Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: H. Harb, M. Ferrandon, T. A. Goetjen, S. Lee, O. K. Farha, M. Delferro and R. Surendran Assary, *Digital Discovery*, 2025, DOI: 10.1039/D6DD00102E.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D6DD00102E

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

Abstract. Accelerating the discovery of new molecules with targeted properties is a central challenge in molecular design. In this contribution, we present an AI-driven molecular discovery framework that integrates Large Language Models (LLMs) for generative molecular design with Machine Learning (ML)-based screening to identify novel Liquid Organic Hydrogen Carrier (LOHC) candidates. Using the developed framework, LOHC molecules were systematically generated, evaluated, and refined iteratively, combining LLM-guided molecular generation and

View Article Online
DOI: 10.1039/D6DD00102E

ML-predicted hydrogenation enthalpies (ΔH), under physicochemical property constraints such as optimal melting points (MP), desired hydrogen storage capacity (wt % H_2), and synthetic accessibility (SA) scores. This approach enabled the discovery of 42 new LOHC candidates in two distinct campaigns, one seeded with experimentally known and another with previously computationally identified LOHCs, respectively. Although we began with different numbers of starting molecules (31 vs. 7 seed molecules), both runs yielded a comparable number of viable candidates, suggesting an influence of chemically intuitive seed molecule selection for success. Selected LOHC molecules, such as 3-methyl pyridine, 1-ethylnapthalene, 1,1-diphenylethane, and benzofuran, were experimentally tested and compared with benchmark LOHCs (toluene and 9-ethylcarbazole) for hydrogenation using a series of commercial supported metal catalysts. The order of conversion into fully hydrogenated products at 200 ˚C was 3-methyl pyridine (100 %) > 9-ethyl carbazole (86.4 %) > 2,3-benzofuran (74 %) > 1,1-diphenylethane (66.9 %) > 1-ethylnapthalene (66.7 %) > toluene (57 %), further validating the AI-guided molecular design. This study demonstrates promise of LLM-driven molecular design in conjunction with ML-based screening for accelerated discovery and design of molecules.

1. Introduction

Generative molecular discovery is rapidly emerging as a transformative approach to navigating the immense molecular space (> 10^{60}).[1–6] Traditional brute-force molecular screening trial-and-error synthesis is infeasible due to the vast chemical space.[6–14] To overcome these limitations, generative artificial intelligence (AI), particularly Large Language Models (LLMs), has gained

View Article Online
DOI: 10.1039/D6DD00102E

prominence.[15–20] Originally developed for natural language tasks, pre-trained LLMs can be adapted to generate and refine molecular structures from simple text prompts, enabling efficient exploration of chemical space beyond the constraints of conventional methods.[1,21,22] In molecular discovery, predictive and generative AI play complementary roles.[23–25] Predictive AI, employing techniques such as random forest regression or graph neural networks (GNNs), forecasts molecular properties and interactions based on historical data, enabling rapid screening without resource-intensive experimental validation.[7,9,26–30] In contrast, generative AI, including LLMs and diffusion models, creates novel molecular structures by learning patterns from training data and generating candidates that may not exist in known databases.[2,3,31–35] These generative models explore uncharted regions of chemical space, guided by learned design principles or optimization rules.[2,3] While predictive models excel at accurately estimating properties within the boundaries of known structure-property relationships, generative approaches transcend these boundaries by proposing entirely new molecules. This synergy between generative and predictive AI holds the potential to revolutionize materials discovery, dramatically accelerating the design process by expanding the search space and efficiently filtering candidates, far surpassing the capabilities of traditional screening methods.

View Article Online
DOI: 10.1039/D6DD00102E

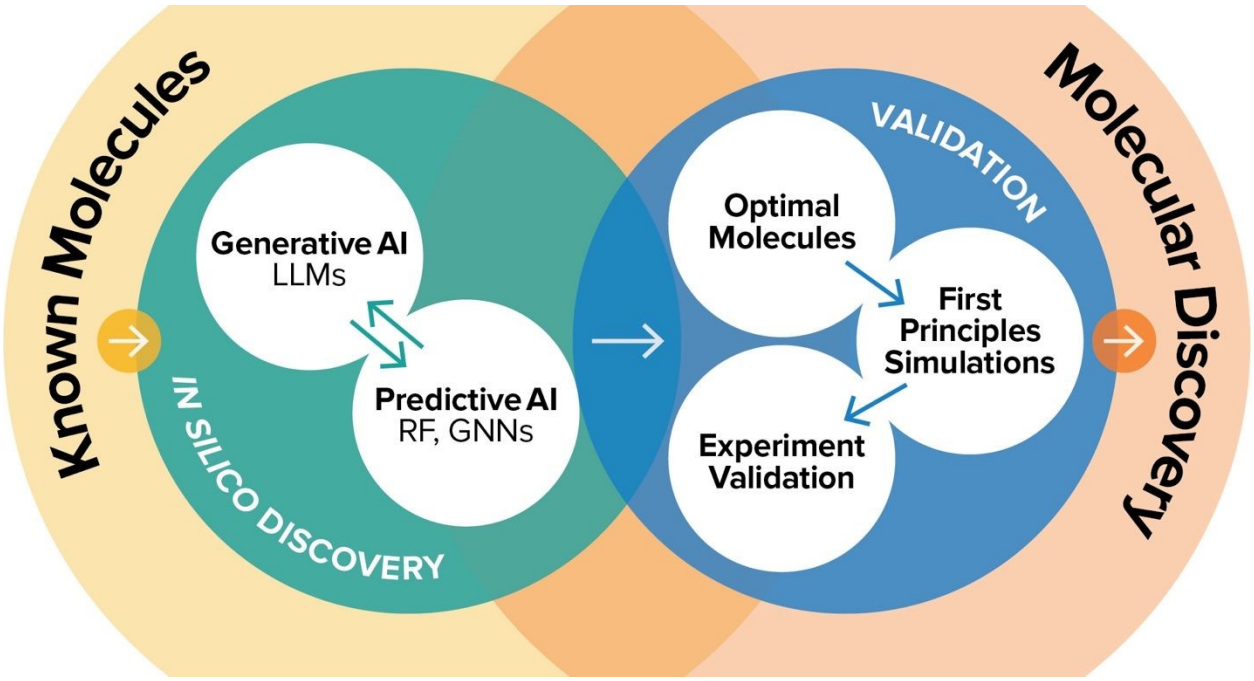
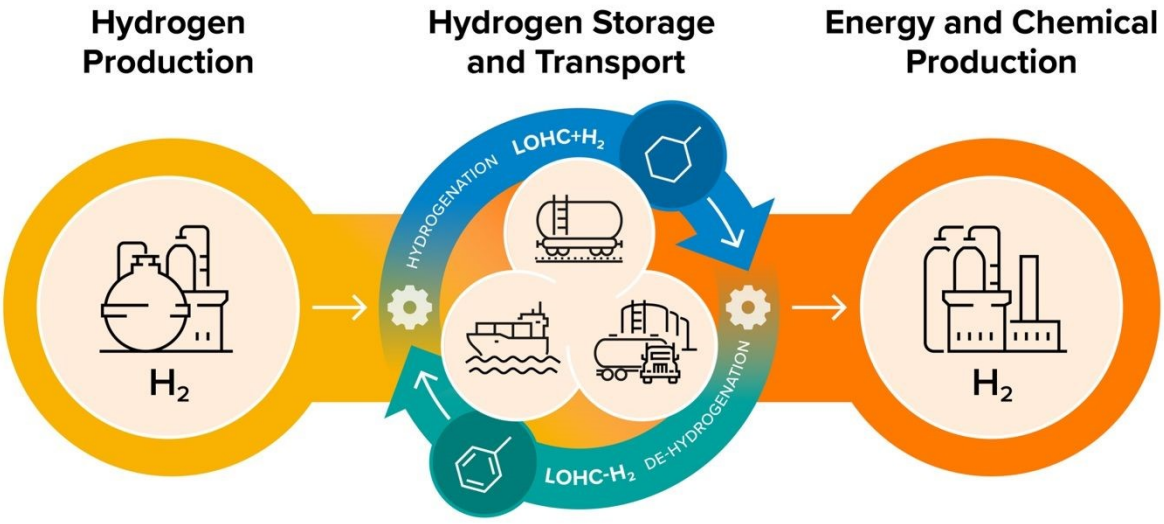


Figure 1: Overview of the molecular discovery workflow depicting the synergy between generative AI, predictive AI, and validation.

As a representative case, AI-driven discovery of Liquid Organic Hydrogen Carrier (LOHC)[36–39] molecules was chosen, which provides a tractable and well-defined system for demonstrating LLM-driven molecular design and discovery.[36,38,40–45] LOHCs are unsaturated organic molecules that chemically bind and release hydrogen via reversible hydrogenation/dehydrogenation cycles, facilitated by catalysts.[38,39,46] Their liquid-phase stability, safety, and compatibility with existing fuel infrastructure make them a practical and scalable energy storage solution.[37,39,47,48] Unlike compressed hydrogen, LOHCs eliminate boil-off losses and enable long-term storage and transport without degradation.[49,50] In terms of research and development, comprehensive overviews of potential LOHC systems have been widely reported in the literature and exemplar systems include benzene, toluene, N-ethyl carbazole, and dibenzyl toluene, among others.[38,47,48,51,52] At present, the

View Article Online
DOI: 10.1039/D6DD00102E

LOHCs face significant challenges and limitations that hinder their widespread application.[48,53,54] An optimal LOHC molecule should possess a combination of properties that ensure efficiency, stability, and practicality for hydrogen storage and transport.[36,37,40] LOHCs must exhibit a gravimetric hydrogen capacity (wt % H_2) above 5.5% to ensure sufficient energy density for transportation applications[36,48] and an optimal enthalpy range (40 – 70 kJ/mol per H_2) for efficient low-temperature cycling between the hydrogenated and dehydrogenated forms.[36,39] In addition, both the hydrogen-lean and hydrogen-rich states should remain liquid at room temperature to facilitate handling and storage.[36,39,48] LOHCs must undergo hydrogenation and dehydrogenation without molecular degradation[36,37,48,55]and have low toxicity to ensure safe and practical implementation.[56,57] These challenges highlight the continuous need for novel LOHC molecules with improved stability, thermodynamics, and catalytic performance, motivating the search for an accelerated and data-driven molecular design approaches to LOHC discovery.[36]



View Article Online
DOI: 10.1039/D6DD00102E

Figure 2. Schematic representation of hydrogen production, transportation, and use in chemical manufacturing using toluene/methylcyclohexane cycle as an LOHC system.

