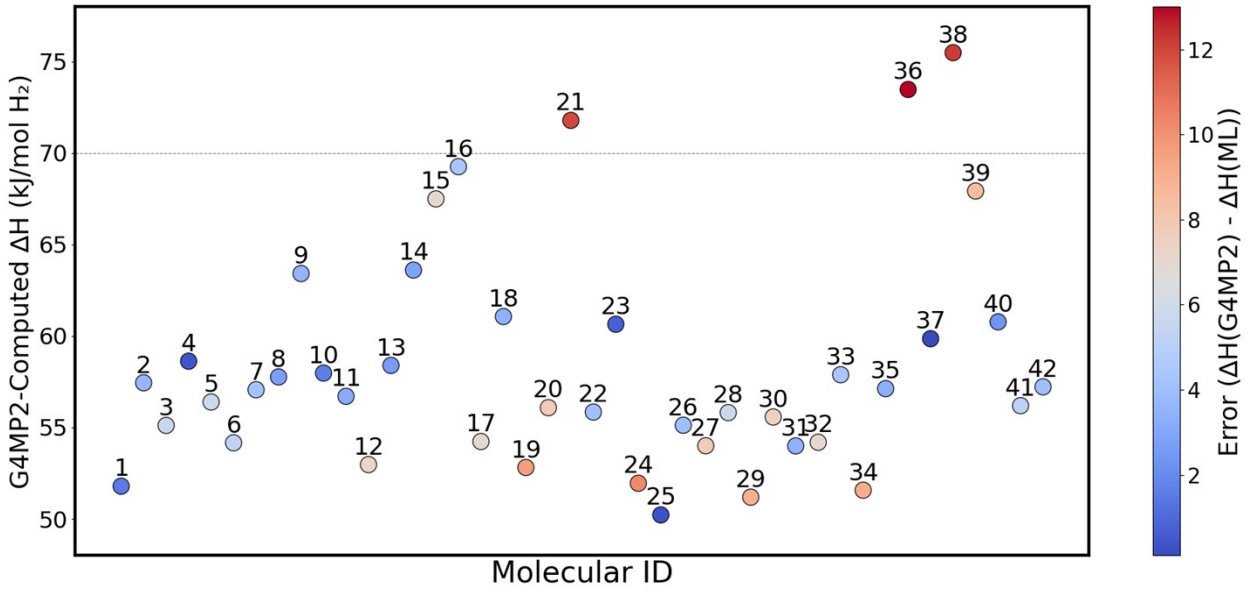


Figure S4: Scatter plot of absolute errors between ML-predicted and G4MP2 calculated hydrogenation enthalpies, in units of kJ/mol H_2 for 42 molecules (Argo-01 … Argo-42, See Figure 5 for 2D structures of the molecules). Red points indicate molecules with $\Delta H \geq 70$ kJ/mol H_2 and their corresponding IDs are given.

