dZiner: Rational Inverse Design of Materials with AI Agents

Mehrad Ansari*
Acceleration Consortium
University of Toronto
80 St George St, Toronto, ON M5S 3H6
mehrad.ansari@utoronto.ca

Jeffrey Watchorn
Acceleration Consortium
University of Toronto
80 St George St, Toronto, ON M5S 3H6
jeff.watchorn@utoronto.ca

Carla E. Brown
Acceleration Consortium
University of Toronto
80 St George St, Toronto, ON M5S 3H6
carla.brown@utoronto.ca

Joseph S. Brown
Acceleration Consortium
University of Toronto
80 St George St, Toronto, ON M5S 3H6
js.brown@utoronto.ca

Abstract

Recent breakthroughs in machine learning and artificial intelligence, fueled by scientific data, are revolutionizing the discovery of new materials. Despite the wealth of existing scientific literature, the availability of both structured experimental data and chemical domain knowledge that can be easily integrated into data-driven workflows is limited. The motivation to integrate this information, as well as additional context from first-principle calculations and physics-informed deep learning surrogate models, is to enable efficient exploration of the relevant chemical space and to predict structure-property relationships of new materials *a priori*. Ultimately, such a framework could replicate the expertise of human subject-matter experts. In this work, we present dZiner, a chemist AI agent, powered by large language models (LLMs), that discovers new compounds with desired properties via inverse design (property-to-structure). In specific, the agent leverages domain-specific insights from foundational scientific literature to propose new materials with enhanced chemical properties, iteratively evaluating them using relevant surrogate models in a rational design process, while accounting for design constraints. The model supports both closed-loop and human-in-the-loop feedback cycles enabling human-AI collaboration in molecular design with real-time property inference, and uncertainty and chemical feasibility assessment. We demonstrate the flexibility of this agent by applying it to various materials target properties including surfactants, ligand and drug candidates, and metal-organic frameworks. Our approach holds promise to both accelerate the discovery of new materials and enable the targeted design of materials with desired functionalities. The methodology is available as an open-source software on https://github.com/mehradans92/dZiner.

1 Introduction

The discovery of new molecules and materials with advanced properties is essential for tackling significant challenges, ranging from therapeutic discovery to addressing climate change. The evolution of materials innovation has gone through four distinct paradigms [1]. Initially, it primarily relied on empirical trial and error. And then, as disciplines like mathematics, chemistry, and physics advanced,

*Denotes corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

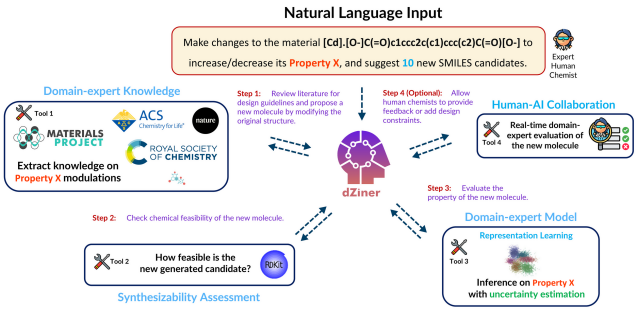

Figure 1: **dZiner workflow overview**. The model starts by inputting the material's initial structure as a textual representation. The AI agent dynamically retrieves domain-knowledge (design guidelines) for Property X from scientific literature, the Internet or other resources. Based on these guidelines, and any additional design constraints provided in *natural language*, the agent proposes a new candidate and assesses its chemical feasibility in real-time. Next, it estimates Property X for the new candidate, incorporating epistemic uncertainty, using a cost-efficient surrogate model. Optionally, as part of a human-in-the-loop process, the human chemist can review the agent's new candidates and chain-of-thoughts, providing feedback and suggesting further modifications or constraints, creating an opportunity for human-AI collaboration to guide the exploration process. The agent continues exploring the chemical space, guided by chemistry-informed rules, until it meets the convergence criteria.

materials innovation began to follow scientific laws. The third paradigm emerged with the advent of computational chemistry, illustrated by tools such as Gaussian 70 for ab initio calculations and density functional theory (DFT) [2, 3]. Currently, the fourth paradigm integrates theoretical, experimental, and computational methodologies using data-driven techniques including data mining, cluster analysis, predictive analytics, machine learning (ML), and materials informatics altogether [4, 5].

One major drawback of the traditional materials discovery methods is that it often involves extensive screening through laboratory experiments or in silico simulations, which are both time-consuming and resource-intensive [6, 7]. On the other hand, the promising data-driven approaches that use machine learning surrogate models to predict material structures and properties [8–10] or suggest novel materials [11, 12] rely heavily on extensive training datasets. However, these models face challenges when such data is unavailable or when there is only a limited budget for conducting experiments or simulations. In contrast, a human expert would be much more effective in such cases, by leveraging their domain knowledge and reasoning from limited examples. This underscores the need for a new materials design paradigm, where models are built to replicate and/or integrate the expertise of human domain experts.

The emergence of large language models (LLMs), which excel at understanding and processing natural language text, presents a promising opportunity to integrate primary sources from complex scientific literature, diverse datasets, and human expertise toward the acceleration of scientific discoveries. LLMs have excelled at various tasks, even those they are not explicitly trained for, which has led to increasing interest in creating LLM-based agents with abilities including human-mimicking reasoning, self-reflection, and decision-making [13–15]. These autonomous agents can be augmented with external tools or action modules, enabling them to surpass conventional text processing and directly interact with the physical world (i.e. robotic manipulation [16–18] and scientific experimentation [19, 20]). By integrating tools such as plugins specific to domain expertise, these agents can overcome the inherent deficiencies of LLMs in specific domains and enhance their overall applicability, performance, and interpretability [21, 22]. For instance, recent studies have demonstrated the use of LLM agents to extract materials datasets and scientific research [23–27], chemical innovation [28], experiment planning [29] and predicting experimental outcomes [30], hypothesis generation [31, 32], and closed-loop or human-in-the-loop molecular discovery [33, 34], among many other applications. An excellent overview of LLM-based autonomous agents in chemistry and materials can be found in reference [35].

In this work, we present dZiner, an agent-based framework for rational inverse design of materials, powered by the state-of-the-art LLMs (Figure 1). Leveraging both human expertise and the existing knowledge contained in scientific publications, dZiner acts as an intelligent chemist research assistant [36], providing feedback on every step of the iterative inverse design process. Our agent starts

by inputting the initial molecule as a textual representation, SMILES (Simplified Molecular Input Line-Entry System) or a sequence string, along with a brief description of the property optimization task (e.g., increase binding affinity, decrease critical micelle concentration (CMC), or increase CO_2 adsorption), all in natural language. The agent then interprets human-provided instructions or any design constraints, along with retrieving chemical knowledge from relevant scientific literature, the Internet or other resources to identify possible chemical modifications that could potentially improve the target molecular property. Following these modulation guidelines, the agent generates a new candidate molecule. However, since SMILES strings generated by LLMs may sometimes deviate from proper SMILES grammar, resulting in invalid structures or potential hallucinations, we implement a validation step. This step serves as a quick check to assess the chemical feasibility, and score synthesizability of the newly generated molecule. After this validation process, the effectiveness of the molecule modulation is assessed using a domain-expert model, potentially physics-based. However, to reduce the computational cost of the framework, we limit our study to use more affordable surrogate data-driven models for evaluation rather than expensive DFT or Free Energy Perturbation calculations. The agent then iteratively reviews the modified materials and the entire modification history, stopping the generation of new candidates once the convergence criteria are met. Optionally, in a human-in-the-loop process, a chemist can review the agent's proposed candidates and reasoning, offering feedback and suggesting additional modifications or constraints. This enables human-AI collaboration, allowing the chemist to better guide and refine the exploration process.

This manuscript is structured as follows: Section 2 presents the benchmarking and evaluation of the model's performance across three distinct materials inverse design tasks. Section 3 follows with a discussion on the implications of our findings, the strengths of our approach, and potential directions for future research. The Supplementary Information (section 4) provides model limitations, details of our methodology, including the agent's toolkits, domain-expert knowledge, and synthesizability assessment. Information on the domain-expert surrogate models used in this study, along with the visualization for the 600 AI-generated molecules, are available in the Supplementary Information (section 4).

2 Results

2.1 Surfactant Design and Critical Micelle Concentration Inference

Surfactant molecules play important roles in a wide variety of disciplines of study, from lubricants and coating to pharmaceuticals and drug delivery systems [37]. This wide applicability of study is due to the role of surfactant molecules which act as compatibilizers between dissimilar materials phases. While there are many metrics that are used to characterize surfactant molecules, the most common is the critical micelle concentration (CMC). CMC is traditionally the experimentally determined concentration at which individual surfactant molecules will self-assemble into larger aggregates (micelles). This value is critically important as the desirable properties of surfactants (solubilizing differing phases, enabling biocompatibility etc.) are typically only enabled when the solution concentration of the surfactant is above the CMC [38]. To design surfactant molecules with a desired CMC, the task is often challenging and relies heavily on domain-knowledge based expertise. Hence, the design task of minimizing CMC is both well-suited for an LLM agent, and a desirable objective to reduce the reliance on domain expertise for chemical synthesis.

Given these considerations, we apply dZiner to the rational design of surfactant molecules, with the objective of generating synthesizable molecules that minimize their expected CMC in water at room temperature. The agent was provided with an initial candidate surfactant-like molecule, for these experiments N-(2-oxotetrahydrofuran-3-yl) decanamide, and was tasked with making additions, substitutions or deletions to reduce CMC. The expected CMC with uncertainty is evaluated via a surrogate model as outlined in the methods (section 4.2.2.2). The design guidelines were determined by the agent via providing exemplary journal articles [39–46] on surfactant design. These general guidelines include; 1. hydrophobic tail length and structure; increasing the length of the tail generally reduced CMC while increasing branching reduces CMC, 2. hydrophilic head group size and polarity; larger and more polar head groups generally increase CMC by increasing aqueous solubility, 3. functionalization with heteroatoms, aromatic moieties or other functional groups; modifications to add silicons, fluorines or other groups such as ethylene oxides to the tail or head respectively, reduces CMC. Additionally, the model is asked to keep the molecular weight of the generated candidates

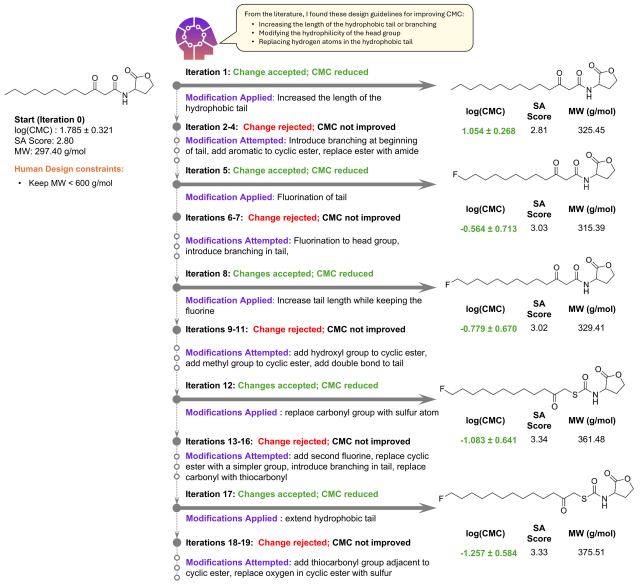

Figure 2: **dZiner's chain-of-thoughts in the closed-loop inverse design of surfactants with lower CMC**. The agent is powered by Claude 3.5 Sonnet. The design guidelines are retrieved from references [39–46], and the model is asked to keep the molecular weight lower than 600 (g/mol) in natural language text. CMC is reduced by two orders of magnitude via iterative agent-suggested chemical modifications.

lower than 600 (g/mol) in natural language text. With this information, the agent was applied to the inverse design task. The resulting iterations of surfactant design powered by Claude 3.5 Sonnet (Figure 2) demonstrated the introduction of several modifications to the initial SMILES structure, that ultimately reduced the expected CMC by roughly two orders of magnitude. In the resulting benchmarking, Sonnet 3.5 agent generally performed better than the one using GPT-4o, generating a larger proportion of chemically valid surfactant molecules (Table 1). The associated chain-of-thoughts analysis for GPT-4o generated surfactants is discussed in detail in the supplementary information (see Figure S1). In analyzing the Sonnet 3.5 agent generated iterations, the agent was able to initially reduce the log(CMC) by first extending the hydrophobic tail of the initial molecule. After this initial improvement, the agent attempts to make a series of modifications to the head group (iterations 2-5) which are rejected. Iterations 5 and 6 yield the largest reduction to the log(CMC) (improvement by 1.618) with the introduction of fluorine heteroatoms to the tail and head of the surfactant molecule, respectively. Notably, during this heteroatom addition Claude 3.5 Sonnet is able to identify potentially invalid molecular structures due to ambiguity in the learned design rules (add heteroatoms to the end of the tail group), and successfully generates molecules that are valid while applying an equivalent modification (in this case, modifying the terminal carbon on the tail group). This behavior was exclusively observed in Claude 3.5 Sonnet agents, and was not reproduced by GPT-4o, likely contributing to the increased performance of 3.5 Sonnet models. Another noteworthy behavior is that the suggested changes from iteration 6 were ultimately reverted, as the addition of head group fluorination dramatically increased the SA score, which the model believed was indicative of a synthetic pathway that would not yield further improvement. Afterwards, additional modifications including the introduction of sulfur heteroatoms and addition groups to the hydrophobic section are able to make modest improvements. Iteration 17 further increased the length of the hydrophobic tail which reduced the log(CMC) to the ultimate value of -1.257, yielding a total improvement of roughly two orders of magnitude over the campaign. The Tanimoto similarity between the initial surfactant molecule and the final molecule was 0.41, demonstrating that for this design task, dZiner is able to make significant and creative changes to the initial molecule over the course of an experiment. Throughout the iterations, the SA score ranged from 2.80 to 3.93, where the candidate molecule with the lowest CMC achieved an SA score of 3.33, only slightly more synthetically complex than the initial candidate molecule. Visualization of the 200 AI-generated molecules in our experiments can be found in Figures S2 and S3 in the Supporting Information.