Recently, we have established a computational screening approach to accelerate LOHC discovery.[40] In this massive in silico molecular screening, we systematically processed 160 billion molecules from the ZINC15[58] and GDB-17[59] molecular databases, leveraging cheminformatics-based selection criteria and accurate quantum chemical calculations.[40] This approach led to the identification of 41 novel LOHC candidates with enhanced hydrogen storage capacity and favorable thermodynamic properties.[40] To address the accurate data deficiency of LOHCs, we have developed the QM9-LOHC dataset that includes 10k dehydrogenation reactions calculated using the high-accuracy G4MP2 quantum chemical method.[60] This dataset supports ML-based prediction of hydrogenation enthalpies and facilitates data-driven LOHC molecular discovery. Paragian and colleagues[42] screened over one million PubChem[61] molecules as potential LOHC candidates using RING[62] approach for structure generation, OPERA[63] for phase property predictions, and ML models for dehydrogenation enthalpies. They identified 14,000 feasible LOHC pairs and selected 37 promising candidates based on hydrogen capacity, synthetic accessibility, and key molecular features analyzed via sparse linear discriminant analysis.[42] Despite these advances, many viable LOHC candidates may still be overlooked due to the scale of chemical space and constraints of the applied screening methods.[40,42,64]

Building upon this foundation, this present study integrated generative AI (LLMs) and predictive ML using Random Forest (RF) regression into an iterative molecular discovery framework, where LLMs generated new molecular candidates from seed structures, and an ML model trained on the

View Article Online
DOI: 10.1039/D6DD00102E

QM9-LOHC dataset[60] predicted the enthalpy of hydrogenation (ΔH) for rapid screening. This AI-driven design process enables efficient generation, evaluation, and refinement of candidate molecules without the need for the costly step of finetuning the LLM. Despite differences in the seed sets used, both approaches converged on a similar number of viable molecules, leading to the discovery of 42 distinct new LOHC structures that satisfy key thermodynamic and MP criteria. From these, 4 LOHC molecules (3-methyl pyridine, 1-ethylnapthalene, 1,1-diphenylethane, and 2,3-benzofuran) were selected and compared with benchmark LOHCs (9-ethylcarbazole and toluene) for the hydrogenation reaction at 150 ˚C and 200 ˚C using commercial catalysts. Based on the experimental studies, the order of conversion into fully hydrogenated products is consistent with computational predictions. These results demonstrate that a well-curated data set of molecules (seed set) is more critical than its size, permitting computationally efficient molecular generation. By combining LLM-driven molecular ideation with ML-based screening, this framework accelerates in silico discovery and materials selection processes for targeted experimentation to complete the discovery loop.

2. Results and Discussion

2.1 Generation of LOHC Molecules

In Figure 3, the computational molecular discovery workflow that combines generative and predictive AI tools in an iterative loop to develop new molecules from system prompt is shown. We begin with carefully selected seed molecules, either experimentally known or computationally discovered LOHCs, that guide the LLM to generate chemically valid and structurally diverse

View Article Online
DOI: 10.1039/D6DD00102E

Simplified Molecular Input Line Entry System (SMILES) strings. The seeds define the initial chemical space and help the model focus on structures that are realistic for hydrogen storage. The LLM samples multiple variations for each seed to explore substitutions, ring patterns, and functional group changes.

Each generated molecule is screened using an ML model, trained on QM9-LOHC dataset to predict hydrogenation enthalpies (ΔH). Details on the ML model are provided in the methods section. The screening step also filters invalid structures. It assigns priority to molecules that stay within known LOHC chemical classes while expanding the space with new motifs. Only molecules that fall within the desired thermodynamic window (40–70%kJ/mol per H_2) and meet physicochemical criteria (melting point ≤ 40 °C, wt H_2 ≥ 5.5%, synthetic accessibility (SA) score ≤ 3.3) are retained. The workflow then feeds these retained structures back into the generator. This feedback step helps steer the next round of sampling toward chemistries with improved ΔH values and higher hydrogen capacity. Each cycle expands the pool of candidates while keeping the search focused on feasible LOHC designs. This refinement process repeats for up to ten iterations or 200 unique molecules. These unique molecules and their enthalpies (ΔH) are validated using first principles calculations, ensuring efficiency and accuracy in identifying LOHCs.

View Article Online
DOI: 10.1039/D6DD00102E

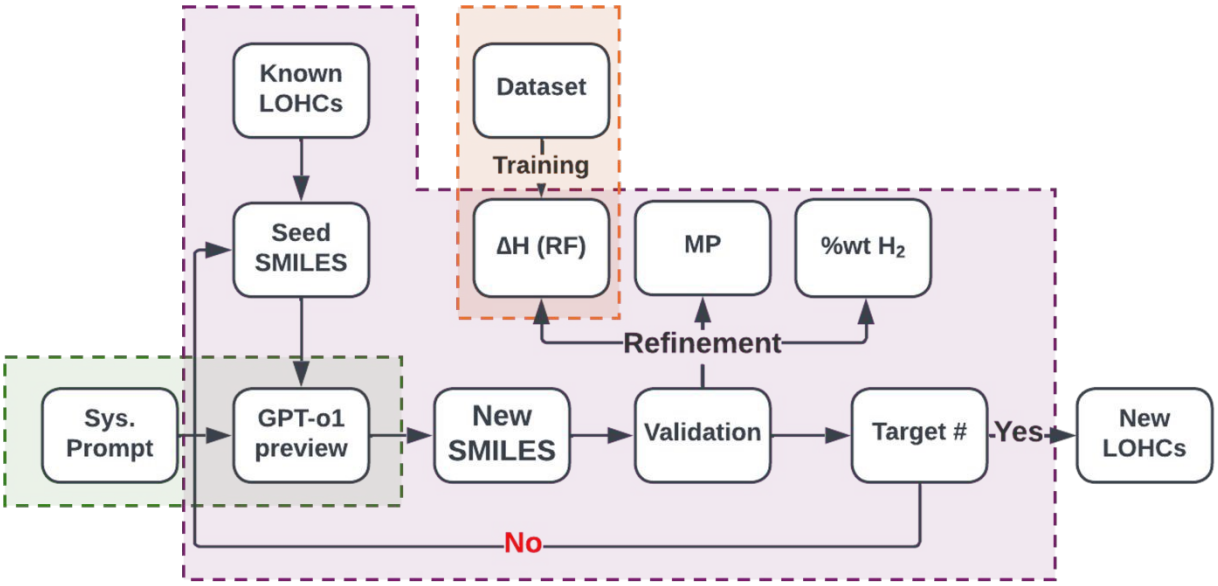


Figure 3. Workflow implemented in this study. The workflow has three main components, shown with the green, purple, and orange shaded regions. The green region is the LLM agent which contains the GPT-o1preview model (accessed via Argo API) and the custom system prompt. The orange region shows the ML part, which contains the dataset and the ML predictive model trained on that dataset and used to evaluate the generated Simplified Molecular Input Line Entry System (SMILES) strings. The purple box shows a generative loop that starts the process with the seed SMILES strings, prompted into LLM agent, and then generates, validates, and refines new SMILES strings. 'Sys. Prompt' refers system prompt, ΔH (RF) denotes reaction enthalpies predicted by Random Forest model, MP denotes melting point (predicted using the OPERA[63] model), 'wt % H_2' represents hydrogen storage capacity, and Target # represents the user-defined target number of SMILES (set at a maximum of 200).

2.2 Selection of LOHC Molecules

View Article Online
DOI: 10.1039/D6DD00102E

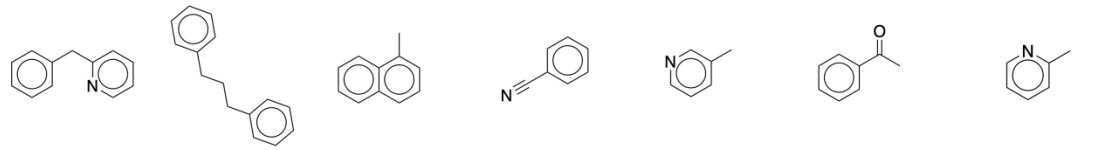
To down select the promising LOHC molecules, we imposed additional criteria for filtering, including practicality and synthesizability. First, both the hydrogen-lean (HL) and hydrogen-rich (HR) forms were required to have MPs below 40%˚C. Second, we imposed a synthesizability filter to gauge the likelihood that these molecules can be feasibly prepared in a laboratory setting. We retained only those molecules that either (a) demonstrated an SA[65] score below 3.3 based on our previously established cutoff for practical synthesis[65] or (b) were already in the PubChem database.[61] The SA score factor reflects complexity and likely number of synthetic steps required, while a PubChem listing indicates prior knowledge or availability of the compound. Additionally, all halogen-bearing structures were removed due to their susceptibility to undesired reactions (e.g., elimination) under LOHC operating conditions.[40] By combining these two criteria, we prioritized LOHC structures that were both thermally suitable and synthetically feasible.

Using the molecular selection workflow (Fig 3), 42 new LOHC molecules were identified, schematically shown in Figure 4. Note that additional details of these molecules, including their melting and boiling points and SA scores for the hydrogen-rich and -lean forms, are presented in the Supporting Information (Table S4). These molecules exhibit a broad spectrum of physical and thermodynamic properties, emphasizing how structural variations can profoundly impact key performance indicators. In most cases, the MPs of both hydrogen-lean (HL) and hydrogen-rich (HR) forms are below 40%˚C, indicating that they remain liquid under ambient conditions. Among these, 11 molecules have MPs below 0%˚C. Notably, **7** (2-methylpyridine), **16** (styrene), and **36** (propenyl benzene) melt at -54.9%˚C, at -30.1%˚C, at -28.3%˚C, respectively, thereby minimizing solidification risks in colder environments. Although their boiling points (BP) span a wide range

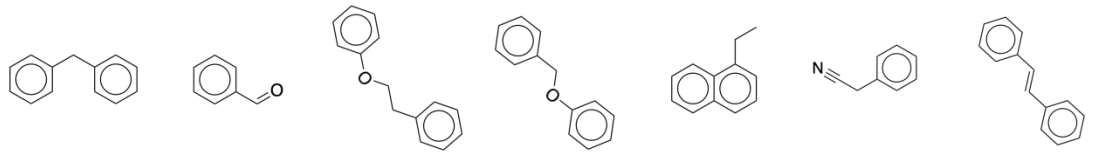
View Article Online
DOI: 10.1039/D6DD00102E

(~130–150%˚C to > 300%˚C), each of these candidates remains liquid at room temperature, further underscoring their use as an LOHC.

