Large Language Models as Molecular Design Engines

Debjyoti Bhattacharya,[†] Harrison J. Cassady,[‡] Michael A. Hickner,[‡] and Wesley F. Reinhart[*,†,¶]

†[1]Materials Science and Engineering, Pennsylvania State University, University Park, PA, USA

‡[2] Department of Chemical Engineering and Material Science, Michigan State University, East Lansing, MI, USA

¶[3]Institute for Computational and Data Sciences, Pennsylvania State University, University Park, PA, USA

E-mail: reinhart@psu.edu

Abstract

The design of small molecules is crucial for technological applications ranging from drug discovery to energy storage. Due to the vast design space available to modern synthetic chemistry, the community has increasingly sought to use data-driven and machine learning approaches to navigate this space. Although generative machine learning methods have recently shown potential for computational molecular design, their use is hindered by complex training procedures, and they often fail to generate valid and unique molecules. In this context, pre-trained Large Language Models (LLMs) have emerged as potential tools for molecular design, as they appear to be capable of creating and modifying molecules based on simple instructions provided through natural language prompts. In this work, we show that the Claude 3 Opus LLM can read, write, and modify molecules according to prompts, with an impressive 97% valid and unique molecules. By quantifying these modifications in a low-dimensional latent space, we systematically evaluate the model's behavior under different prompting conditions. Notably, the model is able to perform guided molecular generation when asked to manipulate the electronic structure of molecules using simple, natural-language prompts. Our findings highlight the potential of LLMs as powerful and versatile molecular design engines.

Introduction

The design of novel molecules and materials remains an important frontier for the scientific community, with new synthetic approaches being developed all the time. Such efforts are crucial across a wide array of applications, including energy storage technologies,[1] alloy design,[2] 2D materials design,[3] and drug discovery.[4] The strategic navigation of this vast chemical space is critical for successful discovery of new material solutions to these challenging problems.[5] Generative machine learning models[6] have been at the forefront of this exploration, offering a glimpse into the future of computational design.

Despite their promise, these models often stumble[7] by producing invalid or irrelevant molecular structures. Furthermore, fine-tuning and retraining these models demand substantial labeled data at times, complicated training procedures, intensive computational resources, and a significant amount of time, making the process costly and sometimes impractical for many tasks. Adding to these hurdles, acquiring training data for these models is challenging, as data must be sourced from disparate materials databases[8] with potentially different formats. These procedural issues further contribute to the challenges of using chemistry-specific generative models to transform raw data into actionable insights for materials discovery.

Nevertheless, the advent of generative AI models[9] marks the beginning of a paradigm shift in the discovery and design of new materials. Among the most promising developments are Large Language Models (LLMs),[9] initially trained on vast amounts of natural language data, now recognized for their disruptive effect in nearly every field. While not explicitly trained to be knowledgeable in chemistry, these models have the advantage of being adaptable and generalizable. The Simplified Molecular Input Line Entry System (SMILES) encodes molecular structures in text form.[10] Thus, SMILES strings can be leveraged to explore LLMs' understanding of cheminformatics and chemistry design principles. LLMs have the additional benefit of being very flexible with the formatting of input data, meaning that some of the challenges associated with chemistry-specific generative models may be circumvented with LLMs.

In this work, we explore the Claude 3 Opus LLM's ability to understand and leverage chemical design rules to perform molecular generation and modification tasks. Through systematic study with quantitative metrics, we offer insights into how well LLMs can design new molecules and navigate the chemical design space in different design scenarios. By leveraging a latent space embedding of the molecules, we perform a nuanced investigation of the molecular modifications applied by the LLM. Additionally, we explore the biases that emerge with different prompts to understand how these prompts will affect the navigation of the chemical space. Through these tasks, we aim to demonstrate systematically that simple, natural language instructions can enable LLMs to generate new molecules with specific characteristics.