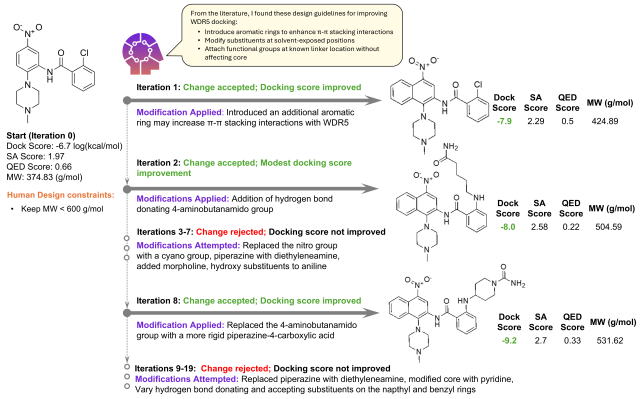

Figure 3: **dZiner's chain-of-thoughts in the closed-loop inverse design of a drug candidate against WDR5 protein target**. The design guidelines are extracted by the agent from references [56, 57], and the model is asked to keep the molecular weight lower than 600 (g/mol) in natural language text. Docking score is reduced by just over two orders of magnitude via iterative agent-suggested chemical modifications (Dock Score $\propto$ log(kcal/mol)).

2.2 Drug Design and Targeted Docking Inference

The discovery of small molecule ligands that bind or inhibit protein targets is ubiquitous to drug development. However, the design and development of drug candidates is challenging and time-consuming given the multi-objective optimization of biological properties including binding affinity (dissociation constant) (K_D), solubility, toxicity, and more. Recent advancements in computational methods have focused primarily on de novo discovery [47–49], while tools to complete hit-to-lead optimization are comparatively lacking. For many novel targets, the discovery and design of small molecules often begins with ligand discovery experiments that discover a "hit" molecule with modest binding affinity to the target. From this hit, multi-objective optimization must be performed to improve the candidate for further biological study, with significant emphasis placed upon potency or binding affinity. Critically, medicinal chemists must be able to synthesize the molecule for testing, somewhat limiting the scope of generative techniques to those that respect synthesizability [50].

Toward this goal, we applied dZiner to the optimization of ligands against WD repeat-containing protein 5 (WDR5). WDR5 is a scaffolding protein that plays a critical role in gene expression and cell differentiation through the assembly of chromatin-modifying complexes, such as the MLL/SET methyltransferase complex [51, 52]. Thus, WDR5 plays a central role in various cancers by supporting oncogenic transcription. Ligands and inhibitors to WDR5 have been studied and reported, ranging in activity with K_D and IC_{50} in the 10s of μM to pM [53, 54]. This rich background of literature allows the evaluation dZiner's performance, as well as opportunity for human-based input and expertise toward the iterative molecular generation and optimization [51, 54]. From this literature, we are able to jump into the position of medicinal chemists that have discovered a hit to WDR5 from high-throughput screening (HTS), demonstrating K_D of 7 ± 1 μM of the native WIN Peptide substrate (see initial starting molecule in Figure 3) [55, 54]. From this HTS hit, dZiner was tasked to perform iterative molecular optimization and improve the binding affinity against WDR5. Various modifications were applied following key guidelines revealed from literature. Additionally, the model is asked to keep the molecular weight of the generated candidates lower than 600 (g/mol) in natural language text. These guidelines, extracted from references [56–58], emphasize the following; 1) enhancing hydrophobic interactions, 2) optimizing solvent-exposed regions, and 3) minimizing steric hindrance without disrupting core binding interactions. To guide dZiner, we used molecular docking, a common method to approximate the binding affinity of a ligand to a protein. Though noisy in its recommendation, molecular docking was selected as a surrogate because it is a rapid and computationally inexpensive way to assess the geometric fit of ligands to protein targets and provide an estimation of binding affinity in kcal/mol [59]. After each iteration, dZiner docked the generated molecule with WDR5 (PDB: 3UVL) using AutoDock Vina [60] and a score was computed for each (see Methods Section 4.2.2.2). For this task, a reduction in docking score indicates lower binding energy, and thus a higher affinity for WDR5.

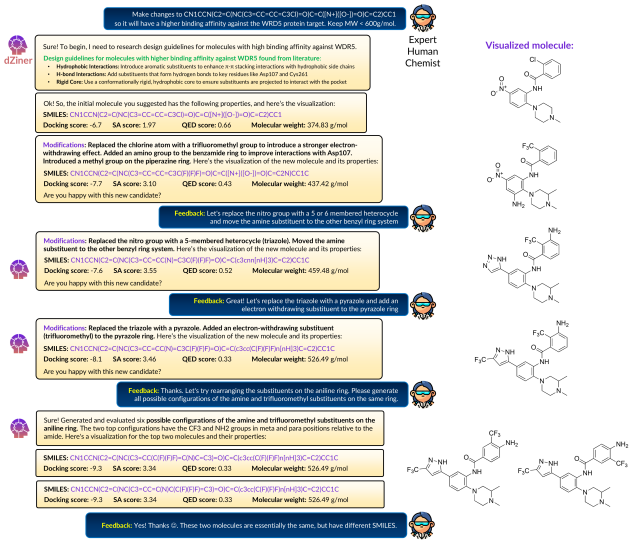

Figure 4: **Human-in-the-loop inverse design of a drug candidate against WDR5 protein target**. The agent is powered by Claude 3.5 Sonnet. dZiner found to be highly cooperative, interpretable, and able to enact changes requested with ease in this human-AI collaboration.

The initial hit provided to dZiner (iteration 0) had a valid structure but relatively modest binding affinity (-6.7). In iteration 1, dZiner adds an aromatic ring to the core to form a rigid napathalene, resulting in a significant improvement in docking score (-7.9). In the next iteration, the chloro-substituent on the benzamide is replaced with a bulky, hydrogen bond donating 4-aminobutanamido group, which has a small positive effect on predicted binding affinity. In the next iteration, the piperazine ring is replaced with an ethylenediamine functionality, significantly compromising the binding affinity (docking score -6.5). This change is reverted for successive analogs. In the following iterations, the -NO_2 group on the napthyl ring is replaced with an alternative electron-withdrawing group (-CN), a morpholino substituent is added to the benzyl ring, and the benzyl ring itself is swapped for a pyridine. Each change is found to be detrimental to the docking score and is thus reverted for the next iteration. The second substantial improvement in docking score occurs in iteration 8 when the 4-aminobutanamido group is replaced with a more rigid piperidine-4-carboxylic acid. The resulting docking score of -9.2 is significantly better than iteration 0 at -6.7. After implementing and reverting several unsuccessful alternatives to the piperazine ring (diethylamine, morpholine), dZiner creates a number of analogs with different H-bond donor and H-bond acceptor groups at variable positions on the benzamide core. These minor modifications result in a total of 10 analogs with a docking score $<$ -8.5. All iterations produced adhered to the human-provided guideline of MW below 600 g/mol. No unstable functional groups were identified in any of the molecules generated by dZiner when using Claude 3.5 Sonnet. Given that docking is a low fidelity method of evaluating binding affinity, each of the 10 analogs with docking score $<$ -8.5 would be promising candidates for synthesis in a hit-to-lead campaign. After 20 iterations, the analogs generated by dZiner have 0.60 Tanimoto similarity to the starting molecule, demonstrating that dziner is capable of making non-trivial modifications to structure to improve binding affinity. Furthermore, these designs generally follow those that were used in the development of OICR-9429 (K_D 24 nM) [52], maintaining the piperizine ring while adding additional hydrophobic or complementary chemical functionalities [54]. When benchmarked against GPT-4o, Claude 3.5 Sonnet generally performed better than GPT-4o, producing fewer invalid SMILES and unstable molecules (see comparison in Table 1). Detailed analysis of the GPT-4o agent generated WDR5 analogs, along with the 200 AI-generated molecules in our experiments can be found in the supplementary information (Figures S4-6).

2.2.1 Human-in-the-loop Design

Collaborative efforts between a human expert and AI agents hold significant promise. In the case of molecular design for WDR5 ligands, we examined human guidance to refine the modifications based on docking scores and structural generation (Figure S7). Similar to the closed-loop analysis (section 2.2), the model initially proposed several modifications to the presented structure in accordance

with the literature-derived guidelines, including the addition of an amine group to promote hydrogen bonding with key WDR5 residue Asp107. A human chemist reviewed this structure and identified the -NO_2 group as a potential toxicophore and priority for replacement. Additionally, the human chemist viewed a published crystal structure of WDR5 (41A9) and hypothesized that the -NH_2 group should be placed on the other benzyl ring to facilitate interaction with Asp107. dZiner was able to competently execute these suggestions and others made by the human chemist to increase binding affinity. dZiner was also able to generate multiple positional isomers when tasked with changing the substitution pattern of an aryl ring. The final molecule generated by dZiner and chemist working in tandem had a significantly improved docking score (-9.3) relative to the starting molecule. Overall dZiner was able to accommodate both general ("add a 5 or 6 carbon heterocycle") and specific ("do not modify the piperazine") feedback, and made several creative suggestions for novel WDR5 analogs. dZiner was found to be highly cooperative, interpretable, and able to enact most changes requested with ease, even following instructions to revert several iterations and make larger-than-average changes to the molecule. dZiner was able to effectively generate molecular ligands to WDR5 while managing input-directed goals. Compared to the closed-loop optimization, the human-in-the-loop optimization enabled more diverse and targeted generation (e.g., location of modification) that could be highly beneficial in a variety of contexts to medicinal chemists investigating structure-activity relationships (SAR). Details on the human-in-the-loop experiment with GPT-4o agent can be found in the supporting information (Figure S7).

2.3 MOF Organic Linker Design and CO_2 Adsorption Capacity Inference

Metal-organic frameworks (MOFs) have gained significant attention in recent years due to their wide range of industrial applications, such as gas adsorption and storage [61], catalysis [62], and drug delivery [63]. These nanocrystalline porous materials are modular [64], consisting of three main building blocks: inorganic nodes, organic nodes, and organic linkers, along with a specific topology that defines the relative positions and orientations of these building blocks that result different MOF properties. We demonstrate the utility of our framework by applying it to the rational

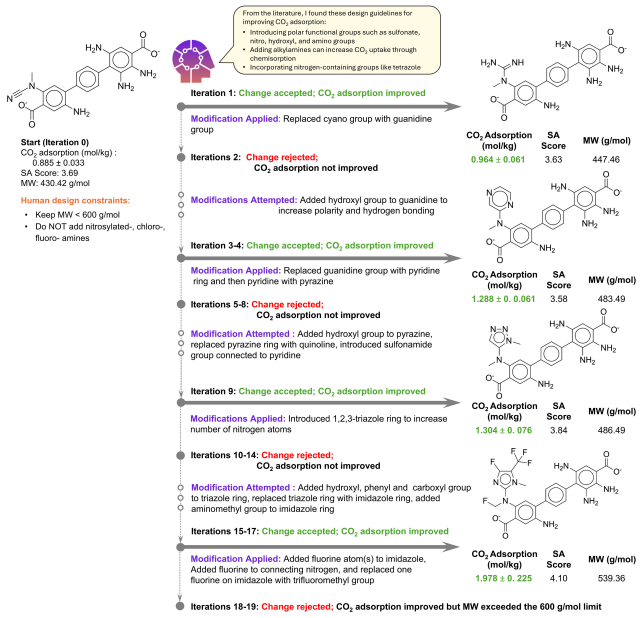

Figure 5: **dZiner's chain-of-thoughts in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity**. The agent is powered by Claude 3.5 Sonnet. Design guidelines were retrieved from references [65–68]. CO_2 adsorption capacity is improved by 85% via iterative agent-suggested chemical modifications, while following additional design constraints.

design of likely synthesizable organic linkers for MOFs with high CO_2 adsorption capacity at 0.5 bar of pressure. These MOFs come with pcu topology and three types of inorganic nodes: Cu paddlewheel, Zn paddlewheel, and Zn tetramer (three most frequent node-topology pairs in the hMOF dataset [69]). In this case, the surrogate CO_2 adsorption predictor is an ensemble of fine-tuned

MOFormers [70] trained on extracted SMILES of the organic linkers from their MOFids in the hmof dataset (see section 4.2.2.2). Throughout the design iterations, as shown in Figure 5, various functional groups were introduced to the initial organic-linker's structure to enhance CO_2 adsorption capacity. These modifications are derived from guidelines emphasizing the alteration of functional groups, incorporation of nitrogen sites, optimization of pore structures and alkylamines to improve interactions and uptake, as automatically extracted from references [65–68]. Additionally, the model includes a set of human design constraints in natural language, specifying to keep the molecular weight lower than 600 (g/mol), and *not* to add nitrosylated, chloro-, fluoro- amines, which are chemically-unstable functional groups (see Figures S9 and S10 for the case study without the latter constraint) The initial organic linker (iteration 0) showed a moderate CO_2 adsorption of 0.885. In iteration 1, nitrogen-containing functional groups like guanidine were introduced, replacing the cyano group to increase polarity and nitrogen content, which significantly improved CO_2 adsorption to 0.964. By iteration 3, further modifications involved replacing guanidine with nitrogen-rich heterocycles like pyridine and pyrazine, boosting adsorption to 1.288. However, adding polar groups like hydroxyl or carboxyl (iterations 5 and 12) slightly decreased CO_2 capture, suggesting that polarity alone was not sufficient for improvement.