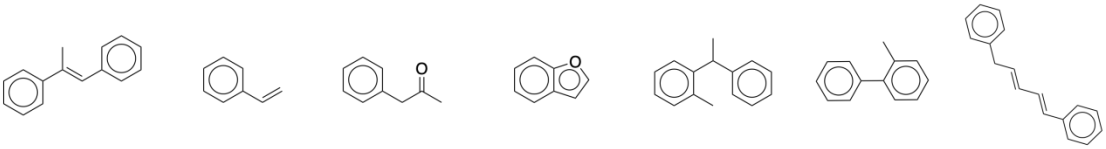
View Article Online
DOI: 10.1039/D6DD00102E



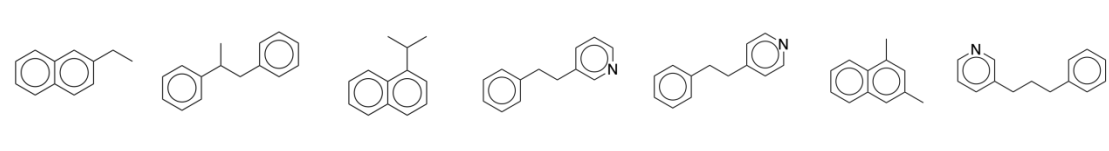
1 L, B+ Δ 50.3, $\mathcal{H}$ 6.67
2 L, B+ Δ 61.1, $\mathcal{H}$ 5.80
3 S, B− Δ 60.8, $\mathcal{H}$ 6.62
4 L, B− Δ 59.1, $\mathcal{H}$ 8.90
5 L, B− Δ 50.5, $\mathcal{H}$ 6.10
6 L, B− Δ 59.4, $\mathcal{H}$ 6.29
7 L, B− Δ 52.8, $\mathcal{H}$ 6.10



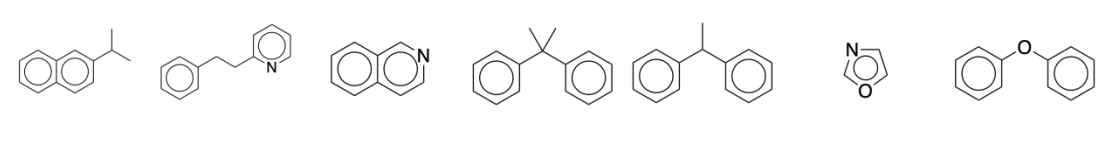
8 L, B+ Δ 60.5, $\mathcal{H}$ 6.71
9 L, B− Δ 59.9, $\mathcal{H}$ 7.06
10 S, B+ Δ 59.6, $\mathcal{H}$ 5.75
11 S, B+ Δ 59.8, $\mathcal{H}$ 6.16
12 L, B+ Δ 60.1, $\mathcal{H}$ 6.06
13 L, B− Δ 55.8, $\mathcal{H}$ 7.92
14 S, B+ Δ 60.7, $\mathcal{H}$ 7.26



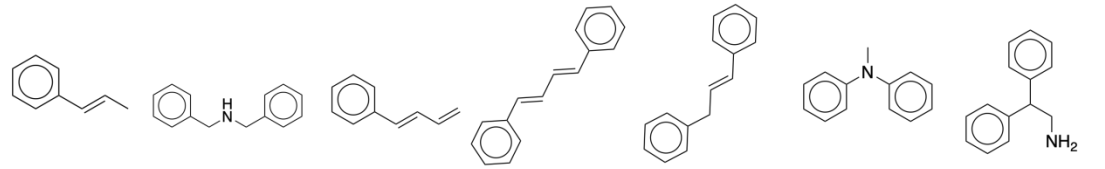
15 S, B+ Δ 60.5, $\mathcal{H}$ 6.77
16 L, B− Δ 64.9, $\mathcal{H}$ 7.19
17 L, B− Δ 61.2, $\mathcal{H}$ 5.67
18 L, B− Δ 64.5, $\mathcal{H}$ 6.39
19 L, B+ Δ 62.4, $\mathcal{H}$ 5.80
20 L, B+ Δ 63.9, $\mathcal{H}$ 6.71
21 S, B+ Δ 59.8, $\mathcal{H}$ 6.82



22 L, B+ Δ 60.0, $\mathcal{H}$ 6.06
23 S, B+ Δ 61.4, $\mathcal{H}$ 5.80
24 L, B+ Δ 62.2, $\mathcal{H}$ 5.59
25 L, B+ Δ 49.9, $\mathcal{H}$ 6.19
26 L, B+ Δ 51.1, $\mathcal{H}$ 6.19
27 S, B+ Δ 61.8, $\mathcal{H}$ 6.06
28 L, B+ Δ 50.1, $\mathcal{H}$ 5.78



29 L, B+ Δ 60.2, $\mathcal{H}$ 5.59
30 L, B+ Δ 48.1, $\mathcal{H}$ 6.19
31 L, B− Δ 50.6, $\mathcal{H}$ 7.24
32 S, B+ Δ 61.2, $\mathcal{H}$ 5.80
33 L, B+ Δ 53.5, $\mathcal{H}$ 6.22
34 L, B− Δ 60.7, $\mathcal{H}$ 5.52
35 S, B+ Δ 60.5, $\mathcal{H}$ 6.63



36 L, B− Δ 60.5, $\mathcal{H}$ 6.39
37 L, B+ Δ 59.8, $\mathcal{H}$ 5.78
38 L, B− Δ 63.3, $\mathcal{H}$ 7.19
39 S, B+ Δ 59.4, $\mathcal{H}$ 7.25
40 S, B+ Δ 58.4, $\mathcal{H}$ 6.77
41 L, B+ Δ 61.3, $\mathcal{H}$ 6.19
42 S, B+ Δ 61.2, $\mathcal{H}$ 5.78

View Article Online
DOI: 10.1039/D6DD00102E

Figure 4. Schematics of structures (entries 1 to 42)/important properties of LOHCs identified in this study. All molecules are available in PubChem. Labels: Physical state (Liquid = L; Solid = S at room temperature), boiling point (B+, B–), enthalpy of hydrogenation (Δ, in kJ/mol H_2), and hydrogen storage capacity ($\mathcal{H}$, in wt %). Note: Each molecule exhibits properties (see Table S4, Supporting Information) that qualify it as a promising hydrogen storage candidate based on thermodynamic and physicochemical characteristics.

Beyond phase behavior, all 42 compounds shown in Figure 4 feature ΔH in the desired 40–70%kJ%mol^{-1}%H_2 range. The entries in Figure 4, **16** (styrene) and **18** (benzofuran), sit near the upper end (~64–65%kJ%mol^{-1}%H_2). Despite this higher ΔH, both remain attractive: benzofuran has previously been shown to undergo efficient hydrogenation, even before its relevance to LOHC applications was established.[66] Benzofuran is also a key structural motif in bio-derived compounds.[67] Styrene's very low MP potentially operates at sub ambient temperatures, compensating for its higher enthalpy requirement by. By contrast, entries **30**, **25**, and **28** (all belonging to the phenylethyl- or phenylpropyl-pyridine family) exhibit lower ΔH values (~48–50%kJ%mol^{-1}%H_2), potentially enabling easier dehydrogenation at moderate temperatures. Meanwhile, gravimetric hydrogen capacities (wt %% H_2) run from about 5.5 % to nearly 9 %%, with **4** (benzonitrile) standing out at 8.9 %.% H_2, the highest among this set.

Finally, synthesizability and availability were assessed using PubChem listings and SA scores, indicating that all 42 molecules are likely available in known databases or accessible via standard synthetic routes. Notably, each hydrogen-lean (HL) and hydrogen-rich (HR) form meets our

View Article Online
DOI: 10.1039/D6DD00102E

practical synthesis cutoff (SA <%3.3, as discussed in our prior work[40,65]) or appears in PubChem, minimizing potential barriers to laboratory or pilot-scale validation. This combination of low MPs, balanced ΔH values, and synthetic accessibility underscores their commercial potential as next-generation LOHCs.

2.3 Experimental validation

Five LOHC molecules, **4** (benzonitrile), **5** (3-methyl pyridine), **12** (1-ethylnapthalene), **18** (2,3-benzofuran), and **33** (1,1-diphenylethane) were selected for experimental validation and compared to toluene and 9-ethylcarbazole as benchmark LOHCs. These molecules were selected based on their high wt %% H_2 (above 6%), low MPs (all below 0 ˚C), and their availability in PubChem. These molecules were chosen as representative candidates spanning different chemical classes to demonstrate the viability of the AI-driven framework; other promising candidates, (e.g. benzonitrile; 8.9 wt% H_2), remain targets for future experimental validation. One of the most studied LOHCs is 9-ethylcarbazole.[68] This heterocyclic compound can theoretically take up 6 moles of equivalent hydrogen (5.7 %). However, it converts into 4 partially hydrogenated products.[69] Schemes of the hydrogenation reactions for all the LOHC molecules are shown in Figure S5. Hydrogenation catalysts were chosen according to literature.[70] 10 wt % Pd on carbon and alumina, 5 wt % Pt on NU-1000, as an example of metal-organic framework, and 5 wt % Rh on alumina were employed as hydrogenation catalysts at 150 ˚C and 200 ˚C for 12 h at 300 psi of H_2. Table 1 shows the conversion of 3-methyl pyridine into 3-methyl piperidine and toluene into methyl cyclohexane at 150 ˚C and 200 ˚C for 12 h, respectively. All 4 catalysts exhibited greater conversion of 3-methyl pyridine compared to toluene to fully hydrogenated product, with

View Article Online
DOI: 10.1039/D6DD00102E

Rh/Al_2O_3 being the most active catalyst at 150 ˚C. At 200 ˚C, full conversion is achieved for 3-methyl pyridine using Pd/C and Rh/Al_2O_3 and almost full conversion for Pd/Al_2O_3. Comparison with other substrates (1-ethylnapthalene, 2,3-benzofuran, and 1,1-diphenylethane) using the most performant catalyst, Rh/Al_2O_3, at 200 ˚ C is listed in Table 2. All three substrates are fully converted to either fully or partially hydrogenated products, with 74% selectivity for 8H-benzofuran. 9-ethylcarbazole converts into fully hydrogenated product (86.4%). However, the catalyst also converts into the de-ethylated product (9.2%). The hydrogenation of benzonitrile leads to the formation of dibenzylamine with high conversion using the Rh/Al2O3 catalyst in comparison with the Pt/NU-1000 (Table 3). The order of conversion into fully hydrogenated products at 200 ˚C is 3-methyl pyridine > 9-ethyl carbazole > 2,3-benzofuran > 1,1-diphenylethane > 1-ethylnapthalene > toluene. Recycling experiments show that Rh/Al2O3 maintain a high activity (Table 3).

