ChemComm

Design of CO2-philic Molecular Units with Large Language Models

Journal:	*ChemComm*
Manuscript ID	CC-COM-05-2025-002652.R2
Article Type:	Communication

ChemComm

COMMUNICATION

Design of CO_2-philic Molecular Units with Large Language Models

Konstantinos D. Vogiatzis*[a]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

We demonstrate the use of large language models (LLMs) for designing molecules with high CO_2 affinity. By integrating LLM-generated candidates with DFT-based evaluation, we identified promising physisorption agents, showcasing emergent molecular strategies and highlighting the synergy between AI and expert-guided chemical research.

The design of molecular units that selectively interact with CO_2 via noncovalent interactions is central for the development of physiosorbed-based CO_2 separation[1] and direct air capture technologies.[2] Such technologies involve (semi)crystalline materials such as porous metal-organic frameworks (MOFs)[3] where CO_2-philic groups selectively bind CO_2, molecular sieves such as or polymers of intrinsic microporosity (PIMs)[4] or passive polymeric membranes where favorable noncovalent interactions enhance the diffusion of the CO_2 molecules through the membrane.[5,6] Computational methods can accelerate the design of CO_2-affine molecules, from small-scale studies to large-scale chemical space exploration using AI, large language models (LLMs), and other data-driven workflows.[7]

In recent years, LLMs have transformed many aspects of education, work, research, and daily life.[8-10] LLMs are deep learning models trained on massive datasets, capable of performing diverse natural language processing (NLP) tasks such as question answering, text classification, translation, and code generation. They operate through user-provided prompts, and the quality of their output depends on the prompt's clarity and specificity, a concept known as prompt engineering. Pretrained LLMs can be fine-tuned through strategies like labeled datasets, interactive question/answering, or text classification. Iterative prompting and the inclusion of relevant databases can further refine their responses and adaptability.[11]

The field of chemical sciences is also undergoing significant transformations with the integration of LLMs.[12] These applications range from the use of pre-trained models as chatbots[13] to fine-tuned models designed for specific chemical tasks,[14-17] molecular design,[18] or broader chemical applications.[19-20] In all cases, the expertise of the chemist remains essential, as domain knowledge is required for the fine-tuning of the LLM, ultimately enhancing their accuracy, applicability, and overall utility.

Here, we present a hybrid approach where pretrained LLM models were tested on a specific chemical question, and they were further tuned by the user to adapt to the problem and examine unexplored areas of the chemical space. We have tested three popular free LLMs (GPT-4o via ChatGPT,[21] Llama-3.2[22] via Ollama, Gemini 2.0[23]) for the design of molecular units with enhanced affinity for CO_2 that can aid carbon capture via physisorption. We are explaining the role of domain expertise by keeping the "chemist in the loop" and highlighting the cooperativity between user and model for enhanced performance. Our computational workflow is shown on Figure 1 which includes the introduction of the chemical question under consideration with a short literature overview in natural language (prompt), the AI-generated molecular units with potential high CO_2-philicity in SMILES format, automated conformational search with the conformer–rotamer sampling tool (CREST),[24] accurate density functional theory (DFT) calculations for the computation of interaction energies (ΔE_{Int})

[a] *Department of Chemistry, University of Tennessee, 37996 Knoxville, Tennessee, United States.*

Supplementary Information available: prompts and full responses by the LLMs with dates; computational details; list of molecular units and computed interaction energies; additional analysis of best performers. See DOI: 10.1039/x0xx00000x

between each molecule and CO_2, and the further exploration of additional molecular units based on user's expertise.