The most successful modifications came in iteration 15, where electron-withdrawing fluorine atoms were introduced, leading to a substantial improvement (CO_2 adsorption of 1.698), and promoting the electrostatic interactions and hydrogen bonding of the linker. Further fluorination, including the use of difluoro-substituted and trifluoromethyl groups (iterations 16 to 19), continued to enhance CO_2 adsorption, reaching a maximum of 2.289 in iteration 19. This increase is attributed to the electron-withdrawing properties of fluorine, which enhance the molecule's interaction with CO_2. However, adding too much fluorine also introduced higher uncertainty in adsorption values, indicating potential sensitivity to environmental conditions. Ultimately, iteration 19 demonstrated that maximizing fluorine content, particularly with trifluoromethyl groups, was the most effective strategy for improving CO_2 adsorption. It is important to note that the last three iterations were not accepted, as the suggested molecules exceeded the 600 g/mol molecular weight limit. The final accepted linker (iteration 17) showed a significant improvement in CO_2 adsorption (1.978) with just 0.50 Tanimoto similarity to the initial molecule, demonstrating that substantial and creative changes were made during the experiment. Detailed analysis of the GPT-4o agent generated organic linkers, along with the 200 AI-generated molecules in our experiments can be found in the supplementary information (Figures S8, S11 and S12).

3 Discussion

Our workflow, dZiner, represents an agent-based computational framework for accelerated materials discovery by replicating and incorporating the expertise of human domain experts across various inverse design tasks and target properties, including surfactants, ligand and drug design, and metal-organic frameworks. The inclusion of other expert tools is easy, and our examples demonstrate that dZiner is generally able to adapt across various property-to-structure problems in materials discovery. To better assess the impact of domain-knowledge in our experiments, we repeated all three case studies by removing the design guidelines retrieval tool from the scientific literature. This served as the baseline. In this setting, the agent's modifications to the core structure of the molecules were primarily restricted to the addition or removal of random functional groups and elements, due to the limited non-domain-specific knowledge of the stand-alone LLM at the training time. Tables 1 and 2 provide a detailed breakdown of dZiner's performance across three inverse design tasks—CMC, WDR5 docking, and CO_2 adsorption—evaluating its success with and without domain-knowledge (each was used to generate 600 molecules across all tasks). Across these tasks, dZiner, powered by Claude 3.5 Sonnet, outperforms GPT-4o significantly in both conditions, with especially high success rates when leveraging domain-knowledge from the literature. This is notable in primary objectives such as improving log(CMC), binding affinity (docking score), and CO_2 adsorption. On the importance of incorporating literature-gathered and human expert-based design principles in the workflow, the baseline runs, which operated without design guidelines, exhibited a high failure rate in terms of both generating valid molecular structures and optimizing the target properties, regardless of the choice of the LLM.

We quantified the success of the model in meeting primary objectives by 1) assessing the average improvement in log(CMC), docking score, CO_2 adsorption (Table 2) for the best candidate in

Table 1: dZiner's success rates for the three inverse design tasks with different target properties, evaluated over the generation of 100 molecules per task. Primary objectives for each task are in bold. The baselines include model runs without retrieving domain-knowledge (design guidelines) from the scientific literature.

Target Property	Criteria	Success Rate without Domain-knowledge (%)		Success Rate with Domain-knowledge (%)	
		GPT-4o (baseline)	Claude 3.5 Sonnet (baseline)	GPT-4o	Claude 3.5 Sonnet
Task 1: CMC	Valid SMILES Generation	77	89	79	96
	Lower log(CMC)	55	81	91	92
	Meeting MW Design Constraint	100	100	95	100
	Lower SA Score	31	24	23	19
Task 2: WDR5 Docking	Valid SMILES Generation	63	96	83	100
	Lower Docking Score	60	89	81	96
	Meeting MW Design Constraint	100	100	99	100
	Lower SA Score	43	19	22	6
	Higher QED Score	47	46	13	18
Task 3: CO_2 Adsorption	Valid SMILES Generation	44	99	86	100
	Higher CO_2 Adsorption	39	98	77	98
	Meeting MW Design Constraint	97	97	97	95
	Lower SA Score	76	98	63	41

each run; 2) by comparing each generated iteration to the initial candidate (iteration 0) to see what percentage of molecules have improved on the target property (Table 1). GPT-4o struggled to generate valid SMILES and to suggest new molecules with improved primary objectives. Claude 3.5 Sonnet consistently performed better than GPT-4o on this metric. For instance, its success rate in valid SMILES generation for CMC and CO_2 adsorption tasks was considerably lower than Claude 3.5 Sonnet, highlighting its limitations in molecule design tasks without specialized guidance. On the other hand, it adhered to the human design constraints (molecular weight and the choice of forbidden functional groups), while still optimizing target properties. This efficiency underscores the value of literature-gathered and human expert-based design principles, in complex molecular design tasks, where balancing various criteria is essential for overall success.

Our approach offers several key contributions; 1) The model's flexibility enables the integration of the complete property optimization task with additional design constraints directly through natural language, making the workflow easily adaptable to different target properties simply by altering the input query (prompt) with the use of proper related surrogate model. 2) The augmented domain-expert surrogate models can be easily customized to target specific properties. This flexibility allows users to either train their own machine learning or deep learning models, or better yet, leverage the existing *state-of-the-art* property predictors from the materials community, avoiding the unnecessary effort of reinventing the wheel. This approach also opens up the possibility of uncertainty estimations of the predicted property via an ensemble of inference models, a capability that typical standalone LLMs do not possess. 3) The model provides *chain-of-thought reasoning*, enabling more interpretable results and a clearer understanding of its chemistry-informed decision-making processes. 4) The workflow supports both closed-loop and human-in-the-loop inverse design. In the human-in-the-loop scenario, a domain expert can interact with the model through natural language to provide feedback on newly suggested candidates, propose modifications, or introduce additional design constraints. 5) Because of the iterative design approach, we observed that most molecular candidates maintained a strong relative amount of synthesizbility compared to completely generative approaches, especially seen in the CMC and WDR5 ligand design.

Table 2: dZiner's improvement rates for the primary objectives in the three inverse design tasks with different target properties, evaluated over the generation of 100 molecules per task. On average, improvements are determined by comparing the best candidate from each run to the initial candidate (iteration 0). The baselines include model runs without retrieving domain-knowledge (design guidelines) from the scientific literature.

Target Property	Criteria	Improvement without Domain-knowledge (%)		Improvement with Domain-knowledge (%)	
		GPT-4o (baseline)	Claude 3.5 Sonnet (baseline)	GPT-4o	Claude 3.5 Sonnet
Task 1: CMC	**Average log(CMC)**	34	86	95	137
Task 2: WDR5 Docking	**Average Docking Score**	16	19	31	31
Task 3: CO_2 Adsorption	**Average CO_2 Adsorption**	41	28	46	108

Data and Code Availability

All data, code and model architectures and fine-tuned weights for the surrogate models used to produce results in this study are publicly available in the following GitHub repository: https://github.com/mehradans92/dZiner.

Acknowledgments

This research was undertaken thanks in part to funding provided to the University of Toronto's Acceleration Consortium from the Canada First Research Excellence Fund: Grant number - CFREF-2022-00042. The authors thank Santha Santhakumar at the Acceleration Consortium for his valuable feedback on the chemical assessment of the AI-generated molecules.

References

[1] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials*, 4(5), 2016.

[2] John A Pople. Quantum chemical models (nobel lecture). *Angewandte Chemie International Edition*, 38(13-14):1894–1902, 1999.

[3] Kevin F Garrity, Joseph W Bennett, Karin M Rabe, and David Vanderbilt. Pseudopotentials for high-throughput dft calculations. *Computational Materials Science*, 81:446–452, 2014.

[4] Amir Mosavi and Atieh Vaezipour. Reactive search optimization; application to multiobjective optimization problems. *Applied Mathematics*, 3(10A):1572–1582, 2012.

[5] S Samudrala, K Rajan, and B Ganapathysubramanian. Informatics for materials science and engineering, 2013.

[6] Geoffroy Hautier, Anubhav Jain, and Shyue Ping Ong. From the computer to the laboratory: materials discovery and design using first-principles calculations. *Journal of Materials Science*, 47:7317–7340, 2012.

[7] Edward O Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annual Review of Materials Research*, 45(1):195–216, 2015.

[8] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[9] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[10] Mehrad Ansari and Andrew D White. Learning peptide properties with positive examples only. *Digital Discovery*, 3(5):977–986, 2024.

[11] Z Ren, SIP Tian, J Noh, F Oviedo, G Xing, J Li, Q Liang, R Zhu, AG Aberle, S Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. matter 5, 314–335, 2022.

[12] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[14] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[15] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022.

[16] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[17] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

[18] Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, et al. Organa: A robotic assistant for automated chemistry experimentation and characterization. *arXiv preprint arXiv:2401.06949*, 2024.

[19] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

[20] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

[21] Geemi P Wellawatte and Philippe Schwaller. Extracting human interpretable structure-property relationships in chemistry using xai and large language models. *arXiv preprint arXiv:2311.04047*, 2023.

[22] Seongmin Kim, Joshua Schrier, and Yousung Jung. Explainable synthesizability prediction of inorganic crystal structures using large language models. *chemrxiv preprint 10.26434/chemrxiv-2024-ltncz*, 2024.

[23] Mehrad Ansari and Seyed Mohamad Moosavi. Agent-based learning of materials datasets from scientific literature. *arXiv preprint arXiv:2312.11690*, 2023.

[24] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

[25] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: Large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*, 2024.

[26] Shi Xuan Leong, Sergio Pablo-García, Zijian Zhang, and Alán Aspuru-Guzik. Automated electrosynthesis reaction mining with multimodal large language models (mllms). 2024.

[27] Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge, 2024. URL https://arxiv.org/abs/2409.13740.

[28] Kexin Chen, Junyou Li, Kunyi Wang, Yuyang Du, Jiahui Yu, Jiamin Lu, Guangyong Chen, Lanqing Li, Jiezhong Qiu, Qun Fang, et al. Towards an automatic ai agent for reaction condition recommendation in chemical synthesis. *arXiv preprint arXiv:2311.10776*, 2023.

[29] Michael H Prince, Henry Chan, Aikaterini Vriza, Tao Zhou, Varuni K Sastry, Matthew T Dearing, Ross J Harder, Rama K Vasudevan, and Mathew J Cherukara. Opportunities for retrieval and tool augmented large language models in scientific facilities. *arXiv preprint arXiv:2312.01291*, 2023.

[30] Mayk Caldas Ramos, Shane S Michtavy, Marc D Porosoff, and Andrew D White. Bayesian optimization of catalysts with in-context learning. *arXiv preprint arXiv:2304.05341*, 2023.

[31] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.

[32] Xuemei Gu and Mario Krenn. Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models. *arXiv preprint arXiv:2405.17044*, 2024.

[33] Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models. *arXiv preprint arXiv:2406.13163*, 2024.

[34] Andrew D McNaughton, Gautham Ramalaxmi, Agustin Kruel, Carter R Knutson, Rohith A Varikoti, and Neeraj Kumar. Cactus: Chemistry agent connecting tool-usage to science. *arXiv preprint arXiv:2405.00972*, 2024.

[35] Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. A review of large language models and autonomous agents in chemistry. *arXiv preprint arXiv:2407.01603*, 2024.

[36] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

[37] Sourav De, Susanta Malik, Aniruddha Ghosh, Rumpa Saha, and Bidyut Saha. A review on natural surfactants. *RSC advances*, 5(81):65757–65767, 2015.

[38] Diego Romano Perinelli, Marco Cespi, Nicola Lorusso, Giovanni Filippo Palmieri, Giulia Bonacucina, and Paolo Blasi. Surfactant self-assembling and critical micelle concentration: one approach fits all? *Langmuir*, 36(21):5745–5753, 2020.

[39] Adam Czajka, Gavin Hazell, and Julian Eastoe. Surfactants at the design limit. *Langmuir*, 31 (30):8205–8217, 2015.

[40] Theophile Gaudin, Patricia Rotureau, Isabelle Pezron, and Guillaume Fayet. New qspr models to predict the critical micelle concentration of sugar-based surfactants. *Industrial & Engineering Chemistry Research*, 55(45):11716–11726, 2016.

[41] Anna Mozrzymas and Bożenna Różycka-Roszak. Prediction of critical micelle concentration of cationic surfactants using connectivity indices. *Journal of mathematical chemistry*, 49(1): 276–289, 2011.