Table 1. Results from hydrogenation experiments at 150 ˚C (300 psi H_2) and 200 ˚C (600 psi H_2) for 12 h.

Catalyst	3-Methyl pyridine conversion (%)		Toluene conversion (%)	
	150 °C	200 °C	150 °C	200 °C
10 wt % Pd/C	35.8	100.0	16.2	45.3
5 wt % Rh/Al_2O_3	58.9	100.0	24.9	57.0
10 wt % Pd/Al_2O_3	38.5	96.8	7.0	45.6
5 wt % Pt/NU-1000	29.1	75.4	5.8	39.7
None	0	0	5.0	9.0

Table 2. Results from hydrogenation experiments using 10 mg Rh/Al_2O_3 at 200 ˚C for 12 h at 600 psi H_2.

View Article Online
DOI: 10.1039/D6DD00102E

Substrate	Fully hydrogenated (%)	Partially hydrogenated (%)	Other (%)
1-Ethylnapthalene	66.7	31.3	-
2,3-Benzofuran	74.0	26.0	-
1,1-Diphenylethane	66.9	33.1	-
9-Ethyl carbazole	86.4	4.4	9.2

Table 3. Recycling experiments using 5 wt % Rh/Al_2O_3 and 5 wt % Pt/NU-1000. Results from hydrogenation experiments at 200 °C (600 psi H_2) for 12 h.

Catalyst	3-Methyl pyridine conversion (%)		Benzonitrile conversion[a] (%)	
	#1	#2	#1	#2
5 wt % Rh/Al_2O_3	100.0	100.0	98.1	97.9
5 wt % Pt/NU-1000	75.4	0	0.5	0.5

[a] conversion to dibenzylamine

Conclusions

This study demonstrates that combining LLM-driven molecular generation with ML-based screening provides an efficient path for discovering new LOHC molecules. The workflow identified chemically meaningful structures, evaluated them with rapid predictive models, and refined the candidates through an iterative loop that focuses on thermodynamic and physicochemical targets. Using this approach, 42 new LOHC candidates were discovered and 39 of them were validated using high-accuracy G4MP2 calculations. Four of these molecules were tested experimentally and showed hydrogenation performance that aligns with the AI predictions. These results show that a small but well-chosen seed set can guide LLMs toward novel and viable LOHC designs. By combining pretrained generative models with modular screening, this

View Article Online
DOI: 10.1039/D6DD00102E

framework enables efficient exploration of chemical space and can be readily extended to diverse molecular discovery tasks.

In the present work, the LLM was guided solely through a system prompt and seed molecules represented as SMILES strings, without access to an external knowledge base. An alternative and potentially complementary strategy is retrieval-augmented generation (RAG), in which the LLM is provided with relevant chemical context retrieved from a structured database at generation time.[71,72] For example, Zhang et al. recently demonstrated a RAG-based framework for solid-state hydrogen storage materials, where a knowledge base of over 30,000 literature-extracted entries was used to inform LLM-driven candidate generation and iterative refinement.[73] Incorporating a similar RAG approach could enrich the chemical context available to the LLM, potentially improving the diversity, novelty, and relevance of generated candidates, and represents a promising direction for future work.

LOOKING AHEAD, SEVERAL DIRECTIONS CAN FURTHER STRENGTHEN AND EXTEND THIS FRAMEWORK. THE WORKFLOW IS INHERENTLY MODULAR AND NOT RESTRICTED TO THE LOHC CANDIDATES PRESENTED HERE. BY ADJUSTING THE SEED MOLECULES, FILTERING CRITERIA, AND ML MODELS TRAINED ON THE RELEVANT TARGET PROPERTY, THE SAME PROTOCOL CAN BE ADAPTED TO DISCOVER FUNCTIONALIZED LOHCS, ALTERNATIVE HYDROGEN CARRIERS, OR MOLECULES FOR ENTIRELY DIFFERENT APPLICATIONS SUCH AS ELECTROLYTES, SOLVENTS, OR ORGANIC SEMICONDUCTORS. FURTHERMORE, THE ITERATIVE DESIGN LOOP IMPLICITLY STEERS THE LLM TOWARD FAVORABLE REGIONS OF CHEMICAL SPACE, AS ACCEPTED MOLECULES FROM ONE CYCLE SERVE AS SEEDS FOR THE NEXT, PROGRESSIVELY NARROWING THE SEARCH TOWARD STRUCTURES THAT SATISFY THE

View Article Online
DOI: 10.1039/D6DD00102E

PHYSICOCHEMICAL CONSTRAINTS. WHILE THIS DOES NOT CONSTITUTE DIRECT MOLECULAR OPTIMIZATION, IT PROVIDES A PRACTICAL MECHANISM FOR DISCOVERING STRUCTURES WITH TARGETED PROPERTY COMBINATIONS, AND THE FILTERING THRESHOLDS FOR PROPERTIES SUCH AS MELTING POINT AND HYDROGEN CAPACITY CAN BE SYSTEMATICALLY TIGHTENED OR RELAXED TO FURTHER GUIDE THIS PROCESS. ADDITIONALLY, INTEGRATING CATALYST PREDICTION INTO THE WORKFLOW WOULD PROVIDE A MORE COMPLETE PICTURE OF THE LOHC DESIGN PROCESS. TRAINING ML MODELS ON CATALYST PERFORMANCE DATA COULD ENABLE SIMULTANEOUS OPTIMIZATION OF BOTH THE CARRIER MOLECULE AND ITS CATALYTIC SYSTEM, BRIDGING MOLECULAR DISCOVERY WITH PROCESS-LEVEL DESIGN. THIS INCLUDES COMPOSITION, METAL–SUPPORT INTERACTIONS, AND REACTION CONDITIONS. FUTURE EFFORTS SHOULD ALSO ACCOUNT FOR TOXICITY, ENVIRONMENTAL IMPACT, AND SCALABILITY OF SYNTHESIS TO ENSURE THAT DISCOVERED CANDIDATES ARE NOT ONLY THERMODYNAMICALLY FAVORABLE BUT ALSO SAFE AND INDUSTRIALLY VIABLE.

ASSOCIATED CONTENT

Supporting Information

Additional detail is provided in the Supporting Information, including an overview of the G4MP2 method, seed sets with calculated and experimental hydrogenation enthalpies, comparisons for down-selected molecules, chemical and physical properties of new LOHCs, LLM and ML performance analyses, design rules, hydrogenation schemes, and synthesis details for Pd/NU-1000. All code used in this study is present on GitHub: https://github.com/HydrogenStorage/LLM_LOHC

View Article Online
DOI: 10.1039/D6DD00102E

AUTHOR INFORMATION

Corresponding Authors

Hassan Harb – *Materials Science Division*

Argonne National Laboratory, Lemont, Illinois 60439, United States, orcid.org/0000-0002-6016-3122

Email: hharb@anl.gov

Rajeev Surendran Assary – *Materials Science Division*

Argonne National Laboratory, Lemont, Illinois 60439, United States, orcid.org/0000-0002-9571-3307

Email: assary@anl.gov

Authors

Magali Ferrandon – *Chemical Sciences and Engineering*

Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; orcid.org/0000-0003-2544-6466

Email: ferrandon@anl.gov

Massimiliano Delferro – *Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, Illinois 60439, United States; orcid.org/0000-0002-4443-165X;*

View Article Online
DOI: 10.1039/D6DD00102E

Email: delferro@anl.gov

Timothy A. Goetjen – *Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States, orcid.org/0000-0001-8023-9107*

Email: tim.goetjen@u.northwestern.edu

Seryeong Lee – *Chemical Sciences and Engineering Division, Argonne National Laboratory, Lemont, IL 60439, United States, Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States, orcid.org/0000-0003-0621-6024*

Email: SeryeongLee2026@u.northwestern.edu

Omar K. Farha – *Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States, orcid.org/0000-0002-9904-9845*

Email: o-farha@northwestern.edu

Author Contributions

H.H. and R.S.A. conceived the project and performed all computational, machine learning, and LLM-based molecular discovery work. T.A.G., S.L., and O.K.F. contributed expertise on catalyst design, synthesis, and mechanistic analysis. M.S.F. and M.D. carried out experimental validation and characterization studies. All authors discussed the results and contributed to the final manuscript.

Funding Sources

U.S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357 (LDRD, LCRC).

View Article Online
DOI: 10.1039/D6DD00102E

Catalyst Design for Decarbonization Center (CD4DC), an Energy Frontier Research Center, under Award No. DE-SC0023383.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, U.S. Department of Energy, under Contract No. DE-AC02-06CH11357. We acknowledge computing resources provided by "BEBOP," a cluster operated by the Laboratory Computing Resource Center (LCRC) at Argonne National Laboratory. Prompting was conducted using Argo, Argonne's internal generative AI chatbot, operated by the Business and Information Services (BIS) division. Experimental work was supported by the Catalyst Design for Decarbonization Center (CD4DC), an Energy Frontier Research Center funded by the DOE Office of Science, Basic Energy Sciences (BES), under Award No. DE-SC0023383. ChatGPT v5.1 was used to assist with editing this manuscript.

Methods

Selection of Seed Molecules: To initiate the LLM-guided molecular generation, selected seed molecules based on the design rules established in previous work[40] were chosen (Figure 5; Text S3; see Supporting Information). Note that this group contains experimentally evaluated LOHC

View Article Online
DOI: 10.1039/D6DD00102E

candidates (Expt-31; known ΔH values) as well as molecules identified from prior computational screening (LOHC-7).[40] The latter consists of the molecules discovered by using high-throughput screening, quantum chemical calculations, and practical down-selection. This dual selection strategy balances experimentally validated structures with computationally curated discoveries, allowing us to assess whether the LLM can generate novel candidates that align with empirical and theoretical insights in LOHC chemistry (Figure 5).