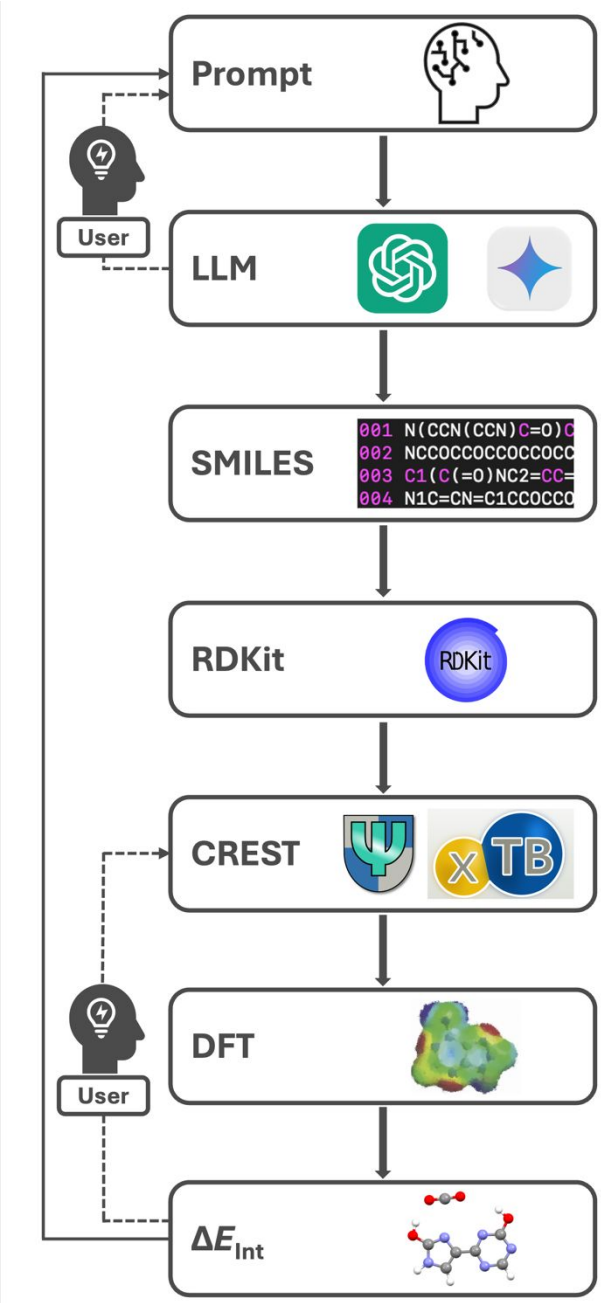

Figure 1. Computational workflow applied in this study. From top to bottom: the user provides a prompt to the LLM and new molecular units with potential strong affinity for CO_2 binding are generated in SMILES format, which are converted to Cartesian coordinates with RDKit. Conformational analysis for both molecule and molecule-CO_2 supersystem is performed with CREST, the energetically most stable conformers are optimized with DFT, and interaction energies ΔE_{int} are computed. Based on results, new molecular units are generated via new prompts to the LLM, or via user's experience.

First, we set an optimization parameter as a specific target for the LLM model. We have selected as target the interaction energy between CO_2 and the molecular unit at -10.0 kcal/mol, which was recommended by Jiang *et al.* as the threshold for optimum physiosorbed-based CO_2 separation,[7] while our pragmatic target was to surpass the threshold of -7.5 kcal/mol.

We then reviewed the literature for representative CO_2-philic molecules, which can be organized into two groups. The first group is single unit interactions (one unit with one CO_2, or 1:1 ratio) and include oxygen-containing molecules (-3.2 kcal/mol),[25] carboxylic groups (-3.8 up to -6.2 kcal/mol),[26,27] and nitrogen-containing heterocyclic compounds which can go up to -7.0 kcal/mol.[28-31] The second group includes cooperative effects between two molecular units and CO_2 (2:1 ratio), such as ethylene oxide oligomers[32] or dipeptides,[33] where the interaction with CO_2 can exceed the -9.1 kcal/mol. Here, we focused on 1:1 capture since it can form the basis of cooperative binding (2:1) through further molecular engineering.