[42] Paul DT Huibers, Victor S Lobanov, AR Katritzky, DO Shah, and M Karelson. Prediction of critical micelle concentration using a quantitative structure–property relationship approach. *Journal of Colloid and Interface Science*, 187(1):113–120, 1997.

[43] Xuefeng Li, Gaoyong Zhang, Jinfeng Dong, Xiaohai Zhou, Xiaoci Yan, and Mingdao Luo. Estimation of critical micelle concentration of anionic surfactants with qspr approach. *Journal of Molecular Structure: THEOCHEM*, 710(1-3):119–126, 2004.

[44] Li Xuefeng, Zhang Gaoyong, Dong Jinfeng, Zhou Xiaohai, Yan Xiaoci, and Luo Mingdao. Correlation of critical micelle concentration of sodium alkyl benzenesulfonates with molecular descriptors. *Wuhan University Journal of Natural Sciences*, 11:409–414, 2006.

[45] Alexander Moriarty, Takeshi Kobayashi, Matteo Salvalaglio, Panagiota Angeli, Alberto Striolo, and Ian McRobbie. Analyzing the accuracy of critical micelle concentration predictions using deep learning. *Journal of Chemical Theory and Computation*, 19(20):7371–7386, 2023.

[46] Nada Boukelkal, Soufiane Rahal, Redha Rebhi, and Mabrouk Hamadache. Qspr for the prediction of critical micelle concentration of different classes of surfactants using machine learning algorithms. *Journal of Molecular Graphics and Modelling*, 129:108757, 2024.

[47] Lei Huang, Tingyang Xu, Yang Yu, Peilin Zhao, Xingjian Chen, Jing Han, Zhi Xie, Hailong Li, Wenge Zhong, Ka-Chun Wong, et al. A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications*, 15(1):2657, 2024.

[48] Zaixi Zhang, Wanxiang Shen, Qi Liu, and Marinka Zitnik. Pocketgen: Generating full-atom ligand-binding protein pockets. *bioRxiv*, 2024. URL https://www.biorxiv.org/content/10.1101/2024.02.25.581968.

[49] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[50] Michał Koziarski, Andrei Rekesh, Dmytro Shevchuk, Almer van der Sloot, Piotr Gaiński, Yoshua Bengio, Cheng-Hao Liu, Mike Tyers, and Robert A Batey. Rgfn: Synthesizable molecular generation using gflownets. *arXiv preprint arXiv:2406.08506*, 2024.

[51] Xin Chen, Junjie Xu, Xianghan Wang, Guanlu Long, Qidong You, and Xiaoke Guo. Targeting wd repeat-containing protein 5 (wdr5): a medicinal chemistry perspective. *Journal of medicinal chemistry*, 64(15):10537–10556, 2021.

[52] Florian Grebien, Masoud Vedadi, Matthäus Getlik, Roberto Giambruno, Amit Grover, Roberto Avellino, Anna Skucha, Sarah Vittori, Ekaterina Kuznetsova, David Smil, et al. Pharmacological targeting of the wdr5-mll interaction in c/ebpα n-terminal leukemia. *Nature chemical biology*, 11(8):571–578, 2015.

[53] Erin R Aho, Jing Wang, Rocco D Gogliotti, Gregory C Howard, Jason Phan, Pankaj Acharya, Jonathan D Macdonald, Ken Cheng, Shelly L Lorey, Bin Lu, et al. Displacement of wdr5 from chromatin by a win site inhibitor with picomolar affinity. *Cell reports*, 26(11):2916–2928, 2019.

[54] Matthaus Getlik, David Smil, Carlos Zepeda-Velazquez, Yuri Bolshan, Gennady Poda, Hong Wu, Aiping Dong, Ekaterina Kuznetsova, Richard Marcellus, Guillermo Senisterra, et al. Structure-based optimization of a small molecule antagonist of the interaction between wd repeat-containing protein 5 (wdr5) and mixed-lineage leukemia 1 (mll1). *Journal of medicinal chemistry*, 59(6):2478–2496, 2016.

[55] Guillermo Senisterra, Hong Wu, Abdellah Allali-Hassani, Gregory A Wasney, Dalia Barsyte-Lovejoy, Ludmila Dombrovski, Aiping Dong, Kong T Nguyen, David Smil, Yuri Bolshan, et al. Small-molecule inhibition of mll activity by disruption of its interaction with wdr5. *Biochemical Journal*, 449(1):151–159, 2013.

[56] Wei-Lin Chen, Dong-Dong Li, Zhi-Hui Wang, Xiao-Li Xu, Xiao-Jin Zhang, Zheng-Yu Jiang, Xiao-Ke Guo, and Qi-Dong You. Design, synthesis, and initial evaluation of affinity-based small molecular probe for detection of wdr5. *Bioorganic chemistry*, 76:380–385, 2018.

[57] Kevin B Teuscher, Somenath Chowdhury, Kenneth M Meyers, Jianhua Tian, Jiqing Sai, Mayme Van Meveren, Taylor M South, John L Sensintaffar, Tyson A Rietz, Soumita Goswami, et al. Structure-based discovery of potent wd repeat domain 5 inhibitors that demonstrate efficacy and safety in preclinical animal models. *Proceedings of the National Academy of Sciences*, 120 (1):e2211297120, 2023.

[58] Dong-Dong Li, Wei-Lin Chen, Zhi-Hui Wang, Yi-Yue Xie, Xiao-Li Xu, Zheng-Yu Jiang, Xiao-Jin Zhang, Qi-Dong You, and Xiao-Ke Guo. High-affinity small molecular blockers of mixed lineage leukemia 1 (mll1)-wdr5 interaction inhibit mll1 complex h3k4 methyltransferase activity. *European journal of medicinal chemistry*, 124:480–489, 2016.

[59] Brian J Bender, Stefan Gahbauer, Andreas Luttens, Jiankun Lyu, Chase M Webb, Reed M Stein, Elissa A Fink, Trent E Balius, Jens Carlsson, John J Irwin, et al. A practical guide to large-scale docking. *Nature protocols*, 16(10):4799–4832, 2021.

[60] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[61] Hao Li, Kecheng Wang, Yujia Sun, Christina T Lollar, Jialuo Li, and Hong-Cai Zhou. Recent advances in gas storage and separation using metal–organic frameworks. *Materials Today*, 21 (2):108–121, 2018.

[62] Mengjie Hao, Muqing Qiu, Hui Yang, Baowei Hu, and Xiangxue Wang. Recent advances on preparation and environmental applications of mof-derived carbons in catalysis. *Science of the Total Environment*, 760:143333, 2021.

[63] Harrison D Lawson, S Patrick Walton, and Christina Chan. Metal–organic frameworks for drug delivery: a design perspective. *ACS applied materials & interfaces*, 13(6):7004–7020, 2021.

[64] Markus J Kalmutzki, Nikita Hanikel, and Omar M Yaghi. Secondary building units as the turning point in the development of the reticular chemistry of mofs. *Science advances*, 4(10): eaat9180, 2018.

[65] Muhammad Usman, Naseem Iqbal, Tayyaba Noor, Neelam Zaman, Aisha Asghar, Mahmoud M Abdelnaby, Ahmad Galadima, and Aasif Helal. Advanced strategies in metal-organic frameworks for co2 capture and separation. *The Chemical Record*, 22(7):e202100230, 2022.

[66] Christopher A Trickett, Aasif Helal, Bassem A Al-Maythalony, Zain H Yamani, Kyle E Cordova, and Omar M Yaghi. The chemistry of metal–organic frameworks for co2 capture, regeneration and conversion. *Nature Reviews Materials*, 2(8):1–16, 2017.

[67] Witman Matthew, Ling Sanliang, Gladysiak Andrzej, Smit Berend, Slater Ben, Haranczyk Maciej, et al. Rational design of a low-cost, high-performance metal–organic framework for hydrogen storage and carbon capture. 2017.

[68] Kahkasha Parveen and Srimanta Pakhira. Designing organic bridging linkers of metal–organic frameworks for enhanced carbon dioxide adsorption. *New Journal of Chemistry*, 2024.

[69] Christopher E Wilmer, Michael Leaf, Chang Yeon Lee, Omar K Farha, Brad G Hauser, Joseph T Hupp, and Randall Q Snurr. Large-scale screening of hypothetical metal–organic frameworks. *Nature chemistry*, 4(2):83–89, 2012.

[70] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2023.

[71] Tzyy-Shyang Lin, Connor W Coley, Hidenobu Mochigase, Haley K Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L Craig, Jeremiah A Johnson, Julia A Kalow, et al. Bigsmiles: a structurally-based line notation for describing macromolecules. *ACS central science*, 5(9): 1523–1531, 2019.

[72] Inc. Daylight Chemical Information Systems. Smarts-a language for describing molecular patterns, 2007.

[73] Samantha Stuart, Jeffrey Watchorn, and Frank X Gu. Sizing up feature descriptors for macromolecular machine learning with polymeric biomaterials. *npj Computational Materials*, 9(1): 102, 2023.

[74] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[75] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[76] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

[77] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

[78] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.

[79] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[80] OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: 2024-05-26.

[81] Anthropic. Claude sonnet 3.5. https://www.anthropic.com, 2024. https://www.anthropic.com.

[82] Harrison Chase. Langchain, 10 2022. URL https://github.com/langchain-ai/langchain.

[83] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[84] Shiyi Qin, Tianyi Jin, Reid C Van Lehn, and Victor M Zavala. Predicting critical micelle concentrations for surfactants using graph convolutional neural networks. *The Journal of Physical Chemistry B*, 125(37):10610–10620, 2021.

[85] Miguel García-Ortegón, Gregor NC Simm, Austin J Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: easy molecular docking yields better benchmarks for ligand design. *Journal of chemical information and modeling*, 62(15):3486–3502, 2022.

[86] Pamela Zhang, Hwabin Lee, Joseph S Brunzelle, and Jean-Francois Couture. The plasticity of wdr5 peptide-binding cleft enables the binding of the set1 family of histone methyltransferases. *Nucleic acids research*, 40(9):4237–4246, 2012.

[87] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[88] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[89] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

[90] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

[91] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

4 Supplementary Information

4.1 Limitations and Future Work

As demonstrated in the previous sections, dZiner works well as an engine for accelerated molecular discovery and in silico experimentation across a wide range of chemical tasks. Despite this flexibility, there are still areas where dZiner could be further improved. The incorporation of multi-modal data (such as the ability to interpret images and schemes) from supporting literature represents a key advancement that could further improve the performance of the model. This improvement would be especially impactful for small molecule virtual screening tasks. For example, SMILES can be a major oversimplification, especially for complex structures like MOFs. Another limitation inherent to SMILES is that unique SMILES codes can be generated for each substitution pattern, leading to a lack of a one-to-one mapping between molecules and their SMILES representations. These limitations may be balanced by considering that SMILES are likely the most prevalent molecular representations

available in the training data of LLMs, making it easier for LLMs to successfully generate valid new candidates. While the current framework leverages textual representations of molecules, it is not necessarily limited to this format, and the agent can easily incorporate other representations. For example, for stochastic design tasks such as the synthesis of polymers, molecular representations like BigSMILES [71] or SMILES Arbitrary Target Specification (SMARTS) [72], which may better capture the complexity of such systems [73]. It should be noted that given that these other molecular representations are not as popular, the model may lose performance when generating valid candidates. Finally, the number of molecules generated by the workflow is a hyperparameter. Although it is important to note that large numbers can easily exceed the context window of the LLMs used.

For human-in-the-loop runs, the model had limited success in interpreting terms like ortho, meta, and para when a ring has three or more substituents. As an example, it struggled with prompts like "move the -NH_2 group so it is para to the -CF_3 group". This likely a result of the zero-shot learning approach and could be resolved with further domain-specific fine-tuning. Alternatively, adding more instructive human feedbacks or augmenting the agent with additional validation tools to better incorporate chemical rules is expected to alleviate this limitation.

4.2 Methods

4.2.1 AI Agent

In broader terms, an *agent* refers to an entity capable of taking action. The AI agent in this work is powered by an LLM, acting as its brain, expanding its perceptual and action space (environment) through strategies such as multimodal perception and tool utilization [74–78]. In this work, we exploit the *zero-shot learning* capabilities of large language models (LLMs)[79] alongside the ReAct architecture, which supports both reasoning (e.g., chain-of-thought prompting) and taking actions (e.g., generating action plans)[75]. Reasoning traces guide the model in creating, overseeing, and refining action plans, while also addressing exceptions. At the same time, actions enable the model to interact with external sources like knowledge bases or environments to acquire additional information. These knowledge bases are structured as toolkits (see section 4.2.2), enabling the agent to extract relevant molecular design insights from research papers, publicly available datasets, and advanced built-in chemical knowledge, as well as on-the-fly evaluation of modifications with the related domain-expert surrogate models and synthesizability assessment (Algorithm 1). Given that the model is not explicitly provided with labeled data in the input context, the use of these tools is consistent with the broader definition of zero-shot learning. This specialized materials design guidelines lie beyond the scope of typical LLM's training data, enhancing the model's ability to function as an expert chemist in various domains in a zero-shot manner. In this work, we used OpenAI's GPT-4o [80] and Anthropic's Claude 3.5 Sonnet [81] with a temperature of 0.3 as our agent's LLM, and LangChain [82] for the application framework development.