View Article Online
DOI: 10.1039/D6DD00102E

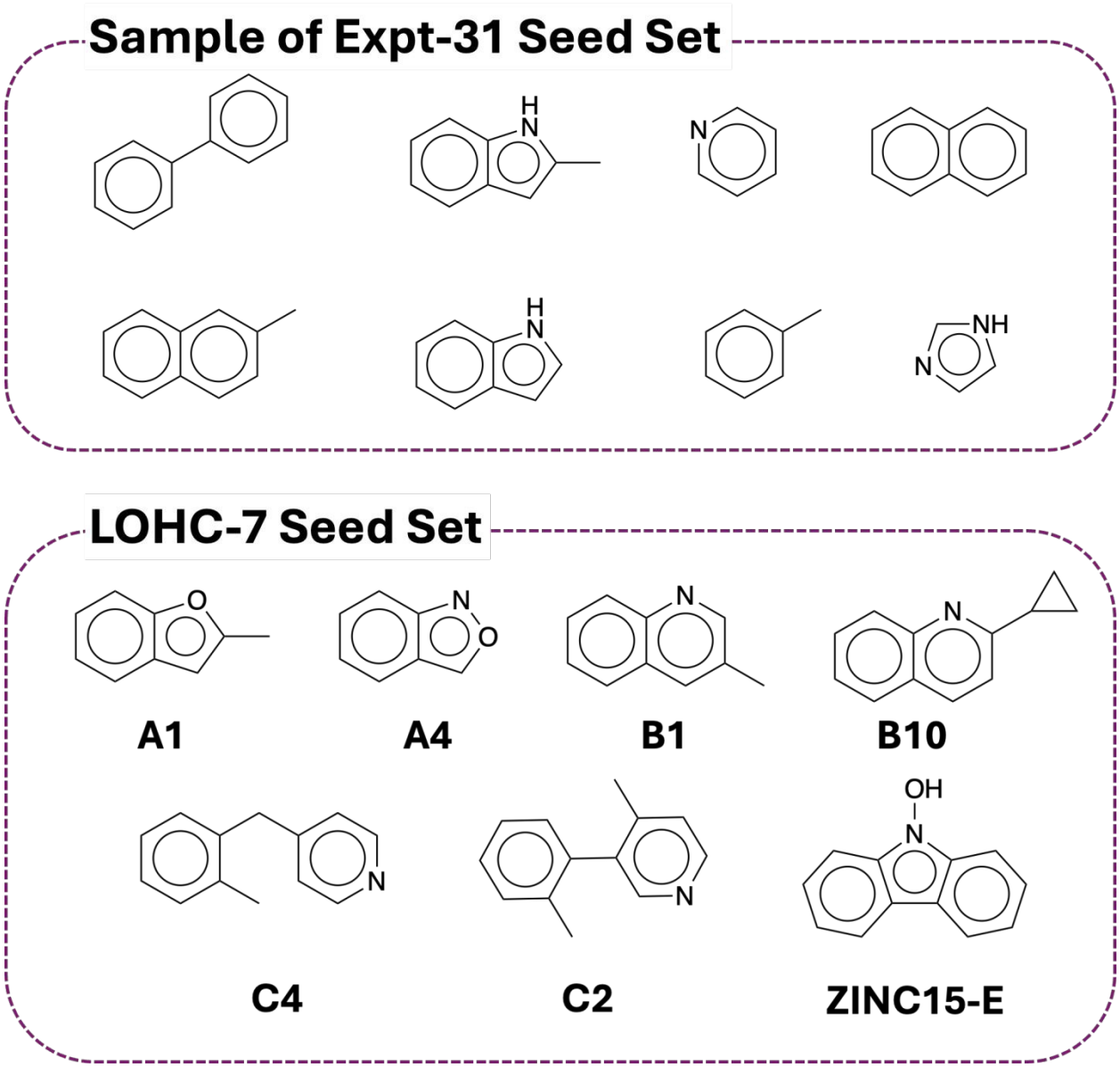


Figure 5. Overview of the seed molecules used to prompt the LLM. The top panel displays eight molecules selected from the Expt-31 dataset, which consists of experimentally known LOHCs. The bottom panel shows seven molecules chosen from previous work on LOHC design.[40]

View Article Online
DOI: 10.1039/D6DD00102E

Machine Learning (ML). As part of the workflow, a ML model capable of predicting hydrogenation enthalpy (ΔH) directly from molecular structures (e.g.: SMILE strings) was developed (Figure 3, orange box). This ML model was trained QM9-LOHC,[60] a high-fidelity computational dataset containing 10,373 dehydrogenation reactions.[60] The dataset includes reactions with hydrogen storage capacities of 5.5 wt % H_2 or higher, and ΔH values computed at the G4MP2[74] level of theory (Fig. 6a). For model training, each molecule was represented using 2048-bit Morgan fingerprints (ECFP4).[75] The dataset was randomly split (80:20), with 80% used for training and 20% reserved for testing. Then RF regressor[76] to predict ΔH was trained using default hyperparameters, and 100 trees were used in the model. The trained model exhibited good predictive accuracy, achieving mean absolute error (MAE) = 4.67 kJ/mol, root mean squared deviation (RMSD) = 7.35 kJ/mol, and coefficient of determination (R^2) = 0.93, with a high correlation between predicted and computed ΔH values. A parity plot shown in Figure 6b, comparing G4MP2 computed vs. ML predicted ΔH confirmed its reliability for screening newly generated LOHC candidates. This ML model was subsequently integrated into the iterative molecular generation process to prioritize molecules within the optimal ΔH range of 40–70 kJ/mol.

View Article Online
DOI: 10.1039/D6DD00102E

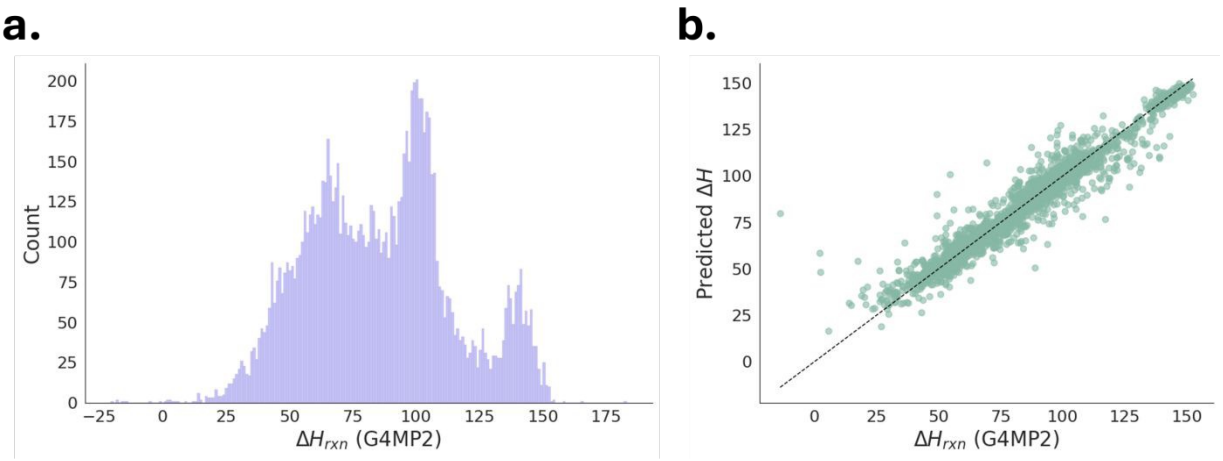


Figure 6. Distribution of (left) computed hydrogenation enthalpies (ΔH, kJ/mol per H_2), and ML predicted ΔH vs. ΔH values using G4MP2 (right). The histogram shows the range of reaction enthalpies (ΔH_{rxn}) in the dataset, while the scatter plot illustrates the predictive accuracy (R^2 = 0.93) of the ML model against G4MP2 values.

LLM Agent. To implement the generative component of the workflow, an LLM-driven molecular design agent capable of proposing new LOHC candidates was established (Figure 3, green box). This agent was built using Argonne's Argo interface to GPT-o1preview model, to systematically generate molecular structures based on provided seed molecules. The LLM agent was constructed with a system prompt that explicitly defines its role, instructing it to act as an expert in molecular design specializing in LOHC. This prompt ensures that the model focuses on functionally relevant molecular motifs rather than arbitrarily generating structures. The core of the agent consists of three steps:

View Article Online
DOI: 10.1039/D6DD00102E

Seeding the Prompt with Known Molecules: The LLM is provided with an initial list of known LOHC SMILES strings, serving as guiding examples to ground the generation process in known chemistry. These molecules establish the structural space in which the LLM operates and generate chemically consistent candidates.

Generating Novel Molecules: The LLM is explicitly prompted to produce 30 SMILE strings, that are unique, chemically valid, and distinct from the provided list. This step ensures that the agent expands molecular diversity while maintaining relevance to LOHC chemistry.

Structured Output Collection: The generated molecules are returned in a JSON format, containing only a structured list of SMILES strings. This was followed by parsing and integration into downstream validation and screening steps. If the LLM deviates from this structure, its response is discarded, and it is re-prompted to ensure compliance with the required format.

The generated molecules serve as inputs for further refinement and validation in subsequent steps of the workflow. The full system prompt is given below:

```
You are an expert in molecular design, specializing in Liquid Organic Hydrogen Carriers (LOHCs).

The user provided these known LOHC SMILES:

{', '.join(initial_smiles)}

Your task is to generate exactly 30 novel LOHC SMILES strings in a structured JSON format:

{{"SMILES": ["SMILES1", "SMILES2", "SMILES3", ..., "SMILES{NEW_LOHC_BATCH_SIZE}"]}}

Ensure that the new SMILES are chemically valid, unique, and not already in the provided list.

Do not include any additional text or explanations. Respond only with the JSON structure.
```

View Article Online
DOI: 10.1039/D6DD00102E

Iterative Generation and Refinement. The LLM-generated molecules go through a multi-stage filtering process to ensure chemical feasibility and suitability as LOHCs (Figure 2, purple box). This iterative workflow was designed to generate and refine molecular candidates during ten iterations. The process continued until either 200 valid SMILES were collected or 10 iterations were completed, whichever came first. To account for variability in the LLM responses, each iteration was given up to three attempts. If no new molecules were generated within an iteration, the LLM was re-prompted to ensure that the search space was sufficiently explored. The first filtering step involved a chemical validity verification, where each generated SMILE string was processed using RDKit to confirm their chemically valid molecular structure. Chemically invalid molecules suggested during this this step was discarded to maintain structural integrity in the dataset. Next, we applied hydrogen storage capacity screening, where the wt % H_2 of each molecule was computed using Equation (1), which quantifies the hydrogen content as a percentage of the molecule's total mass before ($MW_{H\text{-}lean}$) and after full hydrogenation ($MW_{H\text{-}rich}$):

$$\%wt\ H_2 = \frac{MW_{H-rich} - MW_{H-lean}}{MW_{H-rich}} \times 100 \qquad (1)$$

At this stage of screening, only molecules with wt % $H_2 \geq 5.5\%$ were retained. Candidates that passed this threshold were then evaluated for hydrogenation enthalpy (ΔH) using a trained ML model, with only those falling within the desired $40 \leq \Delta H \leq 70$ kJ/mol per H_2 ranges being selected.[36,39] Further, MP predictions were incorporated, as LOHCs must remain in the liquid phase under ambient conditions for ease of handling and transport. Using Leruli's MP prediction platform based on OPERA[77] model, molecules with MP > 40˚C were filtered out. Duplicate removal was

View Article Online
DOI: 10.1039/D6DD00102E

enforced at multiple stages, including after generation, filtering, and before final dataset compilation ensuring that the final LOHC set contained unique and chemically distinct molecules from initial seed set.