A short paragraph was used as prompt which described the chemical problem that we would like to explore with the help of LLMs. The paragraph introduced the background of the problem (physisorption), the role of CO_2-philic groups, representative examples from literature (molecules provided as SMILES), optimization target (interaction energy higher than -7.5 kcal/mol), as well as two specific questions:

1. *Can you compile additional data from the literature with similar CO_2-philic groups. You can either provide their SMILES string or their molecular formula. Listing the corresponding references would be much appreciated.*

2. *Based on the data that you have found and read, can you predict molecules with CO_2 interaction that is lower than -7.5 kcal/mol and closer to -10 kcal/mol?*

First response from GPT-4o was satisfactory since it provided additional molecular units not introduced by the user together with literature references, such as ethylene oxide oligomers,[32] acylamide groups,[34] amines for carbamide formation (which was inaccurate since it redirected to chemisorption as well as it provided wrong reference[35]), and carbonyl and ether groups. Additional questions from the user lead to the recommendation of different sets of known molecules with potential CO_2-philicity, as well as exploration of more exotic molecular units. The last prompt provided by the user contained a list with data collected up to that point (introduced as a pdf file), to further tune the LLM. Overall, an extended dataset of 89 molecules was generated by GPT-4o. It is remarkable that some hypothetical molecules generated by GPT-4o combined a heterocycle with an extended ethylene oxide oligomer chain or two heterocycles with reported CO_2 affinity, which demonstrate its effort to merge known CO_2-philic groups for the discovery of new molecular units with stronger interaction energies.

The response from Llama-3 to the first prompt was unsatisfactory. Proposed CO_2-philic groups from literature were either included in the prompt, or they did not have a clear site for strong CO_2 interaction (e.g. dicyclohexylurea). The five proposed references were redirecting to nonexistent publications, while the three SMILES that were provided had syntax errors. For all these reasons, we decided not to explore further Llama-3.

Gemini 2.0 responses were the most creative since it suggested more complex molecules than those included in our initial prompt. For example, it provided SMILES strings of molecular structures with azacyclobutanes or with nitrogen-containing norbornene units (2-azabicyclo(2.2.1)heptane). Overall, 8 molecules suggested by Gemini 2.0 were added in our database.

A dataset of 97 molecules in SMILES format was compiled from both GPT-4o and Gemini 2.0, which were further converted into Cartesian coordinates with the RDKit package.[36] A detailed analysis of all systems proposed by LLMs will be avoided, and we will only focus on the most interesting molecular cases with enhanced CO_2-philicity. We refer the reader to the Supplementary Information for more details. Overall, five molecules (**05**, **27**, **31**, **79**, **99**, Figure 2) showed stronger CO_2 affinity with interaction energies exceeding the -7 kcal/mol. Structure **05** has the strongest CO_2 interaction (-8.15 kcal/mol) from all molecules examined in this study. CO_2 is attracted by this molecular unit via cooperative interactions from four carboxylic units, where two of them are acting as the primary CO_2 binding sites (distance between oxygens of **05** and carbon of CO_2 are 2.763 Å and 2.915Å, respectively). Structure **27** has interaction energy of -7.04 kcal/mol, where CO_2 is stabilized by two imidazotriazine units and a carboxylic unit. However, the accessibility of CO_2 to the two nitrogen atoms, the primary CO_2-philic site, is partially hindered by the presence of a carboxylic acid group. More interestingly, structure **22** has two unblocked nitrogen atoms of imidazotriazine units that can cooperatively bind CO_2 with a total interaction strength of -6.44 kcal/mol. We observed that one of the two imidazotriazine units has a hydroxyl group that further stabilizes the CO_2 molecule through a weak hydrogen bond.[35] We manually modified structure **22** by substituting a hydrogen atom with an amino (structure **22-a**) or a hydroxo group (structure **22-b**, Figure 2). These modifications were based on the user's intuition and increased further the binding strength of the parent structure **22** to -7.80 and -7.91 kcal/mol, respectively. In particular, the distances between the two nitrogen atoms of **22-b** and the carbon of CO_2 are 2.873Å and 2.965Å, respectively, while the hydroxo-CO_2 distance is only 2.003Å. Similarly, structure **31** (-7.21 kcal/mol, CO_2 binding on five- and six-member nitrogen heterostructures, further enhanced by two neighboring hydroxyl groups) was modified based on user's expertise (structures **31-a** to **31-e**). These modifications resulted to an increased interaction energy that reached up to -7.42 kcal/mol for structure **31-d**. Structure **79** has also two nitrogen sites (distance with carbon of CO_2 of 2.871Å and 2.982Å, respectively) and a hydroxyl group in close proximity (2.136Å), leading to a total interaction energy of -8.04 kcal/mol. Finally, structure **94** that was recommended by Gemini 2.0 also demonstrate strong CO_2-philicity with an interaction energy of -7.24 kcal/mol. In this case, a rigid molecular unit with multiple CO_2-philic sites provides cooperative binding.