Methods

Dataset and representation learning

The dataset for this study includes approximately 1.3 million small molecules from the ZINC database.[11] In total, ZINC contains over 230 million commercially available molecules frequently used in virtual screening for drug discovery. Here we used a subset of small molecules (molecular weight below 200 Daltons) that contain nitrogen and at least one

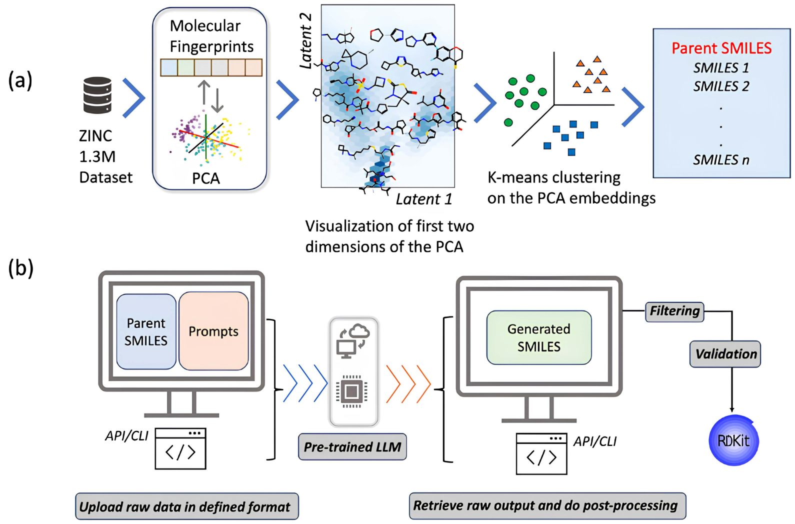

Figure 1: (a) represents the process of parent SMILES generation for the molecular modification process using Claude API. (b) represents the Claude API workflow and how unique validated (by RDKit) SMILES are obtained.

hydrogen bond donor or acceptor, targeting molecules that facilitate proton transport for applications in energy storage technologies.

We employed the counts-based Morgan Fingerprint strategy[12] to featurize the small molecules. This approach involves generating molecular fragments of each molecule, using these fragments as keys, and assigning numbers based on the frequency of occurrence of different substructures to derive a vector of integers representing each molecule. Subsequently, Principal Component Analysis (PCA) was performed to generate a three-dimensional latent embedding for these molecules. Once computed, this permits mapping all possible molecules into continuous coordinates. However, the coordinates are most meaningful for molecules similar to those that produced the PCA (i.e., small organic molecules found in ZINC).

We chose 64 parent molecules via K-means clustering (implemented in `scikit-learn`[13]) on the PCA embeddings to evaluate LLM performance on a diverse group of molecules. These so-called "parent molecules" were then represented by their canonical SMILES for molecular modification via the LLM; the embeddings were only used to quantify relationships between molecules before and after modification. The dataset, embedding scheme, and parent selection process are illustrated schematically in Figure 1a.

LLM interactions

This work utilized Anthropic's Claude 3 Opus model,[14] a state-of-the-art LLM. Interaction with the LLM is facilitated by the Anthropic Python SDK,[15] where requests containing task instructions (prompts) are processed by the pre-trained model on Anthropic's server. We set `temperature=0` so the model always favors the most probable token outputs. This results in generations that are more deterministic and focused, exhibiting less randomness or diversity. However, even with `temperature=0`, the outputs are not entirely deterministic due to the inherent stochasticity of the model's sampling process. The maximum tokens parameter was set to `max_tokens=1024`, which restricts the length of the generated output since we only asked for candidate molecules (represented by relatively compact SMILES) and not any explanations.

The pre-trained model generates SMILES responses for each of the 64 parent molecules based on different prompts. Responses generated by the model are then transmitted back through the Application Programming Interface (API) and post-processed at the requester's end. A simple workflow of the API is shown in Figure 1b.

Base prompts

The following system prompt was provided to the model for every query:

> You are a chemoinformatics expert that can generate new molecules. Please provide only the Python formatted list of SMILES strings, like [SMILES1, SMILES2, SMILES3] without any additional explanations or text.

Additional information was provided in the following format:

> Given the molecule with SMILES representation '`smiles`', generate `n` molecules that are `prompt_detail`. Respond with just the SMILES strings as elements of a Python list.