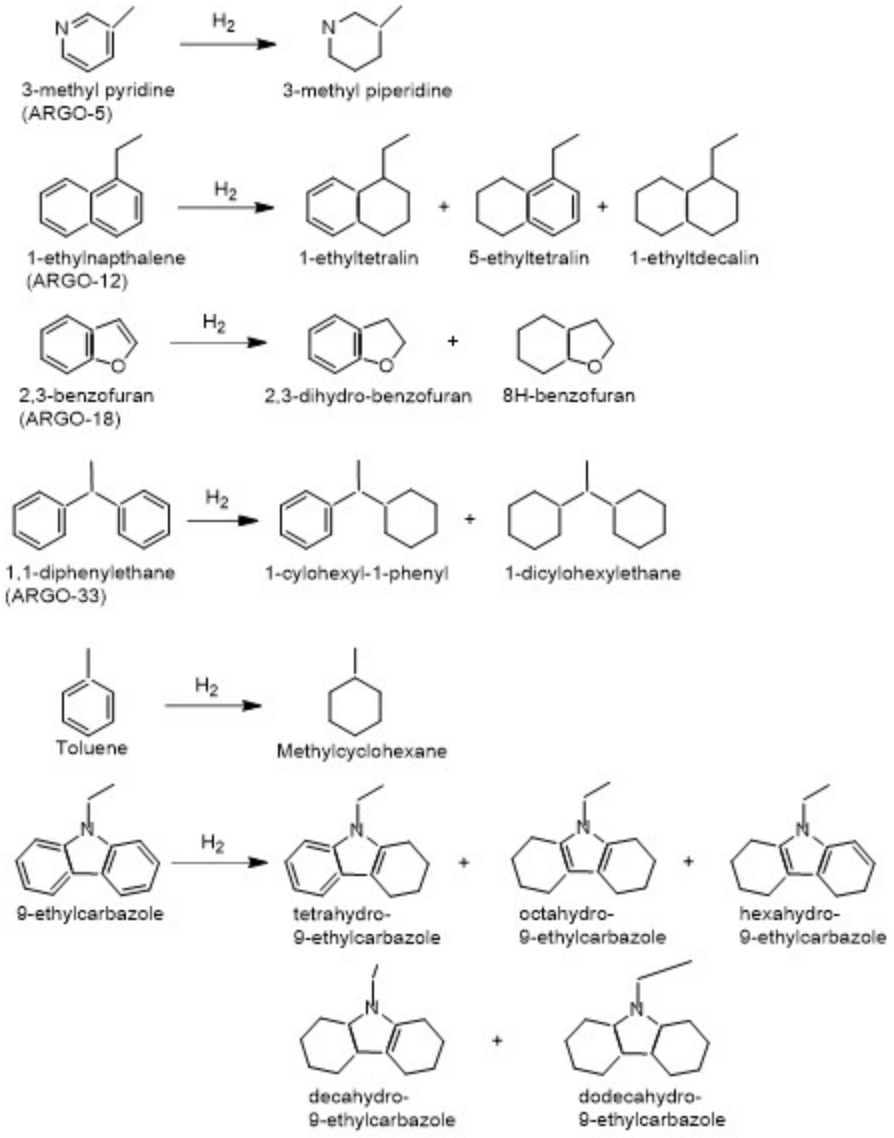


Figure S5: Hydrogenation reactions for the LOHC molecules tested.

Text S5: LOHC Design Rules:

- Limit triple bonds (sp carbons). They destabilize and raise ΔH above the desired range.
- Prefer sp^2 over sp at hydrogenation sites. Multiple π bonds on one carbon increase destabilization.
- sp^3 carbons cannot store H_2 directly but can modulate ΔH through indirect effects.
- Monocyclic rings generally lower ΔH.
  - Non-5-membered monocycles: ~−12 kJ·mol^{-1} H_2.
  - 5-membered monocycles: ~−16 kJ·mol^{-1} H_2 (more favorable than 6-membered due to strain relief).
- Bicyclic fragments strongly increase ΔH (~+50 kJ·mol^{-1} H_2). Use sparingly.
- Polycyclic fragments increase ΔH but less than bicyclic:
  - With a 5-membered ring: ~+35 kJ·mol^{-1} H_2.
  - Without a 5-membered ring: ~+10 kJ·mol^{-1} H_2.
- Heteroatoms in rings:
  - N in bicyclics slightly lowers ΔH (~−3 kJ·mol^{-1} H_2).
  - N in polycyclics raises ΔH (~+5 kcal·mol^{-1}, note units).
  - S/O counts have little net effect, but their positions matter via dipole orientation.
- In 5-membered heterocycles (pyrrole, furan, thiophene), dipole orientation relative to the heteroatom influences ΔH. Mind S/O lone-pair positioning.
- Aromatic substituents have minor impact. Even 1,3-substitution shifts are < 3 kJ·mol^{-1} H_2.
- Avoid 3-membered rings fused to other rings. They increase ΔH by ~+40 kJ·mol^{-1} H_2.
- Overall, tune hybridization, ring count, ring size, and heteroatom placement to keep ΔH in range while maintaining stability.

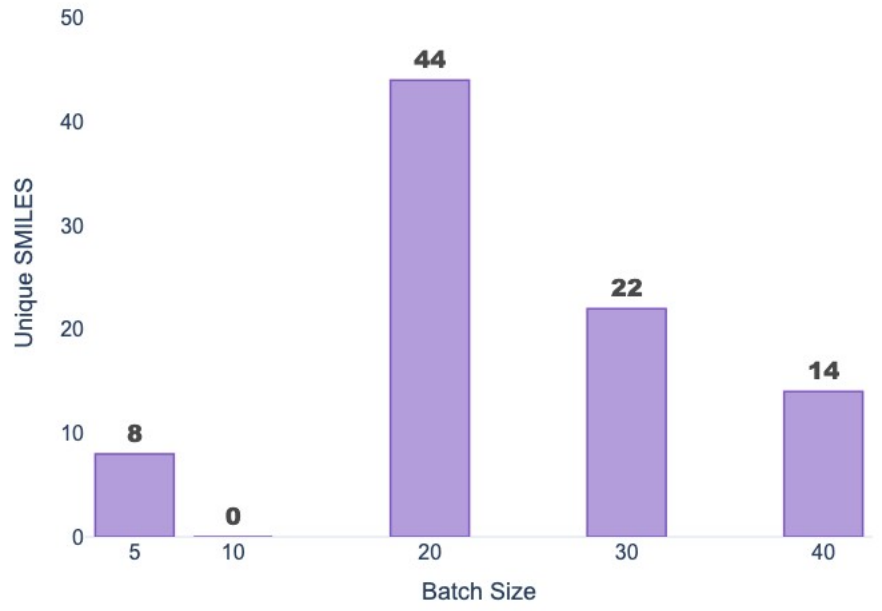
Table S7: Software and package versions used in this work.