At the end of each iteration of the workflow, the molecules that successfully passed filtering were merged with those from previous iterations and used as an updated prompt for the LLM in the subsequent cycle. This process allowed the LLM to iteratively refine and expand the chemical space, generating increasingly viable LOHC candidates. Upon completion of ten iterations, the final dataset was compiled, containing SMILES representations of all validated LOHC candidates, alongside their predicted ΔH values, wt % H_2, and MP. The details of the workflow and generated data are described in the GitHub repository: https://github.com/HydrogenStorage/LLM_LOHC.

Computation of Enthalpies. In order to validate the ΔH predictions using ML, quantum chemical calculations using Gaussian 16 software[78] using accurate G4MP2[74] method were performed. The G4MP2 is a composite method based on the G4 theory that utilizes MP2 perturbation theory to enhance computational efficiency. The minimum energy molecular conformers were first determined using the Universal Force Field (UFF) method in RDKit. The G4MP2 approach relies on geometries optimized at the B3LYP/6-31G(2df,p) level of theory,[79–82] followed by a series of high-level single-point energy calculations. The Zero-point energy (E_{ZPE}) corrections are derived from the B3LYP/6-31G(2df,p) computed vibrational frequencies, scaled by a factor of 0.984 to account for anharmonicity.[74] We confirmed the nature of each located stationary point on the

View Article Online
DOI: 10.1039/D6DD00102E

potential energy surface by the absence of imaginary frequencies. Additional details about the G4MP2 method is described in the supplementary information (Text S1).

Catalytic Hydrogenation of LOHCs.

Four catalysts were tested for the hydrogenation of 3-methyl pyridine and toluene. Two commercial catalysts, 10 wt % Pd/C (Sigma Aldrich), and 5 wt % Rh/Al_2O_3 (Degussa) and two home-made catalysts, 10 wt % Pd/Al_2O_3 and 5 wt % Pd/NU-1000. Pd/Al_2O_3 was prepared using the incipient wetness technique by adding a sufficient amount of a solution containing palladium nitrate solution (Sigma Aldrich) to Al_2O_3 support (Sigma Aldrich) and Pd/NU-100 was synthesized as described elsewhere.[83] Screening of all the catalysts and control (no catalyst) were carried out in 48-well plate using the Screening Pressure Reactor (SPR, Unchained Labs Inc.). 10 mg of catalysts was dispensed in 1/2dr shell vials with 100 uL of either 3-methylpyridne (#5, Figure 5) (≥ 99.5%, Sigma Aldrich) 1-ethylnapthalene (#12) (≥ 95%, Oakwood), 1,1-diphenylethane (#32) (≥ 97%, Ambeed), 9-ethylcarbazole (≥ 97%, Sigma Aldrich), benzofuran (> 99%, Sigma Aldrich), or toluene (≥ 99.5%, Sigma Aldrich) in 1 mL dodecane (≥ 99, Sigma Aldrich). The multiwell plate with vials was then covered with a pinhole graphite gasket and a stainless-steel pinhole plate to ensure gas diffusion but to minimize cross-contamination between the vials. Initially, the SPR was flushed with 500 mL/min N_2 for 15 min at room temperature, at an orbital shaking of 150 rpm and pressurized with H_2 initially to minimize evaporation of the reagents and solvent. The reactor was then heated up slowly (10 ˚C/min ramp rate) to either 150 ˚C or 200 ˚C. Under the given conditions,

View Article Online
DOI: 10.1039/D6DD00102E

the pressure of the reactor reached around 300 psi or 600 psi, respectively. After 12 h, the shaking was stopped, the reactor was cooled down to room temperature and was flushed with 100 mL/min N_2 for 15 min. Aliquots were transferred into filter vials (Whatman Mini-UniPrep Syringeless Filter, 0.2 μm). Aliquots were analysed sequentially by a GC Ultra Gas Chromatograph system equipped with a Tri Plus RSH autosampler, an ISQ MS detector, and a FID (Thermo Scientific). The column used for the MS detector was an Agilent J&W DB-5 column (30 m × 0.25 mm × 0.25μm film thickness) while the column used for the FID was an Agilent J&W DB-5MS column (30 m × 0.25 mm × 0.25μm film thickness). GC data were analysed using the Thermo Xcalibur 2.2 SP1.48 software. The following method was used: a 0.5 μL split injection S5 with a split ratio of 100 run under a constant gas flow of 1 mL/min. The oven temperature profile was as follows: initial temperature = 30 ˚C, hold for 10 minutes, ramp at 20 ˚C/min, final temperature = 250 ˚C. The conversions of the substrates were determined based on the sum of the peaks. For the recycling experiments, the spent catalysts (Rh/Al2O3 and Pt/NU-1000) were washed 3 times with toluene and one time with pentane. After drying the catalysts were re-used.

References

(1) Bhowmik, D.; Zhang, P.; Fox, Z.; Irle, S.; Gounley, J. Enhancing Molecular Design Efficiency: Uniting Language Models and Generative Networks with Genetic Algorithms. *Patterns* **2024**, *5* (4), 100947. https://doi.org/10.1016/j.patter.2024.100947.

(2) Menon, D.; Ranganathan, R. A Generative Approach to Materials Discovery, Design, and Optimization. *ACS Omega* **2022**, *7* (30), 25958–25973. https://doi.org/10.1021/acsomega.2c03264.

View Article Online
DOI: 10.1039/D6DD00102E

(3) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative Models for Molecular Discovery: Recent Advances and Challenges. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), e1608. https://doi.org/10.1002/wcms.1608.

(4) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(5) Zimmermann, Y.; Bazgir, A.; Al-Feghali, A.; Ansari, M.; Bocarsly, J.; Brinson, L. C.; Chiang, Y.; Circi, D.; Chiu, M.-H.; Daelman, N.; Evans, M. L.; Gangan, A. S.; George, J.; Harb, H.; Khalighinejad, G.; Takrim Khan, S.; Klawohn, S.; Lederbauer, M.; Mahjoubi, S.; Mohr, B.; Mohamad Moosavi, S.; Naik, A.; Beste Ozhan, A.; Plessers, D.; Roy, A.; Schöppach, F.; Schwaller, P.; Terboven, C.; Ueltzen, K.; Wu, Y.; Zhu, S.; Janssen, J.; Li, C.; Foster, I.; Blaiszik, B. 32 Examples of LLM Applications in Materials Science and Chemistry: Towards Automation, Assistants, Agents, and Accelerated Scientific Discovery. *Mach. Learn. Sci. Technol.* **2025**, *6* (3), 030701. https://doi.org/10.1088/2632-2153/ae011a.

(6) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine Learning. *Nat. Rev. Chem.* **2020**, *4* (7), 347–358. https://doi.org/10.1038/s41570-020-0189-9.

(7) Dandu, N. K.; Ward, L.; Assary, R. S.; Redfern, P. C.; Curtiss, L. A. Accurate Prediction of Adiabatic Ionization Potentials of Organic Molecules Using Quantum Chemistry Assisted Machine Learning. *J. Phys. Chem. A* **2023**, *127* (28), 5914–5920. https://doi.org/10.1021/acs.jpca.3c00823.

(8) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(9) Rashidi, M. M.; Alhuyi Nazari, M.; Harley, C.; Momoniat, E.; Mahariq, I.; Ali, N. Applications of Machine Learning Methods for Boiling Modeling and Prediction: A Comprehensive Review. *Chem. Thermodyn. Therm. Anal.* **2022**, *8*, 100081. https://doi.org/10.1016/j.ctta.2022.100081.

(10) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121* (16), 9816–9872. https://doi.org/10.1021/acs.chemrev.1c00107.

(11) Sajjan, M.; Li, J.; Selvarajan, R.; Sureshbabu, S. H.; Kale, S. S.; Gupta, R.; Singh, V.; Kais, S. Quantum Machine Learning for Chemistry and Physics. *Chem. Soc. Rev.* **2022**, *51* (15), 6475–6573. https://doi.org/10.1039/D2CS00203E.

(12) Patel, L.; Shukla, T.; Huang, X.; Ussery, D. W.; Wang, S. Machine Learning Methods in Drug Discovery. *Molecules* **2020**, *25* (22), 5277. https://doi.org/10.3390/molecules25225277.

View Article Online
DOI: 10.1039/D6DD00102E

(13) Gaviria, J. F.; Narváez, G.; Guillen, C.; Giraldo, L. F.; Bressan, M. Machine Learning in Photovoltaic Systems: A Review. *Renew. Energy* **2022**, *196*, 298–318. https://doi.org/10.1016/j.renene.2022.06.105.

(14) Yang, W.; Fidelis, T. T.; Sun, W.-H. Machine Learning in Catalysis, From Proposal to Practicing. *ACS Omega* **2020**, *5* (1), 83–88. https://doi.org/10.1021/acsomega.9b03673.