Algorithm 1 dZiner Algorithm (closed-loop)

Input: x_0: chemical structure of the starting molecule
y_{target}: target property label to optimize (can be a minimization or maximization problem).
$\mathcal{L}$: selection of scientific literature on target property optimization.
$\mathcal{S}(x)$: surrogate domain-expert model for evaluating target property, given x.
$\mathcal{M} := \emptyset$: set of history messages, if any.
$(x_{\text{best}}, \hat{y}_{\text{best}}, \hat{\sigma}^2_{\text{epi}_{best}}) \leftarrow (x_0, (\hat{y}_0, \hat{\sigma}^2_{\text{epi}_0}) : \mathcal{S}(x_0))$: initialize best molecule.

Output: $(x_i, y_i, \hat{\sigma}^2_{\text{epi}_i})$: chemical structure and property of the new molecule.

for $i = 1 : N$ **do**
  $m_i, g_i \leftarrow \text{LLM}(x_{i-1}, \hat{y}_{i-1}, y_{\text{target}}, \mathcal{L}, \mathcal{M})$ ▷ m: modification, g: guideline, $\mathcal{M}$: history
  $\tilde{x}_i \leftarrow \text{LLM: modify structure}(x_{i-1}, g_i)$ ▷ $\tilde{x}_i$: modified molecule
  $x_i, v_i \leftarrow \mathcal{V}(\tilde{x}_i)$ ▷ x_i: new molecule, v_i: validity check
  if v_i is invalid **then**
    return Invalid molecule
    $x_i, \hat{y}_i, \hat{\sigma}^2_{\text{epi}_i} \leftarrow x_{\text{best}}, \hat{y}_{\text{best}}, \hat{\sigma}^2_{\text{epi}_{best}}$ ▷ Revert to best
    continue
  end if
  $\hat{y}_i, \hat{\sigma}^2_{\text{epi}_i} \leftarrow \mathcal{S}(x_i)$ ▷ $\hat{y}_i$: property, $\hat{\sigma}^2_{\text{epi}_i}$: epistemic uncertainty
  if $|\hat{y}_i - y_{\text{target}}| \geq |\hat{y}_{\text{best}} - y_{\text{target}}|$ **then**
    $x_i, \hat{y}_i, \hat{\sigma}^2_{\text{epi}_i} \leftarrow x_{\text{best}}, \hat{y}_{\text{best}}, \hat{\sigma}^2_{\text{epi}_{best}}$ ▷ Revert to best
    continue
  else
    $(x_{\text{best}}, \hat{y}_{\text{best}}, \hat{\sigma}^2_{\text{epi}_{best}}) \leftarrow (x_i, \hat{y}_i, \hat{\sigma}^2_{\text{epi}_i})$ ▷ Update best
  end if
  $h_i \leftarrow \text{create history message}(x_i, \hat{y}_i, \hat{\sigma}^2_{\text{epi}_i}, v_i, g_i)$ ▷ h_i: history message
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{h_i\}$ ▷ Update history
end for

Notation:
N: max number of new molecules
$\mathcal{V}$: Chemical feasibility and synthesizeability assessment; $\mathcal{M}$: history

4.2.2 Agent Toolkits

4.2.2.1 Domain-expert Knowledge

This tool enables the agent to do retrieval-augmented generation (RAG), and extract design guidelines from unstructured text, offering insights on how to modify the core structure of a molecule to optimize a specific property. It identifies the most relevant sentences from research papers in response to a query, focusing on suggestions for molecular modifications that enhance the desired property. The process involves embedding both the paper and the query as numerical vectors, and then selecting the top k passages within the document that either explicitly mention or implicitly hint at adaptations to optimize the band gap property of a MOF. The embedding model used is OpenAI's text-embedding-3-large. Drawing on our previous work [23], k is set to 9 but is dynamically adjusted based on the context's length to prevent OpenAI's token limitation errors. The semantic similarity search is ranked using Maximum Marginal Relevance (MMR) [83], based on cosine similarity, which is defined as:

$$\text{MMR} = \arg \max_{d_i \in R \setminus S} \left[\lambda \cdot \cos(d_i, q) - (1 - \lambda) \cdot \max_{d_j \in S} \cos(d_i, d_j) \right] \tag{1}$$

Here, d_i represents a document from the set of retrieved documents R, S is the set of already selected documents, and q is the query. The parameter λ, which ranges from 0 and 1, controls the balance between relevance to the query and diversity (i.e., novelty compared to the already selected documents). In this work, we use the default value of 0.5. The purpose of MMR is to retrieve documents that are both relevant to the query and diverse, minimizing redundancy in the results.

4.2.2.2 Domain-expert Surrogate Models

For the case studies in sections 2.1 and 2.3, we use ensemble modeling to estimate prediction uncertainty, thus enhancing the predictive capability of the domain-expert surrogate model. For a given data point $\vec{x}$, the ensemble prediction average ($\hat{y}(\vec{x})$) is calculated as follows:

$$\hat{y}(\vec{x}) = \frac{1}{N}\sum_{m}\hat{y}_m(\vec{x}) \tag{2}$$

$$\hat{\sigma}^2_{\text{epi}}(\vec{x}) = \frac{1}{N}\sum_{m}\left(\hat{y}(\vec{x}) - \hat{y}_m(\vec{x})\right)^2 \tag{3}$$

where N is the ensemble size, and m indexes the model in the ensemble. $\hat{\sigma}^2_{\text{epi}}(\vec{x})$ denotes the epistemic uncertainty, quantifying the disagreements amongst model estimations.

Critical Micelle Concentration Inference Model This model is based on the work of Qin et al. [84] without any adaptations. The authors used a Graph Convolutional Neural Network (GCN) to predict the critical micelle concentration (CMC) of surfactants based on their molecular structure as SMILES input. The CMC training data were experimentally measured at room temperature (between 20 and 25 °C) in water for 202 surfactants coming from various classes, including nonionic, cationic, anionic, and zwitterionic. The model architecture leverages graph convolutional layers to process molecular graphs, where atoms are represented as nodes and bonds as edges, effectively capturing both topological and constitutional information. The GCN includes average pooling to aggregate atom-level features into a fixed-size graph-level vector, followed by fully connected layers with ReLU activations for the final regression of the log CMC value. In terms of performance, the GCN has a root-mean-squared-error (RMSE) of 0.23 and an R^2 of 0.96 on test data for nonionic surfactants, outperforming previous quantitative structure-property relationship (QSPR) models. The ensemble of models used in our study are based on an 11-fold cross-validation with mean RMSE of 0.32. For more details on the model, refer to reference [84].

Targeted Docking Inference Model This model utilizes on Dockstring [85], to predict the fit and binding affinity of small molecules (ligands) bind to target proteins by using molecular docking. Dockstring is a user-friendly Python wrapper for AutoDock Vina [60]. WDR5 (PDB: 3UVL) [86] was accessed on May 30th 2024 and was prepared for molecular docking in MGLTools / Python Molecular Viewer (1.5.7) by removing the Histone-lysine N-methyltransferase MLL3 peptide ligand, cofactors, and water. The protein was protonated at pH 7.4 by adding Polar Only Hydrogens. Kollmann charges were added. Dockstring uses a targeted version as opposed to blind docking and a 30 Angstrom grid box was defined around the central binding pocket of WDR5. Docking was performed using default exhaustiveness and energy range. Docking results were returned and used without any rescoring in single measurements.

CO_2 Adsorption Inference Model This model is based on MOFormer, a self-supervised Transformer model developed for predicting the properties of Metal-Organic Frameworks (MOFs) using a structure-agnostic approach [70]. Unlike traditional models that rely on 3D atomic structures, MOFormer uses a text-based representation of MOFs, known as MOFid. The model utilizes the self-attention mechanism of Transformers to capture complex relationships within MOFs and is pretrained on over 400,000 MOF structures using self-supervised learning with Barlow-Twin loss [87]. This pretraining improves the prediction accuracy, as it aligns the textual-based representations of MOFormer with the structure-based representation leaning of a Crystal Graph Convolutional Neural Network (CGCNN) [88]. For the specific task of predicting CO2 adsorption capacity at 0.5 bar, MOFormer achieved a mean absolute error (MAE) of 0.545 mol/kg, whereas our fine-tuned ensemble of models with 5-fold training on the SMILES of the organic linkers have a MAE of 0.894 mol/kg. For more details on the model, refer to reference [70].

4.2.2.3 Synthesizability Assessment

This tool uses RDKit [89] to convert a SMILES string into an RDKit *Mol* object and performs several validation steps, including syntax parsing, atom and bond validation, checking atomic

valences, verifying ring closure notation, adding implicit hydrogens, and detecting aromatic systems. These processes ensure the basic chemical validity of the molecule. In addition to the chemical feasibility assessment, we use a heuristic measure of synthesizability: synthetic accessibility score (SA score) [90], which is based on the analysis of one million PubChem molecules and combines fragment contributions from molecular substructures with a complexity penalty that accounts for molecular size and structural features. In section 2.2, we also include a quantitative estimate of drug-likeness (QED) [91], which measures a compound's drug-likeness by integrating molecular properties, such as molecular weight, lipophilicity (logP), polar surface area, and the number of hydrogen bond donors and acceptors, into a single value.

5 Surfactant Design and Critical Micelle Concentration Inference with GPT-4o

Similar to Section 2.1, and starting from the same initial surfactant molecule, we applied dZiner powered by GPT-4o to this property optimization task. The resulting iterations of surfactant design (Figure S1) demonstrated the introduction of several modifications to the initial SMILES structure, that ultimately reduced the expected CMC by roughly two orders of magnitude. Across the first 3 iterations of design, the agent was able to significantly reduce the CMC by introducing additional methyl-type units to the hydrophobic tail (iteration 1), as well as replacing hydrogen atoms in the tail with fluorine (iteration 3).

During this improvement, iteration 2 attempted to introduce branching in the hydrophobic tail, but was rejected after the CMC evaluation did not yield any improvement between iterations 1 and 2. Following several iterations with other rational but ultimately unsuccessful modifications, the agent achieves the largest reduction of log(CMC) in iteration 7 (0.633 to 0.102) by replacing the head group with a series of amide-linked cyclic ethers. Interestingly, the agent completes this modification at the expense of the modification in iteration 3, which ultimately further reduces the CMC beyond what was previously achieved (this behavior also occurred in other benchmarking runs). The final improvement in log(CMC) was achieved in iteration 8 with a further increase to the length of the hydrophobic tail unit to the ultimate value of -0.424. Throughout the experiments the SA score ranged from 2.80 to 4.01, where the candidate molecule with the lowest CMC achieved an SA score of 3.29, only slightly more complex than the initial candidate molecule.

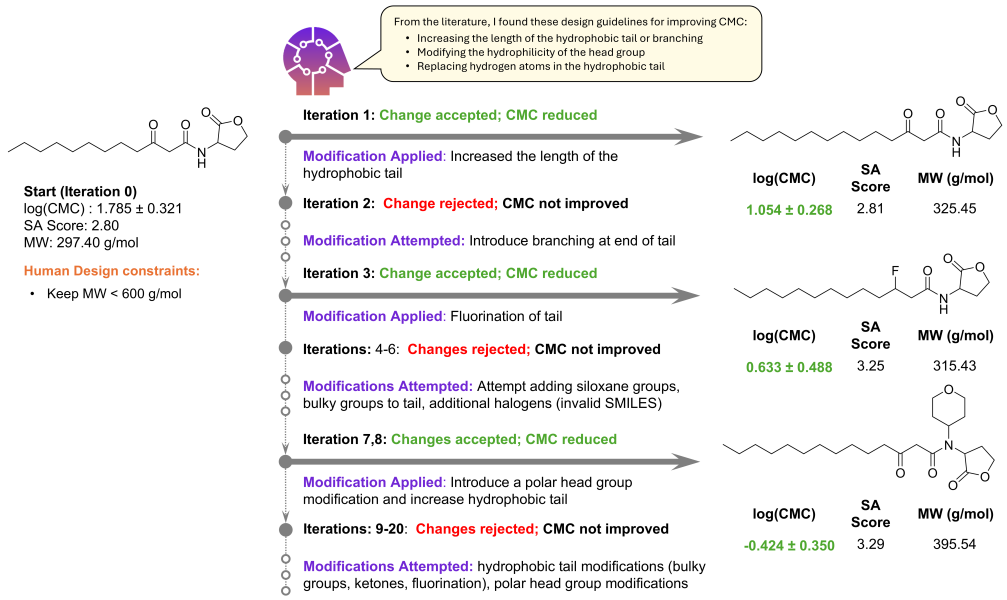

Figure S1: **dZiner's chain-of-thoughts in the closed-loop inverse design of surfactants with lower CMC**. The agent is powered by GPT-4o. The design guidelines are retrieved from literature (same references as in Figure 2), and the model is asked to keep the molecular weight lower than 600 (g/mol) in natural language text. CMC is reduced by two orders of magnitude via iterative agent-suggested chemical modifications. The accepted molecule bears 0.74 similarity (Tanimoto) to the starting molecule after 8 iterations.