Software/Package	Version
Python	3.10.9
RDKit	2024.09.5
NumPy	1.26.4
scikit-learn	1.7.2
PyTorch	2.2.2
PyTorch Geometric	2.7.0
Gaussian	16, Revision C.01
LLM	GPT-o1-preview

Text S6: Synthesis of Pd/NU-1000

NU-1000 was synthesized as previously described[13] and used as prepared. An automated synthesis platform (CM3 Core Module deck, Unchained Laboratories Inc.) housed in a custom-built N_2-filled glovebox (MB 200B, MBruan) was used for catalyst synthesis. The CM3 performed both solid and liquid dispensing within a 0.5% tolerance. A surface organometallic chemistry (SOMC) process was used for grafting palladium onto NU-1000.[14] In detail, 20 mg of NU-1000 was first dispensed into a 4 mL vial. Palladium (II) hexafluoroacetylacetonate was then dispensed to 5.0 wt % metal, after which 4 mL of dried toluene was added. The vial was then placed manually onto a heated unit atop a shaker plate and left to metalate for 72 h at 60 °C and 400 rpm. After cooling, the sample was washed in toluene and centrifuged (Speedvac Concentrator, SPD121P, ThermoElectron) five times. Solvent exchange using pentane was done in the last washing step. After removal of the supernatant fluid, the vial was then removed from the glovebox and vacuum-dried at 60 °C overnight.

Figure S6: Unique SMILES generated as a function of seed-set size.



Text S7: As shown in Figure S6, the number of unique SMILES exhibits a non-monotonic dependence on seed-set size. Increasing the seed set from 5 to 20 leads to a strong increase in diversity, reaching a maximum of 44 unique molecules. This indicates that moderate seed sizes provide sufficient chemical context for effective generation. Further increases in seed-set size

result in reduced diversity, with 22 and 14 unique SMILES obtained for 30 and 40 seeds, respectively. This trend suggests diminishing returns and possible redundancy in the conditioning information. The absence of valid outputs at a seed size of 10 indicates sensitivity to seed selection. These results support the presence of an optimal intermediate seed-set regime, beyond which performance does not improve and may decline.

Table S8: Random Forest model performance for ΔH prediction under different data splitting strategies.

Split Strategy	MAE (kJ/mol)	RMSD (kJ/mol)	R^2
Random	4.7	7.4	0.935
Scaffold-based	7.8	10.5	0.784

Text S8: The scaffold-based split yields a more conservative estimate of model generalization, which is the relevant regime for screening LLM-generated molecules with potentially novel scaffolds. The scaffold-split MAE of 7.8 kJ/mol remains well within the ability to rank and filter candidates within the target ΔH window of 40–70 kJ/mol per H_2 (a 30 kJ/mol range). Moreover, the ML model serves as a pre-screening tool in the discovery workflow, with promising candidates subsequently validated through quantum chemical calculations and experimental testing.

Table S9: Benchmarking of LLM-generated molecular candidates across 19 commercial and open-source large language models.