In the above, `smiles` was replaced with the parent SMILES, `n` with the target number of candidates (10), and `prompt_detail` with a specific task as described below.

The task for the LLM was to generate 10 molecules that adhere to the criteria specified in the accompanying prompt descriptions, as given in Table 1. The model was then instructed to return the SMILES strings of these molecules in a Python list format. To develop these prompts, we utilized a "prompting for prompts" approach, engaging Claude-3 Opus to suggest eight distinct prompts for inducing either minor (fine) or major (coarse) modifications to a given molecule's SMILES representation. Fine prompts were characterized by the phrase "similar molecules", whereas coarse prompts were distinguished by the phrase "completely different molecules."

One potential benefit of incorporating LLM feedback in this meta-task is their capacity to complement human expertise by offering a different perspective. Unlike human experts, who may be constrained by a finite set of known molecular generation rules, LLMs can leverage their extensive training on diverse datasets to propose innovative approaches and solutions. This capability enables them to identify potential design and modification opportunities that might not be immediately evident to human experts. Furthermore, LLMs might facilitate the exploration of the full spectrum of chemistry design rules (based on the very large corpus of documents seen in their training), a task that could require the combined efforts of multiple experts.

Prompts for guided generation

Beyond the base prompts described above, we consider more detailed prompts that specify modifications to the electronic structure of the molecules. Specifically, we ask for Electron Donating Groups (EDGs) or Electron Withdrawing Groups (EWGs) to be incorporated into the generated molecules. This represents a crucial advantage of the LLM-based approach since natural language can be used to express these details, while conventional methods would require crafting substitution rules by hand, and other generative methods would require

Table 1: Detailed sub-prompts used to describe how the molecular modification task should be carried out.

Identifier	Prompt detail text
A	similar molecules by changing one or two atoms or bonds to produce closely related structures
B	similar molecules by tweaking only the side chains
C	similar molecules with minimal structural changes to find similar but new candidates
D	similar molecules with slight variations on functional groups while maintaining the backbone structure
E	completely different molecules by changing multiple atoms or bonds
F	completely different molecules by significantly altering the core structure and introducing completely new functional groups
G	completely different molecules that significantly vary in size and functional groups
H	completely different molecules with significant structural changes to find new candidates

either consideration of this requirement at training time to perform conditional sampling or use a very inefficient sampling at inference time to identify candidates that match the requirements. The prompts are based on the fine base prompts above (A-D) and are displayed in Table 2.

Validation and metrics

After receiving candidate SMILES strings from the LLM, they are validated by RDKit,[16] an open-source cheminformatics toolkit. We first ensure the strings represent valid molecules and then convert them to canonical form, as SMILES are not bijective mappings. The canonical SMILES undergo further filtering, eliminating duplicates and removing instances where the unmodified parent appears within the generated set. This ensures that unique and valid SMILES appear in the list of generated molecules (and only these are considered in the evaluation metrics). We evaluated the resulting molecules primarily using three metrics: Euclidean distance, validity ratio, and chemical diversity.

We calculated the Euclidean distance between parents and children for each prompt using

Table 2: Detailed sub-prompts used to describe how the molecular modification task should be carried out, in the specific case of adding electron-donating groups and electron-withdrawing groups.

Identifier	Prompt detail text
I	Similar molecules by changing one or two atoms or bonds to produce closely related structures focusing on incorporating electron donating groups (EDGs) to find new candidates
J	Similar molecules by tweaking only the side chains to produce closely related structures focusing on incorporating electron donating groups (EDGs) to find new candidates
K	Similar molecules with minimal structural changes to produce closely related structures focusing on incorporating electron donating groups (EDGs) to find new candidates
L	Similar molecules with slight variations on functional groups while maintaining the backbone structure to produce closely related structures focusing on incorporating electron donating groups (EDGs) to find new candidates
M	Similar molecules by changing one or two atoms or bonds to produce closely related structures focusing on incorporating electron withdrawing groups (EWGs) to find new candidates
N	Similar molecules by tweaking only the side chains to produce closely related structures focusing on incorporating electron withdrawing groups (EWGs) to find new candidates
O	Similar molecules with minimal structural changes to produce closely related structures focusing on incorporating electron withdrawing groups (EWGs) to find new candidates
P	Similar molecules with slight variations on functional groups while maintaining the backbone structure to produce closely related structures focusing on incorporating electron withdrawing groups (EWGs) to find new candidates