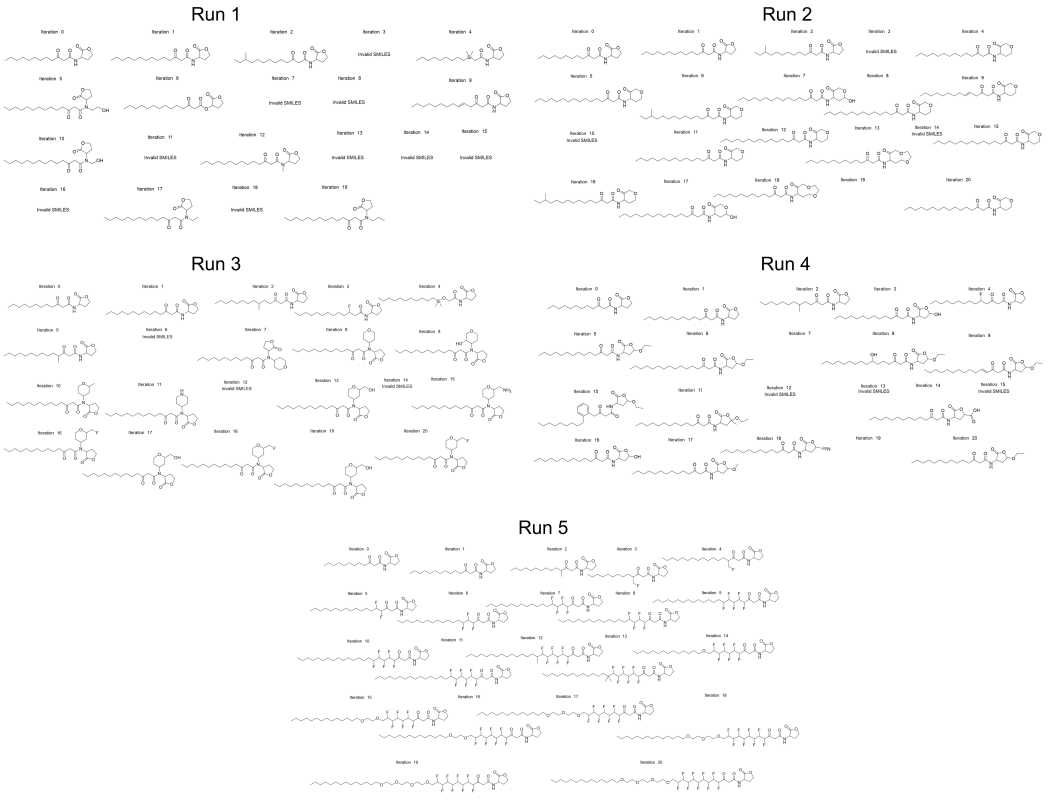

Figure S2: Visualization of the 100 molecules generated in the closed-loop inverse design of surfactants with lower CMC. The agent is powered by GPT-4o. No potentially unstable functional groups were found. Invalid SMILES generated are marked as invalid.

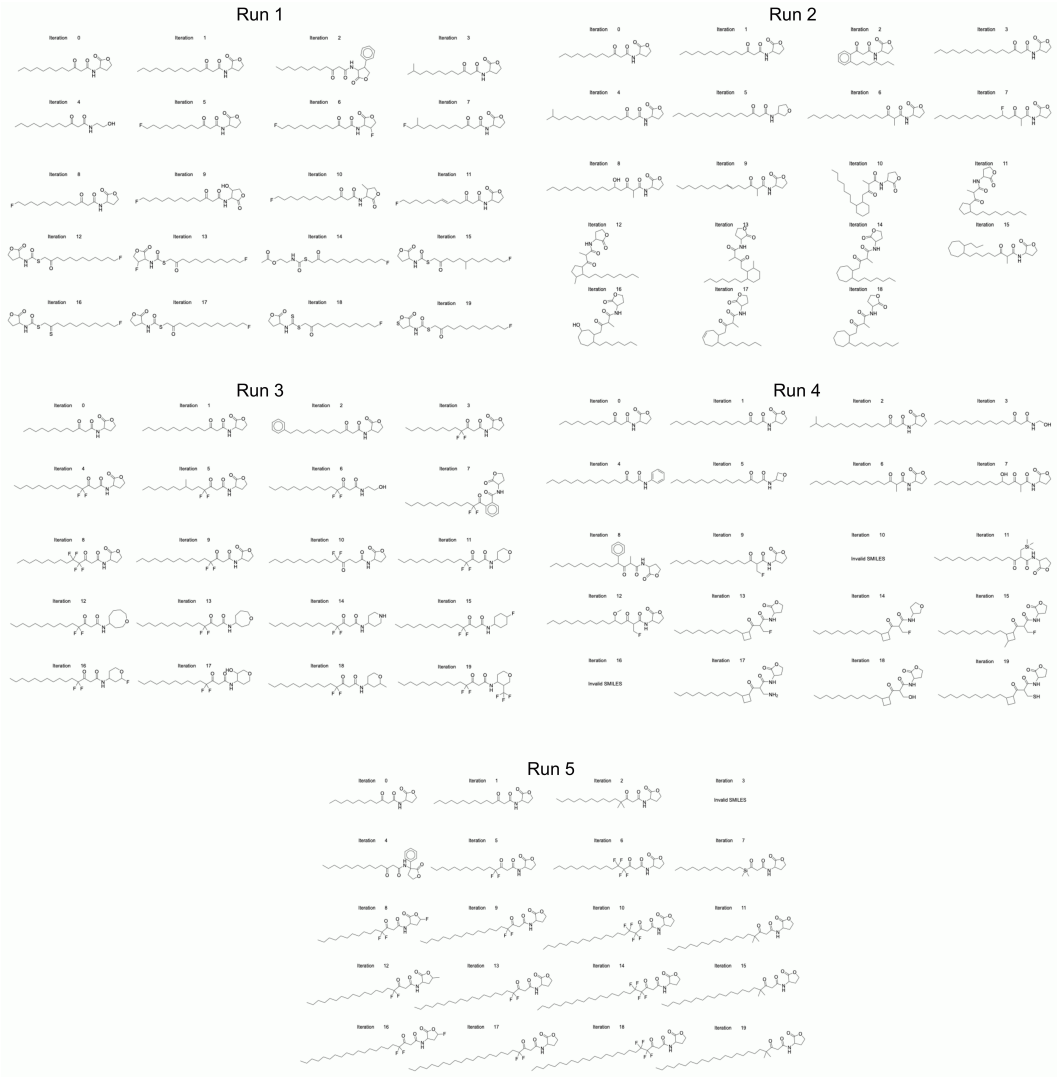

Figure S3: Visualization of the 100 molecules generated in the closed-loop inverse design of surfactants with lower CMC. The agent is powered by Claude 3.5 Sonnet. No potentially unstable functional groups were found. Invalid SMILES generated are marked as invalid.

6 Drug Design and Targeted Docking Inference with GPT-4o

Similar to Section 2.2, and starting from the same initial HTS hit, we applied dZiner powered by GPT-4o to improve docking against WDR5 (see Figure S4). In iteration 1, an aromatic ring was added to strengthen hydrophobic interactions, improving the docking score to -7.2. Further modifications in iteration 3 included replacing the nitro group with a cyano group, yielding a docking score of -7.4. In iterations 4-8, functional groups like methoxy, ethoxy, and butoxy were added to enhance hydrophobic interactions, with the best improvement seen in iteration 8, where a butoxy group raised the docking score to -8.2.

Subsequent iterations aimed to fine-tune the structure by replacing the butoxy group with pentoxy and hexoxy groups, though these did not lead to further improvements. In iteration 11, a trifluoromethyl group was added to the butoxy-substituted structure, yielding the highest docking score of -8.4. This modification optimized interactions within the binding pocket. Other attempts, such as adding a trifluoromethoxy group in iteration 18 and a trifluoromethylthio group in iteration 20, showed varying results but did not surpass the best docking score. Overall, the case study demonstrated

that introducing electron-withdrawing and hydrophobic groups, particularly in iterations 8 and 11, significantly enhanced binding affinity, aligning with the guidelines for targeting WDR5.

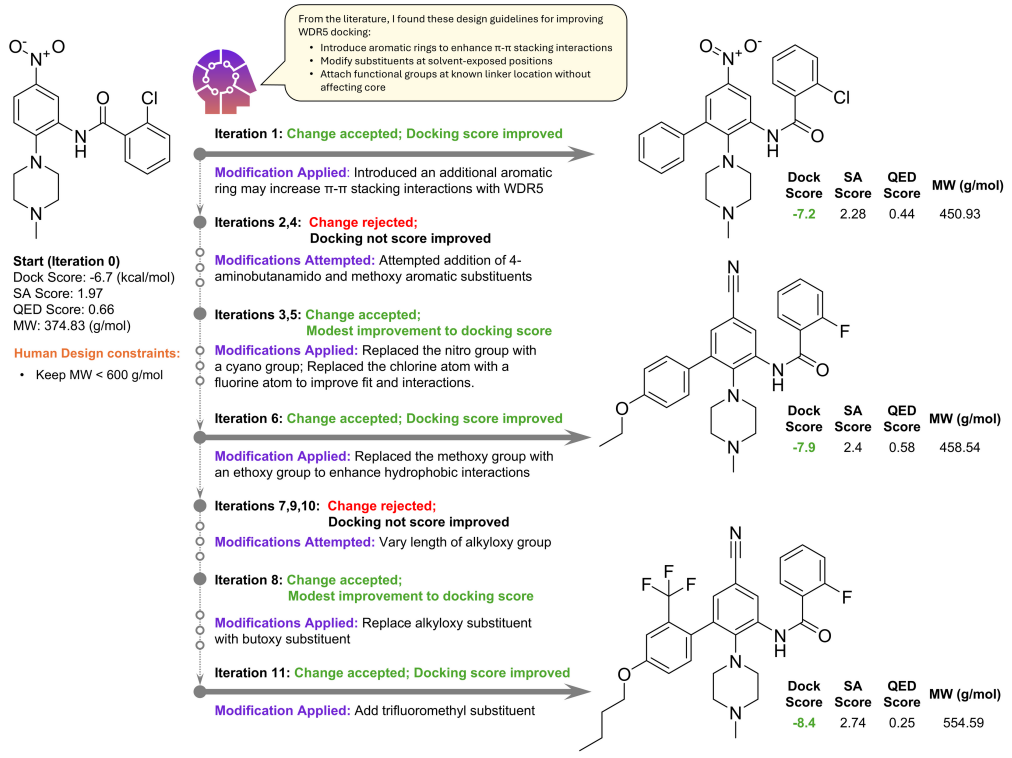

Figure S4: **dZiner's chain-of-thoughts in the closed-loop inverse design of a drug candidate against WDR5 protein target**. The agent is powered by GPT-4o. The design guidelines are extracted by the agent from the same references as in Figure 3, and the model is asked to keep the molecular weight lower than 600 (g/mol) in natural language text. Docking score is reduced by just over two orders of magnitude via iterative agent-suggested chemical modifications (Dock Score = log(kcal/mol)). The accepted molecule has a Tanimoto similarity score of 0.46 compared to the initial molecule, indicating that substantial changes have been made to the structure in the process of improving binding affinity.

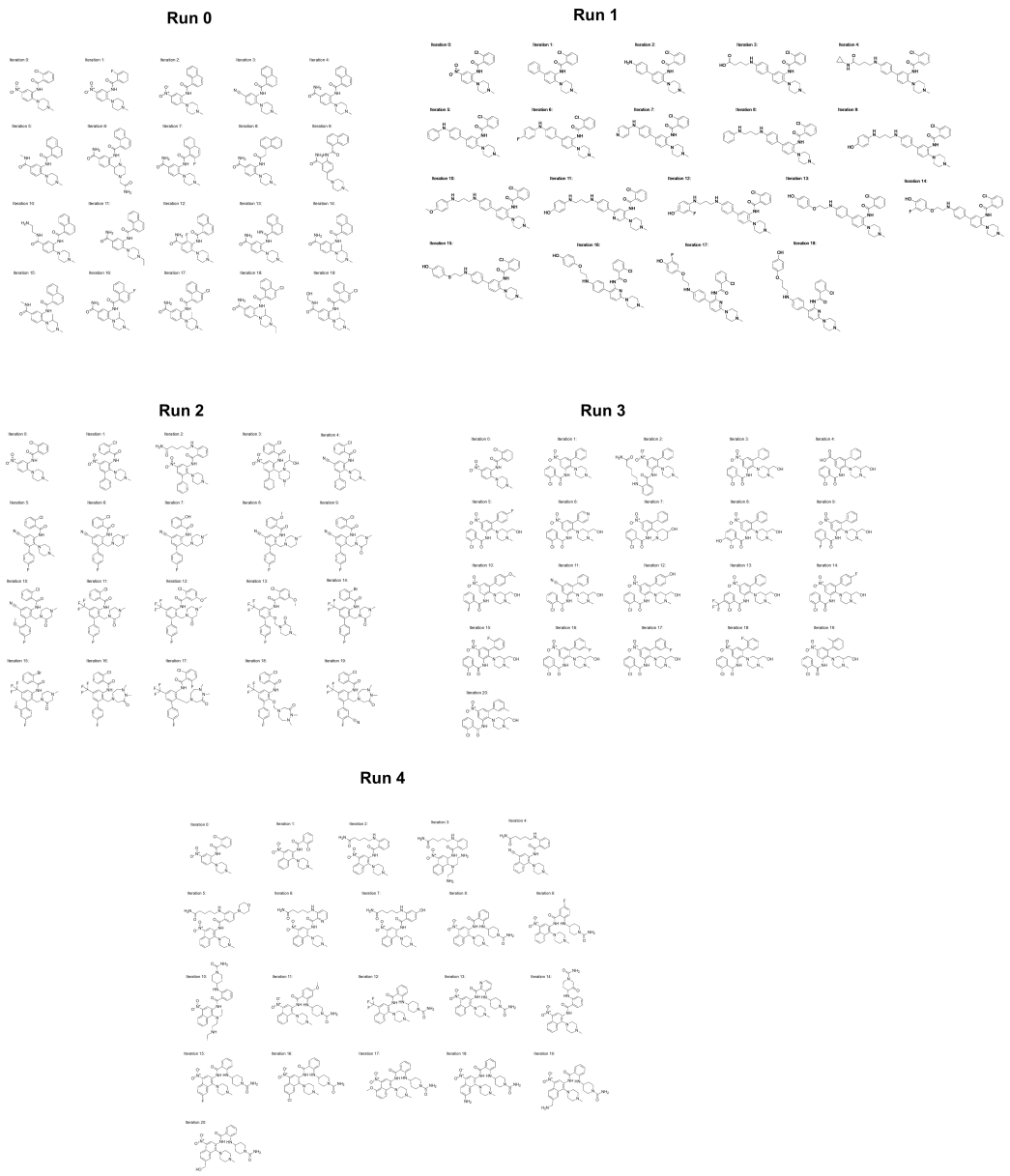

Figure S5: Visualization of the 100 molecules generated in the closed-loop inverse design of drug molecules with high binding affinity against WDR5. The agent is powered by Claude 3.5 Sonnet. No molecule was found to be invalid or contain potentially unstable functional groups.