Model	Type	Reasoning	# SMILES	Avg ΔH (kJ/mol)	Avg MP (°C)	Avg %wt H_2
Claude Opus 4	Commercial	No	19	51.3 ± 7.2	13.8 ± 21.4	5.97 ± 0.47
Claude Opus 4.1	Commercial	No	33	55.6 ± 6.9	11.4 ± 18.8	6.26 ± 0.44
Claude Sonnet 4	Commercial	No	3	49.3 ± 3.2	2.5 ± 20.6	6.59 ± 0.64
GPT-3.5	Commercial	No	22	57.9 ± 7.3	20.5 ± 14.5	6.21 ± 0.36
GPT-3.5-Large	Commercial	No	22	51.9 ± 6.5	16.6 ± 18.2	6.47 ± 0.48
GPT-4-Large	Commercial	No	2	62.8 ± 0.2	31.3 ± 9.1	6.67 ± 0.00
GPT-4-Turbo	Commercial	No	3	48.7 ± 4.0	19.7 ± 13.3	5.72 ± 0.02
GPT-4.1	Commercial	No	2	51.0 ± 0.1	27.4 ± 0.0	5.71 ± 0.00
GPT-4.1-mini	Commercial	No	22	50.3 ± 5.8	15.9 ± 14.3	6.05 ± 0.34
GPT-5.1	Commercial	No	2	59.6 ± 11.1	-15.2 ± 35.0	5.73 ± 0.03
GPT-5.2	Commercial	No	19	51.8 ± 5.6	17.6 ± 16.9	6.14 ± 0.61
GPT-5.4	Commercial	No	43	56.4 ± 7.9	22.3 ± 15.2	6.08 ± 0.39
GPT-o1	Commercial	Yes	12	56.3 ± 5.7	14.8 ± 19.9	6.40 ± 0.73
GPT-o1-mini	Commercial	Yes	9	49.2 ± 2.7	10.3 ± 11.4	6.15 ± 0.44

GPT-o3	Commercial	Yes	12	50.1 ± 6.1	10.9 ± 18.2	6.38 ± 0.62
GPT-o3-mini	Commercial	Yes	21	52.9 ± 5.4	15.0 ± 19.1	5.97 ± 0.36
Llama-3.3-70B	Open-Source	No	12	55.9 ± 7.0	18.8 ± 11.8	5.92 ± 0.30
Llama-4-Scout-17B	Open-Source	No	22	60.0 ± 4.9	23.5 ± 10.5	6.80 ± 0.41
Qwen3-32B	Open-Source	Yes	8	58.3 ± 4.8	24.6 ± 9.2	6.26 ± 0.41

Text S9: Analysis of LLM benchmarking results

Table S9 summarizes the predicted thermodynamic and physicochemical properties of LLM-generated molecular candidates across 19 models spanning three commercial families: GPT (OpenAI), Claude (Anthropic), and Gemini (Google), and three open-source families: Llama (Meta), Qwen (Alibaba), and DeepSeek. For each model, the number of unique valid SMILES generated, along with the mean ± standard deviation of predicted enthalpy of formation (ΔH), melting point (MP), and %wt H_2, are reported. All predictions were obtained under identical prompting conditions using a single fixed seed (dd7) to isolate model identity as the sole source of variation. Models are further categorized by reasoning capability: five models (GPT-o1, GPT-o1-mini, GPT-o3, GPT-o3-mini, and Qwen3-32B) employ explicit chain-of-thought inference, while the remaining fourteen follow standard instruction-based generation.

Across all models, mean ΔH ranged from 48.7 kJ/mol (GPT-4-Turbo) to 62.8 kJ/mol (GPT-4-Large), mean MP from −15.2 °C (GPT-5.1) to 31.3 °C (GPT-4-Large), and mean pH_2 from 5.71 (GPT-4.1) to 6.80 (Llama-4-Scout-17B). When aggregated by model type, commercial models (n = 16) yielded a mean ΔH of 53.7 kJ/mol, MP of 16.1 °C, and pH_2 of 6.17, while open-source models (n = 3) produced modestly higher values of 58.5 kJ/mol, 22.4 °C, and 6.45, respectively. Aggregation by reasoning capability revealed even smaller differences: reasoning models predicted a mean ΔH of 53.2 kJ/mol and MP of 14.7 °C, compared to 54.7 kJ/mol and 17.7 °C for non-reasoning models, with virtually identical mean pH_2 values (6.20 vs. 6.21). These results indicate that while individual model predictions exhibit moderate variability, no systematic divergence was observed between commercial and open-source architectures or between reasoning and non-reasoning paradigms. The consistency across model classes strengthens the generalizability of the findings reported in this work and suggests that the conclusions drawn from GPT-o1preview are not an artifact of model selection.