the PCA latent space embeddings described above. This process involves mapping each molecule to the latent space. The distance metric quantifies the molecular transformation induced by the prompt. Thus, the distance d_Z for a given prompt can be expressed simply as $d_Z = ||z_p - z_c||$, where z_p is the latent space coordinate of the parent molecule, and z_c is that of the child. This distance provides insights into the magnitude of change between generated molecules and their parents, elucidating the impact of the specific prompts on molecular evolution within this space. Note that there is no clear preference for a particular value of d_Z. Rather, it is descriptive, with low values indicating small changes to the molecule and high values indicating significant changes.

Defined as the proportion of chemically valid and unique structures among the generated molecules (excluding any parent SMILES), the validity ratio v is calculated simply as $v = N_{\text{valid}}/N_{\text{gen}}$, where N_{valid} is the number of valid, *unique* molecules obtained from the LLM call (excluding parent SMILES) and N_{gen} is the total number of raw SMILES generated by the LLM before any filtering or validation. As described above, RDKit verifies the validity of SMILES strings, and any duplicates or parent SMILES are removed from the generated SMILES before calculating v. This ratio describes the model's efficiency in producing chemically valid and novel structures from specified inputs and should ideally be close to 1.

Chemical diversity δ_{chem} quantifies the heterogeneity among unique and chemically valid generated molecules (after filtering and validation) and is calculated as:

$$\delta_{\text{chem}} = 1 - \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} T(c_i, c_j), \tag{1}$$

where $T(c_i, c_j)$ represents the Tanimoto similarity between the molecular fingerprints of molecules i and j, and N is the number of molecules considered in the calculation. This formula inverts the average Tanimoto similarity across all unique pairwise combinations into a measure of diversity, with a higher score indicating greater chemical diversity within the set.

There is no clear preference for a particular value of δ_{chem} since there is a trade-off between exploration and exploitation here, as with d_Z. Very low δ_{chem} indicates that the generated molecules are nearly identical so that no significant modification was obtained, while very high δ_{chem} indicates that the molecules are totally different, so the modifications are not semantically meaningful.

Electronic structure calculations

The highest occupied molecular orbital (HOMO) energies were computed using the PM7 semi-empirical quantum chemical method, as implemented in the MOPAC[17,18] program. The MOPAC calculations were performed through the Python interface provided by the Atomic Simulation Environment (ASE) package.[19] The RDKit library was employed to generate the initial 3D molecular structures, and geometry optimizations were carried out using the Universal Force Field (UFF) to obtain the most stable conformers.

The following set of keywords was employed in the MOPAC calculations, in addition to the PM7 method, to achieve an optimal balance between computational efficiency and accuracy: `PRECISE`, `GNORM=0.001`, `SCFCRT=1.D-8`, `DISPERSION=D3H4`, `H-PRIORITY`, `AUX`, and `ITRY=200`. These keywords enforce stricter convergence criteria for the self-consistent field (SCF) procedure, include long-range dispersion corrections, prioritize the treatment of hydrogen atoms, enable auxiliary basis functions, and increase the maximum number of SCF iterations.

For each optimized molecular structure, the HOMO energies were extracted from the MOPAC output files. While the semi-empirical level of theory may not provide the same accuracy as higher-level quantum chemical methods, the HOMO energies obtained from these calculations can serve as a surrogate for more sophisticated calculations and provide insights into the electronic structure of the generated molecules with a minimal computational footprint.