(15) Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

(16) Zimmermann, Y.; Bazgir, A.; Afzal, Z.; Agbere, F.; Ai, Q.; Alampara, N.; Al-Feghali, A.; Ansari, M.; Antypov, D.; Aswad, A.; Bai, J.; Baibakova, V.; Biswajeet, D. D.; Bitzek, E.; Bocarsly, J. D.; Borisova, A.; Bran, A. M.; Brinson, L. C.; Calderon, M. M.; Canalicchio, A.; Chen, V.; Chiang, Y.; Circi, D.; Charmes, B.; Chaudhary, V.; Chen, Z.; Chiu, M.-H.; Clymo, J.; Dabhadkar, K.; Daelman, N.; Datar, A.; Evans, M. L.; Fard, M. G.; Fisicaro, G.; Gangan, A. S.; George, J.; Gonzalez, J. D. C.; Götte, M.; Gupta, A. K.; Harb, H.; Hong, P.; Ibrahim, A.; Ilyas, A.; Imran, A.; Ishimwe, K.; Issa, R.; Jablonka, K. M.; Jones, C.; Josephson, T. R.; Juhasz, G.; Kapoor, S.; Kang, R.; Khalighinejad, G.; Khan, S.; Klawohn, S.; Kuman, S.; Ladines, A. N.; Leang, S.; Lederbauer, M.; Liao, S.-L. M.; Liu, H.; Liu, X.; Lo, S.; Madireddy, S.; Maharana, P. R.; Maheshwari, S.; Mahjoubi, S.; Márquez, J. A.; Mills, R.; Mohanty, T.; Mohr, B.; Moosavi, S. M.; Moßhammer, A.; Naghdi, A. D.; Naik, A.; Narykov, O.; Näsström, H.; Nguyen, X. V.; Ni, X.; O'Connor, D.; Olayiwola, T.; Ottomano, F.; Ozhan, A. B.; Pagel, S.; Parida, C.; Park, J.; Patel, V.; Patyukova, E.; Petersen, M. H.; Pinto, L.; Pizarro, J. M.; Plessers, D.; Pradhan, T.; Pratiush, U.; Puli, C.; Qin, A.; Rajabi, M.; Ricci, F.; Risch, E.; Ríos-García, M.; Roy, A.; Rug, T.; Sayeed, H. M.; Scheidgen, M.; Schilling-Wilhelmi, M.; Schloz, M.; Schöppach, F.; Schumann, J.; Schwaller, P.; Schwarting, M.; Sharlin, S.; Shen, K.; Shi, J.; Si, P.; D'Souza, J.; Sparks, T.; Sudhakar, S.; Talirz, L.; Tang, D.; Taran, O.; Terboven, C.; Tropin, M.; Tsymbal, A.; Ueltzen, K.; Unzueta, P. A.; Vasan, A.; Vinchurkar, T.; Vo, T.; Vogel, G.; Völker, C.; Weinreich, J.; Yang, F.; Zaki, M.; Zhang, C.; Zhang, S.; Zhang, W.; Zhu, R.; Zhu, S.; Janssen, J.; Foster, I.; Blaiszik, B. Reflections from the 2024 Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry. arXiv 2024. https://doi.org/10.48550/ARXIV.2411.15221.

(17) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; Cox, S.; De Jong, W. A.; Evans, M. L.; Gastellu, N.; Genzling, J.; Gil, M. V.; Gupta, A. K.; Hong, Z.; Imran, A.; Kruschwitz, S.; Labarre, A.; Lála, J.; Liu, T.; Ma, S.; Majumdar, S.; Merz, G. W.; Moitessier, N.; Moubarak, E.; Mouriño, B.; Pelkie, B.; Pieler, M.; Ramos, M. C.; Ranković, B.; Rodriques, S. G.; Sanders, J. N.; Schwaller, P.; Schwarting, M.; Shi, J.; Smit, B.; Smith, B. E.; Van Herck, J.; Völker, C.; Ward, L.; Warren, S.; Weiser, B.; Zhang, S.; Zhang, X.; Zia, G. A.; Scourtas, A.; Schmidt, K. J.; Foster, I.; White, A. D.; Blaiszik, B. 14 Examples of How LLMs Can Transform

View Article Online
DOI: 10.1039/D6DD00102E

Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon. *Digit. Discov.* **2023**, *2* (5), 1233–1250. https://doi.org/10.1039/D3DD00113J.

(18) Marvin, G.; Hellen, N.; Jjingo, D.; Nakatumba-Nabende, J. Prompt Engineering in Large Language Models. In *Data Intelligence and Cognitive Informatics*; Jacob, I. J., Piramuthu, S., Falkowski-Gilski, P., Eds.; Algorithms for Intelligent Systems; Springer Nature Singapore: Singapore, 2024; pp 387–402. https://doi.org/10.1007/978-981-99-7962-2_30.

(19) Birhane, A.; Kasirzadeh, A.; Leslie, D.; Wachter, S. Science in the Age of Large Language Models. *Nat. Rev. Phys.* **2023**, *5* (5), 277–280. https://doi.org/10.1038/s42254-023-00581-4.

(20) Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; McHardy, R. Challenges and Applications of Large Language Models. arXiv 2023. https://doi.org/10.48550/ARXIV.2307.10169.

(21) Malusare, A.; Aggarwal, V. Improving Molecule Generation and Drug Discovery With a Knowledge-Enhanced Generative Model. *IEEE Trans. Comput. Biol. Bioinforma.* **2025**, *22* (1), 375–381. https://doi.org/10.1109/TCBB.2024.3477313.

(22) Han, M.; Joung, J. F.; Jeong, M.; Choi, D. H.; Park, S. Generative Deep Learning-Based Efficient Design of Organic Molecules with Tailored Properties. *ACS Cent. Sci.* **2025**, *11* (2), 219–227. https://doi.org/10.1021/acscentsci.4c00656.

(23) Dorna, V.; Subhalingam, D.; Kolluru, K.; Tuli, S.; Singh, M.; Singal, S.; Krishnan, N. M. A.; Ranu, S. TAGMol: Target-Aware Gradient-Guided Molecule Generation. arXiv 2024. https://doi.org/10.48550/ARXIV.2406.01650.

(24) Westermayr, J.; Gilkes, J.; Barrett, R.; Maurer, R. J. High-Throughput Property-Driven Generative Design of Functional Organic Molecules. *Nat. Comput. Sci.* **2023**, *3* (2), 139–148. https://doi.org/10.1038/s43588-022-00391-1.

(25) Kotkondawar, R. R.; Sutar, S. R.; Kiwelekar, A. W.; Kadam, V. J.; Jadhav, S. M. A Generative Framework for Enhancing Drug Target Interaction Prediction in Drug Discovery. *Sci. Rep.* **2025**, *15* (1), 35588. https://doi.org/10.1038/s41598-025-01589-9.

(26) Bhuiyan, F.; Harb, H.; Assary, R.; Vázquez-Mayagoitia, Á. Redox Potential Prediction of Fe(II)/Fe(III) Complexes: A Density Functional Theory and Graph Neural Network Approach. **ChemRxiv 2025, preprint.** https://doi.org/10.26434/chemrxiv-2025-t9kmf.

(27) Chowdhury, A.; Harb, H.; Egele, R.; Alves, C.; Bhuyan, F. H.; Doan, H. A.; Vazquez-Mayagoitia, A.; Assary, R. S.; Balaprakash, P. Automated Learning of GNN Ensembles for Predicting Redox Potentials with Uncertainty. **ChemRxiv 2025, preprint.** https://doi.org/10.26434/chemrxiv-2025-0tq7j-v2.

.

(28) Doan, H. A.; Li, C.; Ward, L.; Zhou, M.; Curtiss, L. A.; S. Assary, R. Accelerating the Evaluation of Crucial Descriptors for Catalyst Screening *via* Message Passing Neural Network. *Digit. Discov.* **2023**, *2* (1), 59–68. https://doi.org/10.1039/D2DD00088A.

(29) Ward, L.; Sivaraman, G.; Pauloski, J. G.; Babuji, Y.; Chard, R.; Dandu, N.; Redfern, P. C.; Assary, R. S.; Chard, K.; Curtiss, L. A.; Thakur, R.; Foster, I. Colmena: Scalable Machine-Learning-Based Steering of Ensemble Simulations for High Performance Computing. In *2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing*

View Article Online
DOI: 10.1039/D6DD00102E

Environments (MLHPC); IEEE: St. Louis, MO, USA, 2021; pp 9–20. https://doi.org/10.1109/MLHPC54614.2021.00007.

(30) Doan, H. A.; Agarwal, G.; Qian, H.; Counihan, M. J.; Rodríguez-López, J.; Moore, J. S.; Assary, R. S. Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chem. Mater.* **2020**, *32* (15), 6338–6346. https://doi.org/10.1021/acs.chemmater.0c00768.

(31) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. arXiv 2013. https://doi.org/10.48550/ARXIV.1312.6114.

(32) Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv 2018. https://doi.org/10.48550/ARXIV.1809.11096.

(33) Zang, C.; Wang, F. MoFlow: An Invertible Flow Model for Generating Molecular Graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: Virtual Event CA USA, 2020; pp 617–626. https://doi.org/10.1145/3394486.3403104.

(34) Loeffler, H. H.; Wan, S.; Klähn, M.; Bhati, A. P.; Coveney, P. V. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *J. Chem. Theory Comput.* **2024**, acs.jctc.4c00576. https://doi.org/10.1021/acs.jctc.4c00576.

(35) Yan, X.; Hudson, N.; Park, H.; Grzenda, D.; Pauloski, J. G.; Schwarting, M.; Pan, H.; Harb, H.; Foreman, S.; Knight, C.; Gibbs, T.; Chard, K.; Chaudhuri, S.; Tajkhorshid, E.; Foster, I.; Moosavi, M.; Ward, L.; Huerta, E. A. MOFA: Discovering Materials for Carbon Capture with a GenAI- and Simulation-Based Workflow. arXiv 2025. https://doi.org/10.48550/ARXIV.2501.10651.

(36) Preuster, P.; Papp, C.; Wasserscheid, P. Liquid Organic Hydrogen Carriers (LOHCs): Toward a Hydrogen-Free Hydrogen Economy. *Acc. Chem. Res.* **2017**, *50* (1), 74–85. https://doi.org/10.1021/acs.accounts.6b00474.

(37) 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.

(38) Chu, C.; Wu, K.; Luo, B.; Cao, Q.; Zhang, H. Hydrogen Storage by Liquid Organic Hydrogen Carriers: Catalyst, Renewable Carrier, and Technology - A Review. *Carbon Resour. Convers.* **2023**, S2588913323000248. https://doi.org/10.1016/j.crcon.2023.03.007.

(39) 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.

(40) 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, 2 (5), 1233–1250.** https://doi.org/10.1039/D3DD00123G.

(41) Vishwakarma, G.; Hachmann, J. *Liquid Organic Hydrogen Carriers: High-Throughput Screening of Homogeneous Catalysts*; **ChemRxiv 2023, preprint.** https://doi.org/10.26434/chemrxiv-2023-s8pkf;..

View Article Online
DOI: 10.1039/D6DD00102E

(42) Paragian, K.; Li, B.; Massino, M.; Rangarajan, S. A Computational Workflow to Discover Novel Liquid Organic Hydrogen Carriers and Their Dehydrogenation Routes. *Mol. Syst. Des. Eng.* **2020**, *5* (10), 1658–1670. https://doi.org/10.1039/D0ME00105H.

(43) Modisha, P.; Bessarabov, D. Aromatic Liquid Organic Hydrogen Carriers for Hydrogen Storage and Release. *Curr. Opin. Green Sustain. Chem.* **2023**, *42*, 100820. https://doi.org/10.1016/j.cogsc.2023.100820.

(44) Teichmann, D.; Stark, K.; Müller, K.; Zöttl, G.; Wasserscheid, P.; Arlt, W. Energy Storage in Residential and Commercial Buildings via Liquid Organic Hydrogen Carriers (LOHC). *Energy Environ. Sci.* **2012**, *5* (10), 9044. https://doi.org/10.1039/c2ee22070a.