Additional computational analysis was performed for the 13 top performers (interaction energies with CO_2 lower than -7 kcal/mol plus structure **22**). First, the interaction energies of these molecular units with N_2 ($\Delta E_{Int(N2)}$) and CH_4 ($\Delta E_{Int(CH4)}$) were calculated and compared with those for CO_2 ($\Delta E_{Int(CO2)}$, see Supplementary Information). The ratios $\Delta E_{Int(CO2)}/\Delta E_{Int(N2)}$ and $\Delta E_{Int(CO2)}/\Delta E_{Int(CH4)}$ can be used as an estimation of their potential performance for CO_2/N_2 and CO_2/CH_4 separations. Overall, we found that these ratios are within the 2.0 – 5.2 range for CO_2/N_2, and within the 1.9 – 4.7 range for CO_2/CH_4. It is remarkable that the highest ratios for both CO_2/N_2 and CO_2/CH_4 were obtained for structure **22-b**, where two hydroxyl groups reduce N_2 and CH_4 attraction. On the contrary, the amine group of structure **22**-a was increasing the interaction with both N_2 and CH_4. Second, we have considered two synthesizability scores (synthetic accessibility score,[37] natural product-likeness score[38]), as indicators of the chemical synthesis feasibility of these molecules. Overall, these scores are within reasonable limits apart from structure **94**, the most exotic molecule out of all top performers that was suggested by Gemini 2.0. Finally, we assessed the novelty of the generated molecules by checking their presence in public molecular repositories. A search of the PubChem database confirmed that these structures have not previously reported, indicating that the LLMs functioned not as retrieval tools, but as user-driven generative engines for novel molecular candidates.

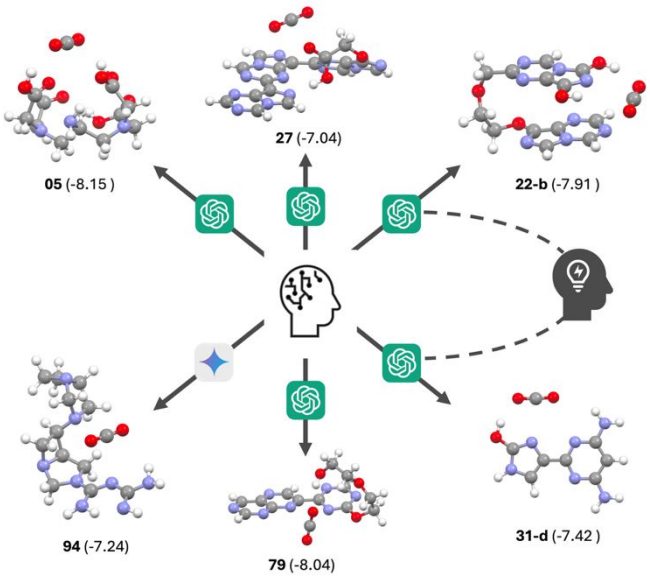

Figure 2. DFT optimized geometries of representative examples from the top molecular performers with enhanced CO_2-affinity, as recommended by GPT-4o (green app logo), Gemini 2.0 (white app logo), or by the user, based on the GPT-4o recommendations (grey human head). Numerical labels of molecules, as well as DFT interaction energies (shown in parenthesis, in kcal/mol) between these five units and CO_2 are shown under each molecular geometry.