Results and Discussion

Representative examples

Overall, all prompts generated reasonable sets of children molecules compared to the parent molecule depending on the prompt instructions. Prompts A through D yielded child

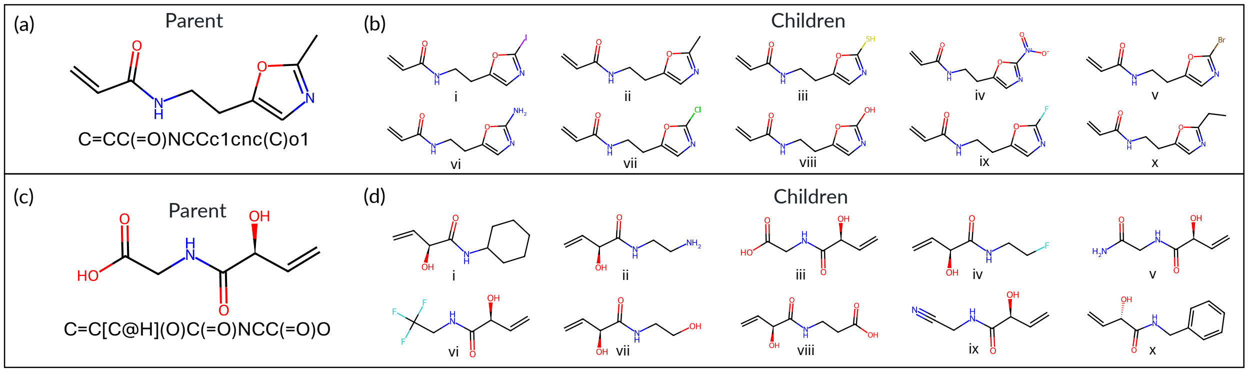

Figure 2: A representative selection of molecules generated by the prompt D ("similar molecules by changing one or two atoms or bonds to produce closely related structures") for two different parent molecules. The LLM accepts the parents (a, c) and returns 10 children (b, d) as SMILES; the molecules are only rendered here to improve human readability.

molecules with slight variations compared to the parent while retaining the overall shape of the parent framework and installing slight variations in functional groups. Prompts E through H, which were instructions for producing children molecules that were different from the parent, yielded child molecules that were a departure from the parent regarding functional groups and framework connectivity without unreasonable departures in terms of molecular size or spurious inclusion of exotic atoms or functional groups. Specifically, prompt E yielded child molecules different from the parent but seemed to have some visual relation to the parent in all 64 cases. Prompts F, G, and H gave children that departed further from the parent molecule than prompt E.

Representative examples of molecules generated by sub-prompt D are shown in Figure 2. In the case of both parents, all 10 generated children are valid molecules (i.e., as verified by RDKit). Note that the LLM outputs text which is later converted to images via RDKit, so the relationships between the children are more subtle than they appear when rendered as images. In the case of prompt D, the "backbone" appears to be interpreted by the LLM as the center of the molecule, which is never modified. Instead, functional groups are attached to the methyl group on the right-hand side of parent (a) and to the left-hand side of parent (c). For parent (c), this includes the addition of some bulky rings and sometimes the truncation of the carbonyl group at the left, which the LLM does not always interpret to be part of the "backbone."

It is interesting to note that the right-most methyl group of parent (a) does not correspond to the last character in the SMILES, but instead to the `(C)` token, which appears third from last in the sequence. This illustrates a non-trivial understanding of the SMILES syntax by the LLM, which must "imagine" the spatial structure of the molecule to some degree to identify suitable substitution sites.

A Python-based viewer for displaying the input molecules and the generated output molecules for each prompt was developed and included in our Zenodo repository.[20]