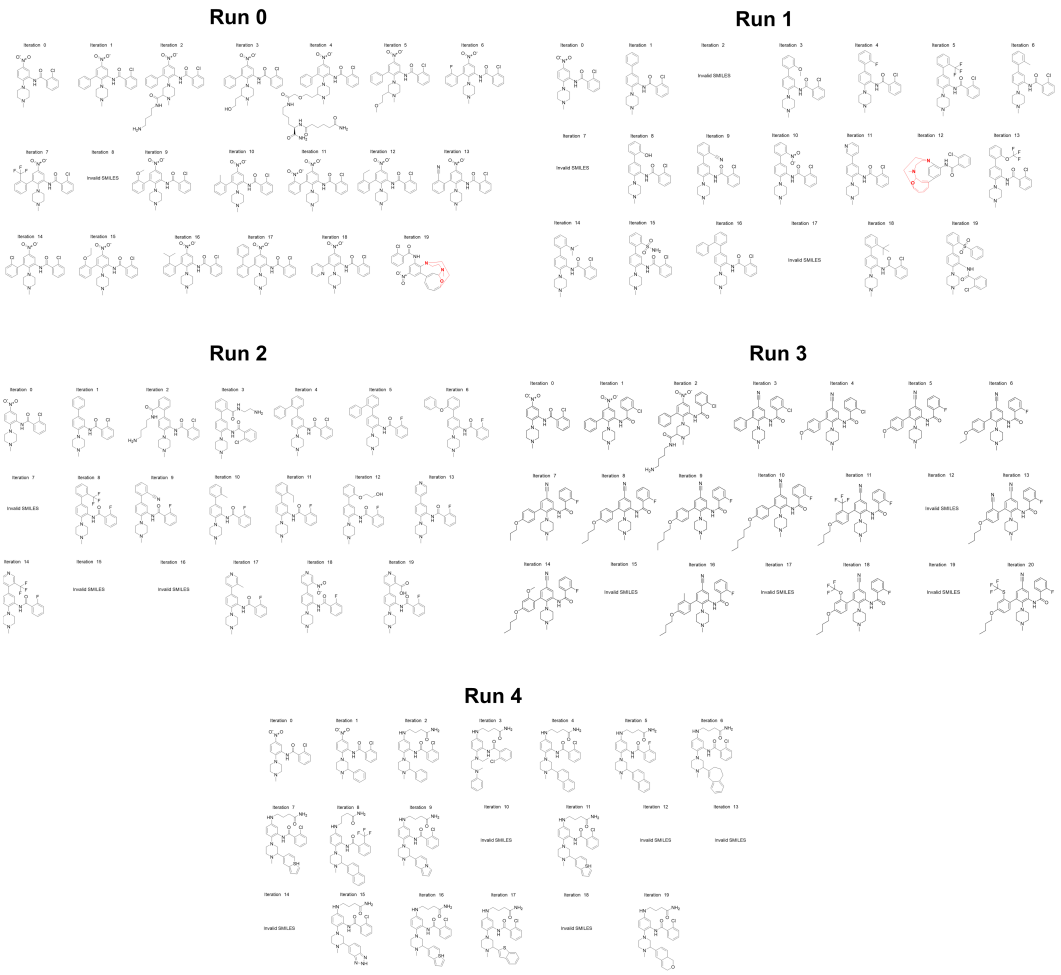

Figure S6: Visualization of the 100 molecules generated in the closed-loop inverse design of drug molecules with high binding affinity against WDR5. The agent is powered by GPT-4o. Color red indicates a potentially unstable functional group. Invalid SMILES generated are marked as invalid.

6.1 Human-in-the-loop design with GPT-4o

As in closed-loop analysis (section 6), the model was provided an initial structure. Examples of initial modifications include adding hydrophobic groups or optimizing solvent-exposed regions to the same starting molecule. After each iteration, the human reviewed the changes, provided specific feedback—such as focusing on different molecular regions or improving solubility—and suggested further modifications like adding functional groups or tweaking $\pi - \pi$ stacking interactions. Accepted changes, like adding a phenyl or trifluoromethyl (-CF_3) group, increased binding affinity and were incorporated, while less effective modifications were rejected based on the docking score and human input. Overall, dZiner was cooperative and could be used by medicinal chemists with no prior AI experience. It was able to enable requested most changes, including instructions to revert to prior iterations. Compared to the closed-loop optimization, human-in-the-loop experiments were able to produce diverse structures with improved WDR5 docking scores.

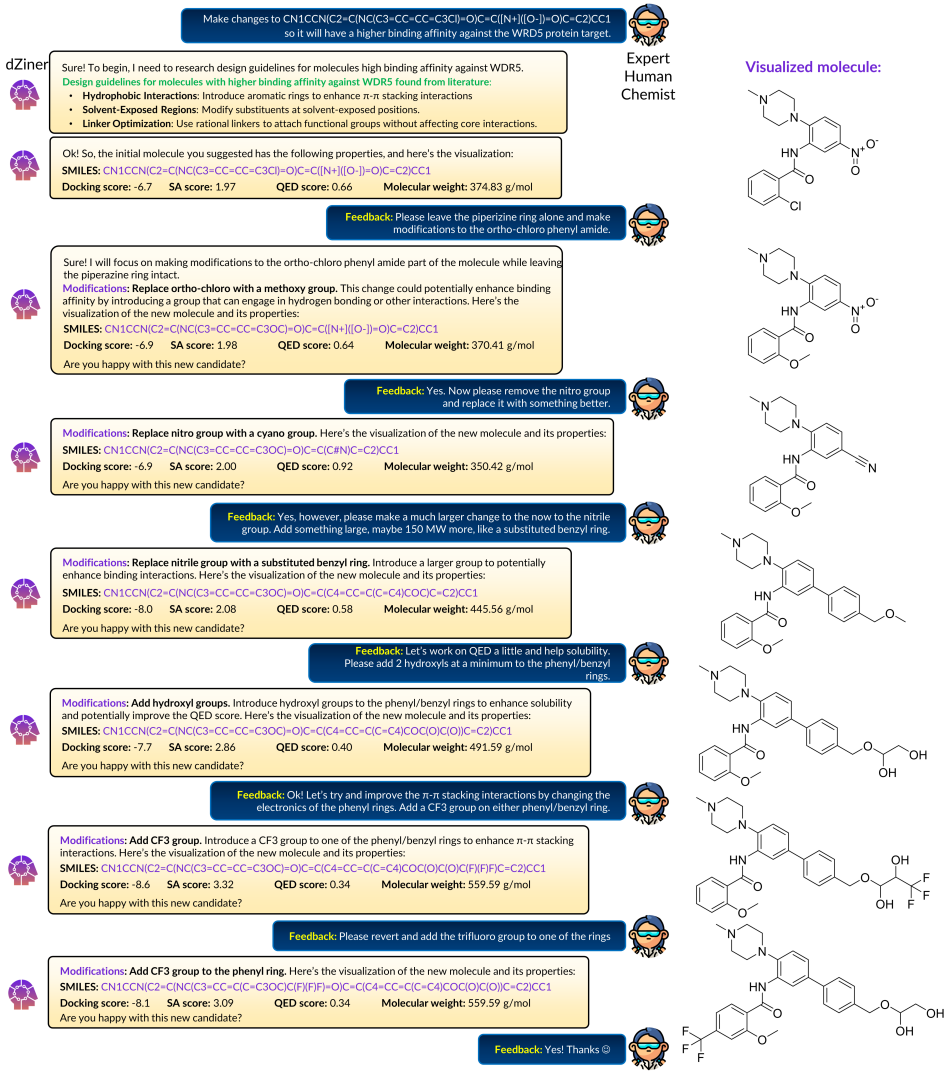

Figure S7: **Human-in-the-loop inverse design of a drug candidate against WDR5 protein target**. Agent is powered by GPT-4o. dZiner was found to be highly cooperative, interpretable, and able to enact changes requested with ease in this human-AI collaboration.

7 MOF Organic Linker Design and CO_2 Adsorption Capacity Inference with GPT-4o

Similar to Section 2.3, and starting from the same initial organic linker, we applied dZiner powered by GPT-4o to enhance CO_2 adsorption capacity (see Figure S8). In iteration 1, the introduction of hydroxyl (-OH) and amino (-NH_2) groups improved the adsorption to 0.992. However, in iteration 2, the addition of a sulfonate group resulted in a slight decrease in performance. Significant improvements were achieved in iteration 7 by incorporating a pyridine ring, which increased nitrogen interactions and boosted CO_2 adsorption to 1.278. The highest adsorption, 1.644, was observed in iteration 8 when a fluorine atom was introduced, leveraging its high electronegativity to enhance CO_2 capture. A chlorine atom was also added in iteration 9, resulting in a CO_2 adsorption of 1.409. Overall, the combination of electronegative atoms and nitrogen-containing functional groups proved most effective in enhancing CO_2 adsorption. Throughout the optimization, the molecular weight increased from 430.424 g/mol (iteration 0) to 516.945 g/mol (iteration 9). The SA score also fluctuated, peaking at 4.595 in iteration 5 after the addition of hydroxyl groups, indicating increased synthetic complexity.

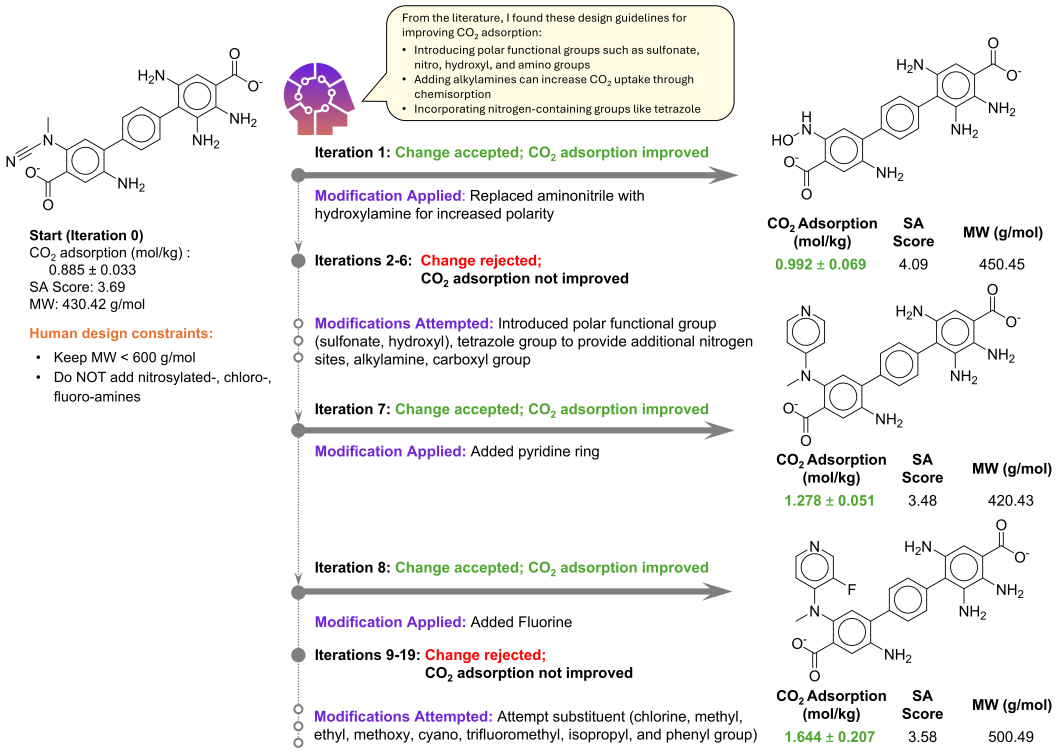

Figure S8: **dZiner's chain-of-thoughts in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity**. The agent is powered by GPT-4o. Design guidelines were retrieved from scientific literature (same as in Figure 4). The model is asked to keep the molecular weight lower than 600 (g/mol), and *not* to add nitrosylated, chloro-, fluoro- amines to the molecule in natural language text. The accepted molecule bears 63