References

(1) Harb, H.; Elliott, S. N.; Ward, L.; Foster, I. T.; Klippenstein, S. J.; Curtiss, L. A.; Assary, R. S. Uncovering Novel Liquid Organic Hydrogen Carriers: A Systematic Exploration of

Chemical Compound Space Using Cheminformatics and Quantum Chemical Methods. *Digit. Discov.* 2023. https://doi.org/10.1039/D3DD00123G.
(2) He, T.; Pei, Q.; Chen, P. Liquid Organic Hydrogen Carriers. *J. Energy Chem.* 2015, *24* (5), 587–594. https://doi.org/10.1016/j.jechem.2015.08.007.
(3) Biniwale, R.; Rayalu, S.; Devotta, S.; Ichikawa, M. Chemical Hydrides: A Solution to High Capacity Hydrogen Storage and Supply. *Int. J. Hydrog. Energy* 2008, *33* (1), 360–365. https://doi.org/10.1016/j.ijhydene.2007.07.028.
(4) Rao, P. C.; Yoon, M. Potential Liquid-Organic Hydrogen Carrier (LOHC) Systems: A Review on Recent Progress. *Energies* 2020, *13* (22), 6040. https://doi.org/10.3390/en13226040.
(5) Konnova, M. E.; Li, S.; Bösmann, A.; Müller, K.; Wasserscheid, P.; Andreeva, I. V.; Turovtzev, V. V.; Zaitsau, D. H.; Pimerzin, A. A.; Verevkin, S. P. Thermochemical Properties and Dehydrogenation Thermodynamics of Indole Derivates. *Ind. Eng. Chem. Res.* 2020, *59* (46), 20539–20550. https://doi.org/10.1021/acs.iecr.0c04069.
(6) Cui, Y.; Kwok, S.; Bucholtz, A.; Davis, B.; Whitney, R. A.; Jessop, P. G. The Effect of Substitution on the Utility of Piperidines and Octahydroindoles for Reversible Hydrogen Storage. *New J. Chem.* 2008, *32* (6), 1027. https://doi.org/10.1039/b718209k.
(7) Clot, E.; Eisenstein, O.; Crabtree, R. H. Computational Structure–Activity Relationships in H $_2$ Storage: How Placement of N Atoms Affects Release Temperatures in Organic Liquid Storage Materials. *Chem Commun* 2007, No. 22, 2231–2233. https://doi.org/10.1039/B705037B.
(8) Niermann, M.; Beckendorff, A.; Kaltschmitt, M.; Bonhoff, K. Liquid Organic Hydrogen Carrier (LOHC) – Assessment Based on Chemical and Economic Properties. *Int. J. Hydrog. Energy* 2019, *44* (13), 6631–6654. https://doi.org/10.1016/j.ijhydene.2019.01.199.
(9) Aakko-Saksa, P. T.; Cook, C.; Kiviaho, J.; Repo, T. Liquid Organic Hydrogen Carriers for Transportation and Storing of Renewable Energy – Review and Discussion. *J. Power Sources* 2018, *396*, 803–823. https://doi.org/10.1016/j.jpowsour.2018.04.011.
(10) NIST: Reaction Enthalpy Was Calculated Form the Enthalpies of Formation of Reactants and Products, Which Were Obtained from the NIST Chemistry WebBook.
(11) Linstrom, P. NIST Chemistry WebBook, NIST Standard Reference Database 69, 1997. https://doi.org/10.18434/T4D303.
(12) Kariya, N.; Fukuoka, A.; Ichikawa, M. Efficient Evolution of Hydrogen from Liquid Cycloalkanes over Pt-Containing Catalysts Supported on Active Carbons under "Wet–Dry Multiphase Conditions." *Appl. Catal. Gen.* 2002, *233* (1–2), 91–102. https://doi.org/10.1016/S0926-860X(02)00139-4.
(13) Hackler, R. A.; Pandharkar, R.; Ferrandon, M. S.; Kim, I. S.; Vermeulen, N. A.; Gallington, L. C.; Chapman, K. W.; Farha, O. K.; Cramer, C. J.; Sauer, J.; Gagliardi, L.; Martinson, A. B. F.; Delferro, M. Isomerization and Selective Hydrogenation of Propyne: Screening of Metal-Organic Frameworks Modified by Atomic Layer Deposition. *J. Am. Chem. Soc.* 2020, *142* (48), 20380– 20389. https://doi.org/10.1021/jacs.0c08641
(14) Witzke, R. J.; Chapovetsky, A.; Conley, M. P.; Kaphan, D. M.; Delferro, M. Nontraditional Catalyst Supports in Surface Organometallic Chemistry. *ACS Catal.* 2020, *10* (20), 11822– 11840. https://doi.org/10.1021/acscatal.0c03350