Molecular fingerprint latent space

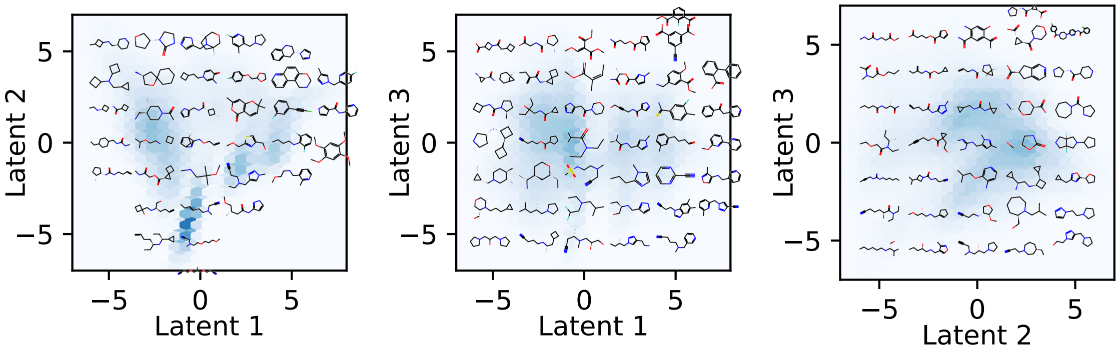

Figure 3: The latent space obtained from featurizing small molecules in the ZINC database with counts-based Morgan Fingerprints and embedding with PCA. Each panel shows a different 2D slice of the embedding, up to three components. Darker colors indicate a higher density of molecules occurring in that cell. Representative molecules are rendered near their 2D embedding.

To quantify the behavior of the LLM when making modifications to molecules, we generate a latent space embedding of molecules based on Morgan fingerprints. The embedding thus yields a three-dimensional coordinate z that describes the molecules by a quantitative feature vector. Representative molecules are rendered throughout this latent space in Figure 3. Generally, latent dimension 1 appears to be related to unconjugated rings at low values and conjugated rings at high values. Latent dimension 2 appears to be related to the prevalence of cycles, with linear or chain-like molecules appearing at low values and ring-containing molecules at high values. Finally, latent dimension 3 appears dominated by ketones, with molecules at high values carrying two or more groups.