In summary, we have explored the capabilities of different popular LLMs for molecular design. Our target was the discovery of molecular units that bind CO_2 in a 1:1 ratio with interaction energies that exceed previously reported values for potential application in physisorbed-based carbon capture. We have applied a limited tuning on pretrained models via chat options, which helped us to identify five molecular units with CO_2 interaction energy less than -7.0 kcal/mol, while further structural engineering increased the number of molecular hits

to 12. It is noteworthy that this human/AI experience revealed new binding motifs with enhanced CO_2-philicity. For example, structure **22-b** stabilizes a CO_2 molecule cooperatively through two nitrogen-based heterocycles that favorably interact with the carbon of CO_2, and two hydroxyl group that further stabilize the superstructure via weak hydrogen bonds, with a total ΔE_{Int} of -7.91 kcal/mol. Structure **22-b** has also large $\Delta E_{Int(CO2)}/\Delta E_{Int(N2)}$ and $\Delta E_{Int(CO2)}/\Delta E_{Int(CH4)}$ ratios and good synthesizability scores, which highlight its potential for practical applications.

To conclude, this study highlights how LLMs can significantly enhance a domain expert's creativity and extend chemical intuition toward novel discoveries. Relevant work has shown that when integrated with automated data extraction tools that convert scientific literature into machine-readable formats, LLMs offer vast opportunities for accelerating chemical research. However, despite their ease of use, these models face limitations akin to conventional chemical machine learning approaches, including dependencies on data quantity, source reliability, and the way chemical information is encoded and processed within AI/ML workflows,[39] an active topic of research. While LLMs hold great promise for the chemical sciences, their adoption must be accompanied by careful validation and an awareness of their inherent constraints. Additionally, a major challenge lies in the reproducibility of LLM-generated responses, as outputs are influenced by model versioning, the scale and quality of pretraining data, and user-specific fine-tuning. Alternatively, LLM consistency and reliability can be improved by aggregating output ensembles, providing valuable guidance for molecular design and discovery. Despite current limitations, LLMs already offer a powerful platform for amplifying human creativity and accelerating scientific discovery. As demonstrated in this study, their integration into molecular design workflows can unlock new avenues of chemical insight and innovation.

This material is based upon work supported by the National Science Foundation under grant no. 2143354 (CAREER: CAS-Climate). K. D. V. is grateful to Prof. Markus Reiher for his accommodation at ETH Zürich, Switzerland. K. D. V. acknowledges the Collegium Helveticum International Fellowship Program that partially supports his Sabbatical visit at ETH Zürich, Switzerland, and the Infrastructure for Scientific Applications and Advanced Computing (ISAAC) of the University of Tennessee for computational resources.

Data Availability Statement

All data supporting this article have been included as part of the Electronic Supplementary Information (exact prompts and responses from the LLMs, computational details for reproducing all results reported in the manuscript, list of molecular units and computed interaction energies).

Conflicts of interest

There are no conflicts to declare.