(45) Valentini, F.; Marrocchi, A.; Vaccaro, L. Liquid Organic Hydrogen Carriers (LOHCs) as H-Source for Bio-Derived Fuels and Additives Production. *Adv. Energy Mater.* **2022**, *12* (13), 2103362. https://doi.org/10.1002/aenm.202103362.

(46) Niermann, M.; Drünert, S.; Kaltschmitt, M.; Bonhoff, K. Liquid Organic Hydrogen Carriers (LOHCs) – Techno-Economic Analysis of LOHCs in a Defined Process Chain. *Energy Environ. Sci.* **2019**, *12* (1), 290–307. https://doi.org/10.1039/C8EE02700E.

(47) Niermann, M.; Timmerberg, S.; Drünert, S.; Kaltschmitt, M. Liquid Organic Hydrogen Carriers and Alternatives for International Transport of Renewable Hydrogen. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110171. https://doi.org/10.1016/j.rser.2020.110171.

(48) 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.

(49) Mulky, L.; Srivastava, S.; Lakshmi, T.; Sandadi, E. R.; Gour, S.; Thomas, N. A.; Shanmuga Priya, S.; Sudhakar, K. An Overview of Hydrogen Storage Technologies – Key Challenges and Opportunities. *Mater. Chem. Phys.* **2024**, *325*, 129710. https://doi.org/10.1016/j.matchemphys.2024.129710.

(50) Mori, D.; Hirose, K. Recent Challenges of Hydrogen Storage Technologies for Fuel Cell Vehicles. *Int. J. Hydrog. Energy* **2009**, *34* (10), 4569–4574. https://doi.org/10.1016/j.ijhydene.2008.07.115.

(51) Brodt, M.; Müller, K.; Kerres, J.; Katsounaros, I.; Mayrhofer, K.; Preuster, P.; Wasserscheid, P.; Thiele, S. The 2-Propanol Fuel Cell: A Review from the Perspective of a Hydrogen Energy Economy. *Energy Technol.* **2021**, *9* (9), 2100164. https://doi.org/10.1002/ente.202100164.

(52) Cho, J.-Y.; Kim, H.; Oh, J.-E.; Park, B. Y. Recent Advances in Homogeneous/Heterogeneous Catalytic Hydrogenation and Dehydrogenation for Potential Liquid Organic Hydrogen Carrier (LOHC) Systems. *Catalysts* **2021**, *11* (12), 1497. https://doi.org/10.3390/catal11121497.

(53) Modisha, P. M.; Ouma, C. N. M.; Garidzirai, R.; Wasserscheid, P.; Bessarabov, D. The Prospect of Hydrogen Storage Using Liquid Organic Hydrogen Carriers. *Energy Fuels* **2019**, *33* (4), 2778–2796. https://doi.org/10.1021/acs.energyfuels.9b00296.

(54) Wei, D.; Shi, X.; Qu, R.; Junge, K.; Junge, H.; Beller, M. Toward a Hydrogen Economy: Development of Heterogeneous Catalysts for Chemical Hydrogen Storage and Release

View Article Online
DOI: 10.1039/D6DD00102E

Reactions. *ACS Energy Lett.* **2022**, *7* (10), 3734–3752. https://doi.org/10.1021/acsenergylett.2c01850.

(55) Dean, D.; Davis, B.; Jessop, P. G. The Effect of Temperature, Catalyst and Sterics on the Rate of N-Heterocycledehydrogenation for Hydrogenstorage. *New J. Chem.* **2011**, *35* (2), 417–422. https://doi.org/10.1039/C0NJ00511H.

(56) Markiewicz, M.; Zhang, Y.-Q.; Empl, M. T.; Lykaki, M.; Thöming, J.; Steinberg, P.; Stolte, S. Hazard Assessment of Quinaldine-, Alkylcarbazole-, Benzene- and Toluene-Based Liquid Organic Hydrogen Carrier (LOHCs) Systems. *Energy Environ. Sci.* **2019**, *12* (1), 366–383. https://doi.org/10.1039/C8EE01696H.

(57) Markiewicz, M.; Zhang, Y. Q.; Bösmann, A.; Brückner, N.; Thöming, J.; Wasserscheid, P.; Stolte, S. Environmental and Health Impact Assessment of Liquid Organic Hydrogen Carrier (LOHC) Systems – Challenges and Preliminary Results. *Energy Environ. Sci.* **2015**, *8* (3), 1035–1045. https://doi.org/10.1039/C4EE03528C.

(58) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(59) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52* (11), 2864–2875. https://doi.org/10.1021/ci300415d.

(60) Harb, H.; Elliott, S. N.; Ward, L.; Foster, I. T.; Klippenstein, S. J.; Curtiss, L. A.; Assary, R. S. Accurate Dehydrogenation Enthalpies Dataset for Liquid Organic Hydrogen Carriers. *Sci. Data* **2025**, *12* (1), 171. https://doi.org/10.1038/s41597-025-04468-0.

(61) PubChem. *PubChem*. https://pubchem.ncbi.nlm.nih.gov/ (accessed 2023-10-09).

(62) Rangarajan, S.; Kaminski, T.; Van Wyk, E.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Algorithms of RING. *Comput. Chem. Eng.* **2014**, *64*, 124–137. https://doi.org/10.1016/j.compchemeng.2014.02.007.

(63) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA Models for Predicting Physicochemical Properties and Environmental Fate Endpoints. *J. Cheminform.* **2018**, *10* (1), 10. https://doi.org/10.1186/s13321-018-0263-1.

(64) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121* (16), 10001–10036. https://doi.org/10.1021/acs.chemrev.0c01303.

(65) Lee, A. S.; Elliott, S.; Harb, H.; Ward, L.; Foster, I.; Curtiss, L.; Assary, R. S. E_{min} : A First-Principles Thermochemical Descriptor for Predicting Molecular Synthesizability. *J. Chem. Inf. Model.* **2024**, acs.jcim.3c01583. https://doi.org/10.1021/acs.jcim.3c01583.

(66) Karakhanov, **?**. A.; Viktorova, E. A. Hydrogenation and Dehydrogenation Reactions of Benzofuran and Its Derivatives (Review). *Chem. Heterocycl. Compd.* **1976**, *12* (4), 367–375. https://doi.org/10.1007/BF00480416.

(67) Teixeira, I. F.; Lo, B. T. W.; Kostetskyy, P.; Ye, L.; Tang, C. C.; Mpourmpakis, G.; Tsang, S. C. E. Direct Catalytic Conversion of Biomass-Derived Furan and Ethanol to Ethylbenzene. *ACS Catal.* **2018**, *8* (3), 1843–1850. https://doi.org/10.1021/acscatal.7b03952.

(68) Eblagon, K. M.; Rentsch, D.; Friedrichs, O.; Remhof, A.; Zuettel, A.; Ramirez-Cuesta, A. J.; Tsang, S. C. Hydrogenation of 9-Ethylcarbazole as a Prototype of a Liquid Hydrogen Carrier.

View Article Online
DOI: 10.1039/D6DD00102E

Int. J. Hydrog. Energy **2010**, *35* (20), 11609–11621. https://doi.org/10.1016/j.ijhydene.2010.03.068.

(69) Sotoodeh, F.; Smith, K. J. Kinetics of Hydrogen Uptake and Release from Heteroaromatic Compounds for Hydrogen Storage. *Ind. Eng. Chem. Res.* **2010**, *49* (3), 1018–1026. https://doi.org/10.1021/ie9007002.

(70) Ramadhani, S.; Dao, Q. N.; Imanuel, Y.; Ridwan, M.; Sohn, H.; Jeong, H.; Kim, K.; Yoon, C. W.; Song, K. H.; Kim, Y. Advances in Catalytic Hydrogenation of Liquid Organic Hydrogen Carriers (LOHCs) Using High-Purity and Low-Purity Hydrogen. *ChemCatChem* **2024**, *16* (24), e202401278. https://doi.org/10.1002/cctc.202401278.

(71) Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv April 12, 2021. https://doi.org/10.48550/arXiv.2005.11401.

(72) Feng, Y.; Wang, J.; He, R.; Zhou, L.; Li, Y. A Retrieval-Augmented Knowledge Mining Method with Deep Thinking LLMs for Biomedical Research and Clinical Support. arXiv March 29, 2025. https://doi.org/10.48550/arXiv.2503.23029.

(73) Zhang, D.; Jia, X.; Tran, H. B.; Jang, S. H.; Zhang, L.; Sato, R.; Hashimoto, Y.; Sato, T.; Konno, K.; Orimo, S.; Li, H. "DIVE" into Hydrogen Storage Materials Discovery with AI Agents. *Chem. Sci.* **2026**, *17* (6), 3031–3042. https://doi.org/10.1039/D5SC09921H.

(74) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory Using Reduced Order Perturbation Theory. *J. Chem. Phys.* **2007**, *127* (12), 124105. https://doi.org/10.1063/1.2770701.

(75) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107–113. https://doi.org/10.1021/c160017a018.

(76) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python.

(77) *Leruli*. https://www.leruli.com/ (accessed 2023-03-13).

(78) 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, Gaussian, Inc., Wallingford CT. **2016**.

View Article Online
DOI: 10.1039/D6DD00102E

(79) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652. https://doi.org/10.1063/1.464913.

(80) Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent Molecular Orbital Methods 25. Supplementary Functions for Gaussian Basis Sets. *J. Chem. Phys.* **1984**, *80* (7), 3265–3269. https://doi.org/10.1063/1.447079.

(81) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent Molecular Orbital Methods. XX. A Basis Set for Correlated Wave Functions. *J. Chem. Phys.* **1980**, *72* (1), 650–654. https://doi.org/10.1063/1.438955.

(82) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100. https://doi.org/10.1103/PhysRevA.38.3098.

(83) McCullough, K. E.; King, D. S.; Chheda, S. P.; Ferrandon, M. S.; Goetjen, T. A.; Syed, Z. H.; Graham, T. R.; Washton, N. M.; Farha, O. K.; Gagliardi, L.; Delferro, M. High-Throughput Experimentation, Theoretical Modeling, and Human Intuition: Lessons Learned in Metal–Organic-Framework-Supported Catalyst Design. *ACS Cent. Sci.* **2023**, *9* (2), 266–276. https://doi.org/10.1021/acscentsci.2c01422.

View Article Online
DOI: 10.1039/D6DD00102E

Data Availability Statement

All data associated with this study are publicly available. The repository includes SMILES strings of the starting materials, the trained Random Forest model, workflow scripts, and representative output files. The code and data can be accessed at https://github.com/HassanHarb92/LLM_LOHC and are archived at https://doi.org/10.5281/zenodo.18853735.