Base prompt performance

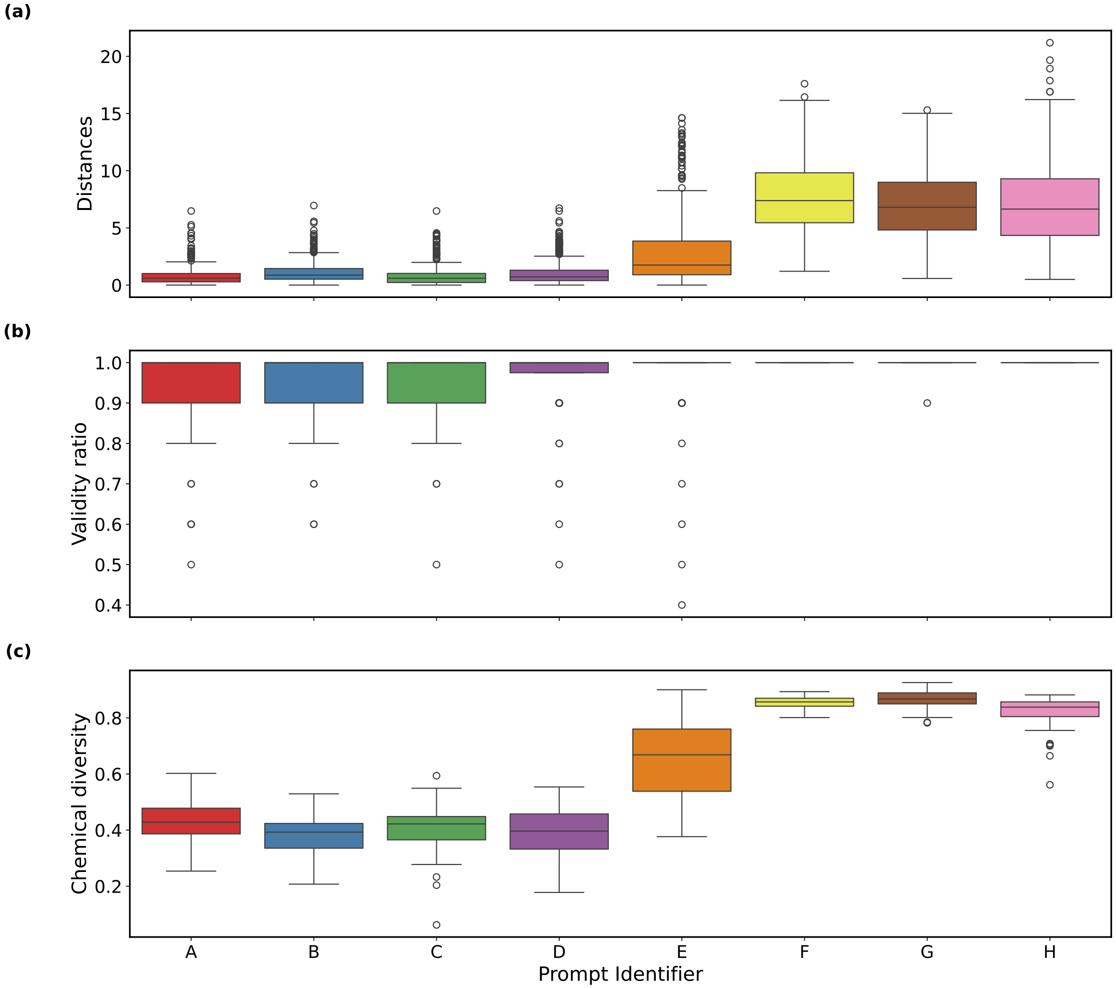

Figure 4: Metrics evaluated on each sub-prompt from Table 1 when the molecular modification task is performed on the same 64 parent molecules. (a) Euclidean distance d_z, (b) validity ratio v, and (c) chemical diversity δ_c, as described in the text.

To systematically evaluate the impact of different prompts on molecule generation, we employed the following metrics (defined above): distance d_Z, validity ratio v, and chemical diversity δ_c. This evaluation yields several key observations regarding the impact of prompt engineering on molecular structure generation.

Analyzing Euclidean distances within the latent space reveals that fine prompts generally result in smaller distances than coarse prompts (see Figure 4a). This observation aligns with the expectation that coarse prompts induce more significant alterations in molecular structures (i.e., via the phrase "completely different molecules" instead of "similar molecules"). However, the extent of these changes varies depending on the specific modification mechanism described within each sub-prompt, emphasizing the importance of prompt engineering in steering the LLM behavior. Among the fine prompts, all variants exhibit similar distances, with prompts B and D showing slightly higher median distances in the latent space than A and C. Coarse prompts F-H exhibit comparable distance distributions, with prompt F having the highest median distance across all prompts. However, prompt E is significantly lower, behaving more like the fine prompts, probably due to the more specific language "atoms or bonds" which implies local changes only, as opposed to more global language in the other sub-prompts.

We found it surprising that the distances d_z were so similar within the groups A-D and F-H, despite very different language specifying how the changes should be made. This indicates that the magnitude of the variation (e.g., "similar" or "completely different") can be somewhat decoupled from the mechanism by which the modification is enforced (e.g., "…by tweaking only the side chains"). If it holds in general, this behavior will make LLM-driven molecular modification an extremely versatile tool for materials design in the future.

The median validity ratio for most prompts remains at more than 0.9, indicating a high rate of chemically valid and unique molecules different from the parent molecules, as shown in Figure 4b. This stands in contrast to generative models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which have reported validity ratios as low as 0.25[21] for the best-performing models. Furthermore, generative models are susceptible to mode collapse, resulting in only $\tilde{2}$% unique molecules.[22] Thus, this LLM-based approach consistently generated more valid and unique molecular structures, even without any fine-tuning.

While it may seem counterintuitive, the fine prompts exhibit a lower validity ratio because they ask the model to produce similar molecules by making only small, localized changes. In contrast, the coarse prompts instruct the model to propose completely different molecules so the model can select molecules that are somewhat unrelated to the parent but are known to be valid molecules. The LLM's occasional failure to generate valid and unique molecules may be attributed to the discrete SMILES representation, which allows many ways for the model to construct invalid strings.