References

1 H. McLaughlin, A. A. Littlefield, M. Menefee, A. Kinzer, T. Hull, B. K. Sovacool, M. D. Bazilian, J. Kim, S. Griffiths, *Renew. Sustain. Energy Rev.*, 2023, **177**, 113215.
2 Y. Abdullatif, A. Sodiq, N. Mir, Y. Bicer, T. Al-Ansari, M. H. El-Naas, A. I. Amhamed, *RSC Adv.*, 2023, **13**, 5687.
3 S. Ali Akbar Razavi, A. Morsali, *Coord. Chem. Rev.*, 2019, **399**, 213023.
4 N. Du, H. B. Park, G. P. Robertson, M. M. Dal-Cin, T. Visser, L. Scoles, M. D. Guiver, *Nat. Mater.*, 2011, **10**, 372.
5 M. Galizia, W. S. Chi, Z. P. Smith, T. C. Merkel, R. W. Baker, B. D. Freeman, *Macromolecules*, 2017, **50**, 7809.
6 W. J. Koros, C. Zhang, *Nat. Mater.*, 2017, **16**, 289.
7 Z. Tian, S. Dai, D. e. Jiang, *WIREs Comput. Mol. Sci.*, 2016, **6**, 173.
8 B. Fecher, M. Hebing, M. Laufer, J. Pohle, F. Sofsky, *AI & Society*, 2023, **40**, 447.
9 S. Dhakal, H. Parry, *Nature*, 2024, **636**, 299.
10 M. Binz et al., *Proc. Natl. Acad. Sci.*, 2025, **122**, e2401227121.
11 H. Naveed et al., arXiv preprint, 2024, DOI: 10.48550/arXiv.2307.06435.
12 M. C. Ramos, C. J. Collison, A. D. White, *Chem. Sci.*, 2025, **16**, 2514.
13 K. M. Jablonka et al., *Digit. Discov.*, 2023, **2**, 1233.
14 D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, *Nature*, 2023, **624**, 570.
15 M. Moret et al., *Nat. Commun.*, 2023, **14**, 114.
16 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, P. Schwaller, *Nat. Mach. Intell.*, 2024, *6*, 525.
17 S. Kim, Y. Jung, J. Schrier, *J. Am. Chem. Soc.*, 2024, *146*, 19654.
18 D. Bhattacharya, H. J. Cassady, M. A. Hickner, W. F. Reinhart, *J. Chem. Inf. Model.*, 2024, **64**, 7086.
19 J. Born, M. Manica, *Nat. Mach. Intell.*, 2023, **5**, 432.
20 J. Van Herck et al., *Chem. Sci.*, 2025, **16**, 670.
21 OpenAI, ChatGPT, 2024.
22 Meta AI, Llama-3.2, 2024.
23 Google, Gemini AI 2.0, 2025.
24 P. Pracht et al., *J. Chem. Phys.*, 2024, **160**, 114110.
25 D. Zhao, X. Liu, *ACS Omega*, 2022, **7**, 17330.
26 R. Selyanchyn, A. Staykov, S. Fujikawa, *RSC Adv.*, 2016, **6**, 88664.
27 A. G. Sylvanus, K. D. Vogiatzis, *ChemPhysChem*, 2023, **24**, e202300027.
28 K. D. Vogiatzis, A. Mavrandonakis, W. Klopper, G. E. Froudakis, *ChemPhysChem*, 2009, **10**, 374.
29 H. M. Lee, I. S. Youn, M. Saleh, J. W. Lee, K. S. Kim, *Phys. Chem. Chem. Phys.*, 2015, **17**, 10925-.
30 K. D. Vogiatzis, W. Klopper, J. Friedrich, *J. Chem. Theory Comput.*, 2015, **11**, 1574.
31 J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas, K. D. Vogiatzis, *Nat. Commun.*, 2020, **11**, 3230.
32 Z. Tian, T. Saito, D. Jiang, *J. Phys. Chem. A*, 2015, **119**, 3848.
33 A. G. Sylvanus, G. M. Jones, R. Custelcean, K. D. Vogiatzis, *ChemPhysChem*, 2025, **24**, e202400498.
34 B. Zheng, J. Bai, J. Duan, L. Wojtas, M. J. Zaworotko, *J. Am. Chem. Soc.*, 2011, **133**, 748.
35 T. Zelenka, K. Simanova, R. Saini, G. Zelenkova, S. P. Nehra, A. Sharma, M. Almasi, *Sci. Rep.*, 2022, **12**, 17366.
36 RDKit: Open-source Cheminformatics; 10.5281/zenodo.591637. https://www.rdkit.org.
37 P. Ertl, A. Schuffenhauer, *J. Cheminform.*, 2009, **1**, 8.
38 P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.*, 2008, **48**, 68.
39 G. M. Jones, B. Story, V. Maroulas, K. D. Vogiatzis, Molecular Representations for Machine Learning, ACS, 2023.

Data Availability Statement for

Design of CO_2-philic Molecular Units with Large Language Models

Konstantinos D. Vogiatzis

The data supporting this article have been included as part of the Supplementary Information. In particular, we provide the following:

- Exact prompts and responses from the LLMs
- Computational details for reproducing all results reported in the manuscript
- List of molecular units and computed interaction energies