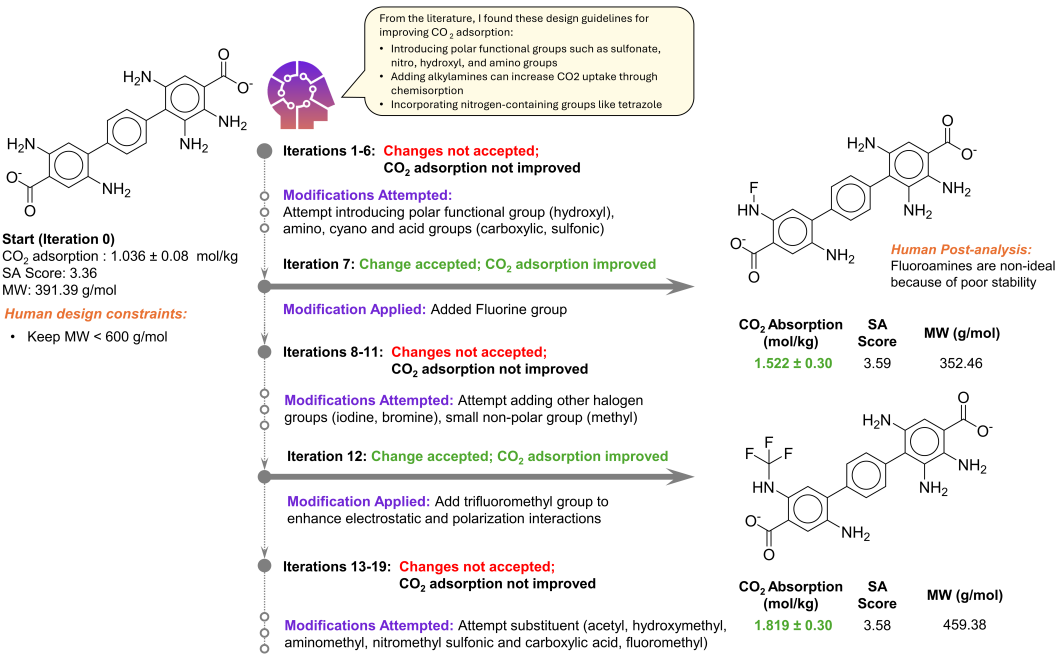

Figure S9: **dZiner's chain-of-thoughts in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity**. The agent is powered by GPT-4o. Design guidelines were retrieved from scientific literature, same as in Figure 4. CO_2 adsorption capacity is improved by 75% via iterative agent-suggested chemical modifications, while following the molecular weight design constraint (MW < 600 g/mol).

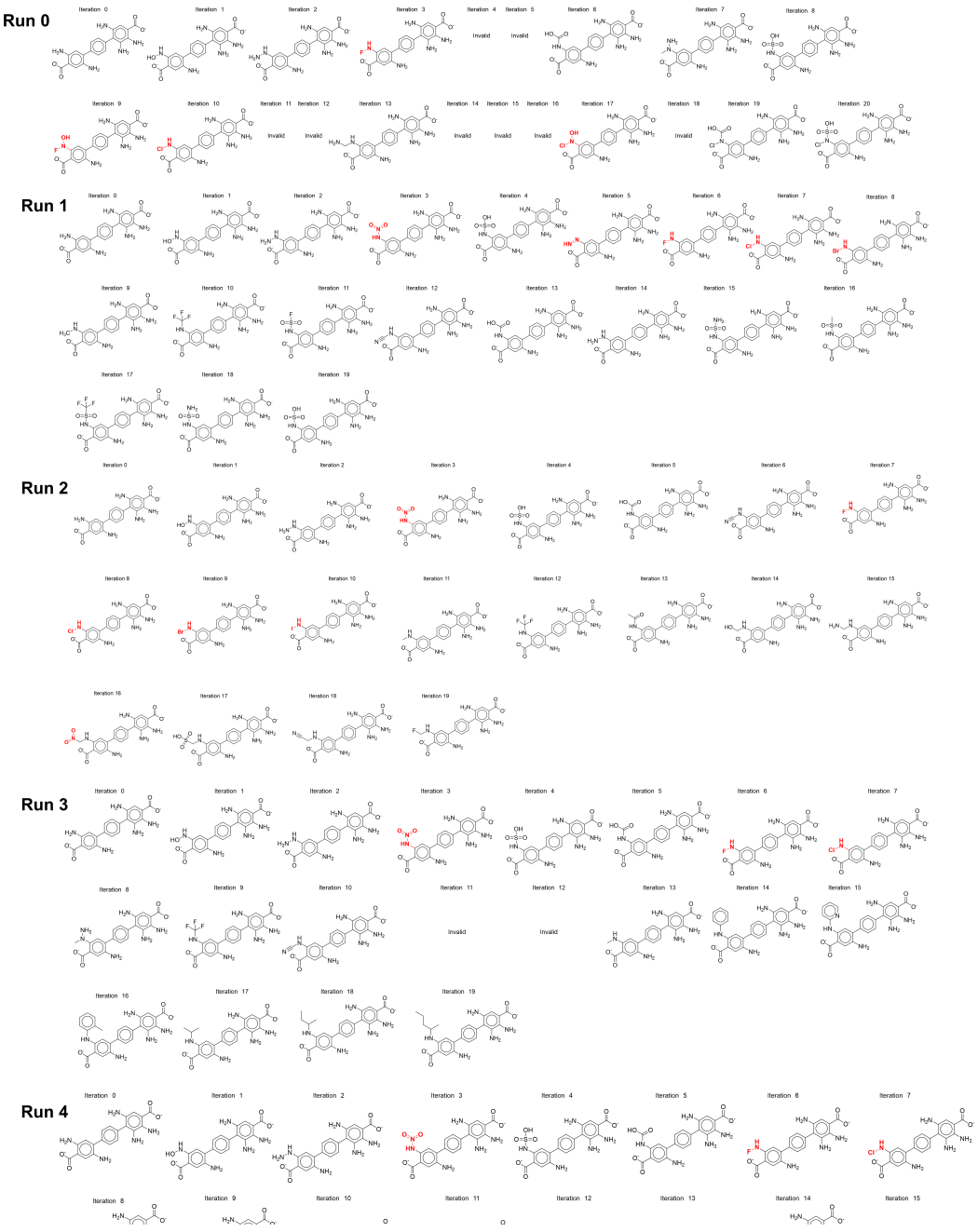

Figure S10: Visualization of the 100 molecules generated in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity (molecular weight design constraint only case study). The agent is powered by GPT-4o. Color red indicates a potentially unstable functional group. Invalid SMILES generated are marked as invalid.

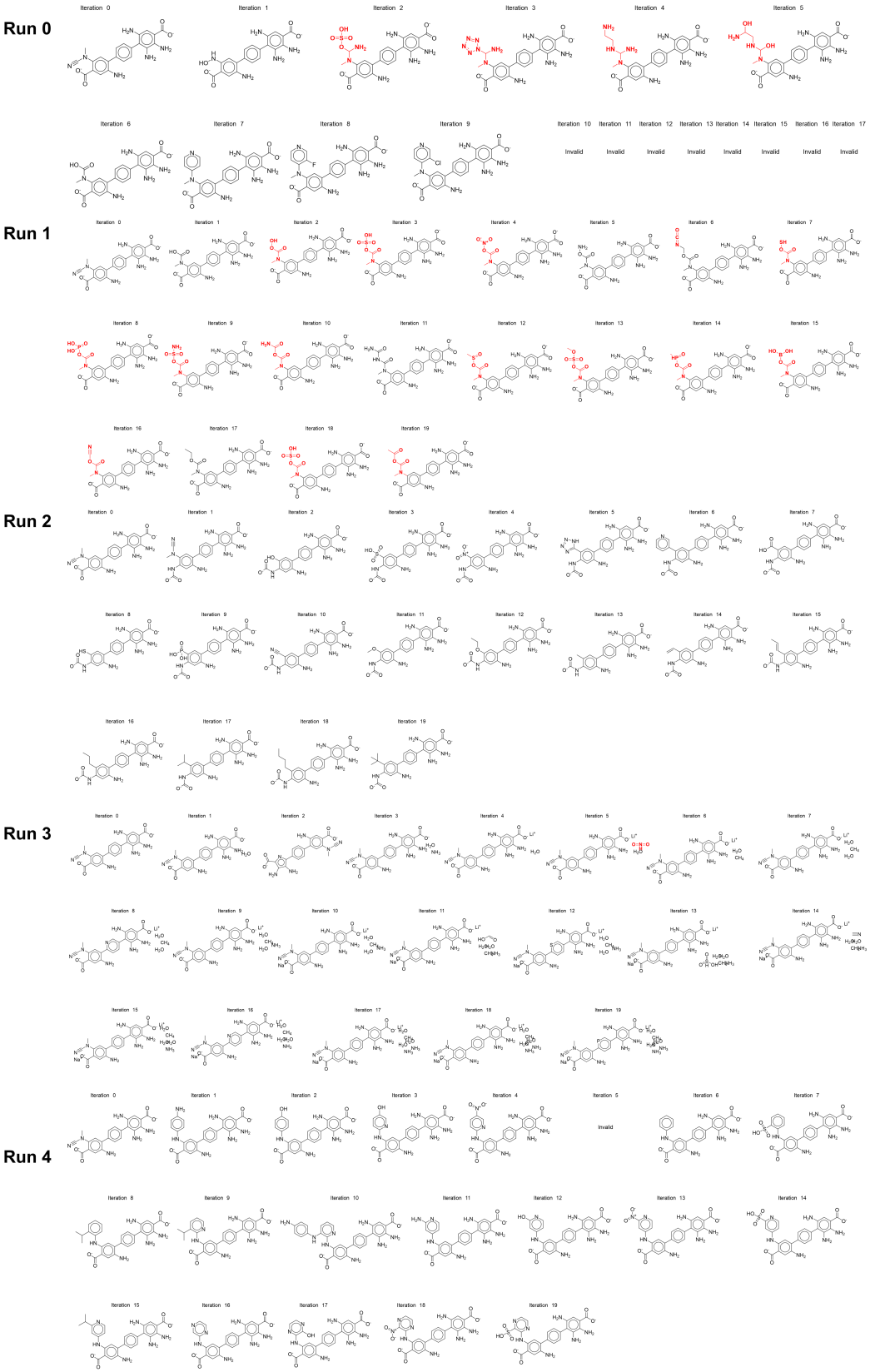

Figure S11: Visualization of the 100 molecules generated in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity (molecular weight and functional groups design constraint case study). The agent is powered by GPT-4o. Color red indicates a potentially unstable functional group. No invalid SMILES were generated.

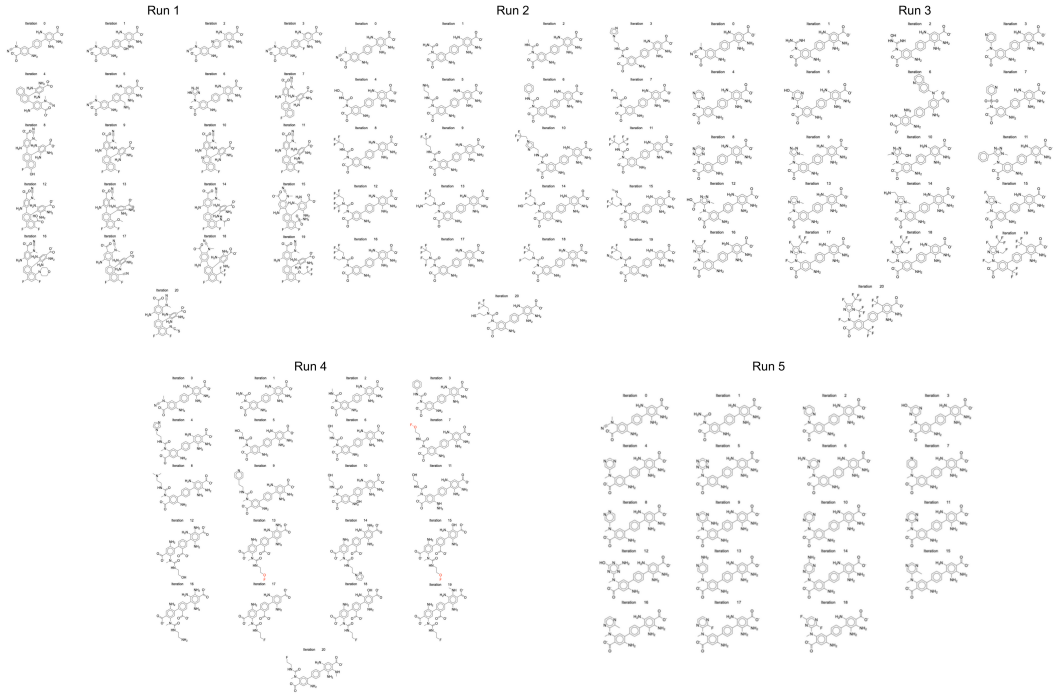

Figure S12: Visualization of the 100 molecules generated in the closed-loop inverse design of organic linkers for MOFs with high CO_2 adsorption capacity (molecular weight and functional groups design constraint case study). The agent is powered by Claude 3.5 Sonnet. Color red indicates a potentially unstable functional group. invalid SMILES generated are marked as invalid.