The LLM's ability to consistently return valid and canonical SMILES with a simple molecular representation like SMILES demonstrates its robustness and reliability in generating molecular structures. However, many factors may contribute to the occasional failure to generate valid molecules. One issue is the phenomenon of hallucination by LLMs,[23] which arises from sometimes competing objectives to balance following detailed instructions in the prompt and obeying grammatical rules for SMILES. Additionally, generating valid molecules in a one-shot setting, as done in our work, is more difficult than a few-shot approach, where the model can iteratively refine its predictions. Another factor contributing to imperfect validity rates is the occasional copying of the parent SMILES when asked to generate new molecules or repeated generated of the same SMILES in one query, leading to duplicate structures. These duplicates and copies of parent SMILES are eventually removed during filtering, decreasing the overall validity ratio.

The chemical diversity metric in Figure 4c indicates that coarse prompts generally lead to higher δ_c than fine prompts. In particular, prompts F-H exhibit the highest levels of chemical diversity across all prompts, with G exhibiting the highest overall. This matches the trend in d_z from Figure 4a since more significant changes are made to the molecules, which can induce higher diversity. Among the fine prompts, A yields a slightly higher δ_c than the others. The diversity metrics for similar prompts show that the generated molecules have a suitable range of structural variations and do not typically exhibit mode collapse. This indicates that the navigation paths through the chemical space, guided by the prompts, yield molecules with distinct structural features, even among prompts within the same category.

Additional performance metrics, such as the duration of the API call, were evaluated and provided in the supplemental information to assess the LLM-based modification scheme's efficiency. The median response time was 10.4 s (to generate 10 molecules), with a mean of 11.5 s and standard deviation of 4.0 s, indicating a long tail. The use of these evaluation metrics is consistent with previous research[24] that has employed similar measures to assess molecule generation tasks.

Evaluating bias

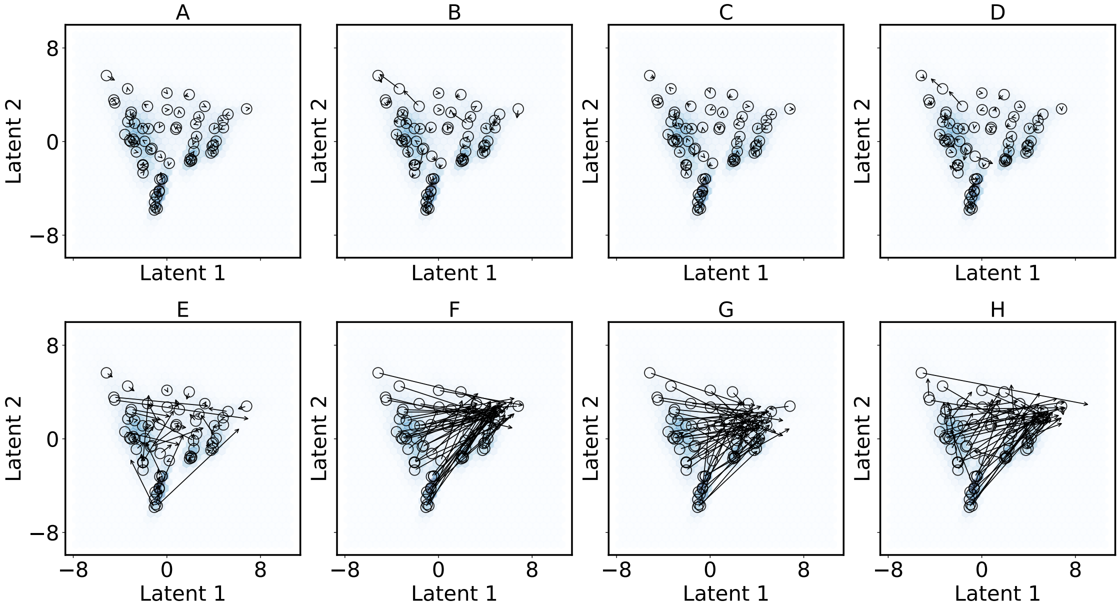

Figure 5: The average displacement in z from all 10 child molecules for each sub-prompt in Table 1. This illustrates prompt-specific bias in molecule generation tasks. Each circle indicates the parent coordinate, while the arrows indicate the average displacement between that parent and its children. Only the first two principal components of the PCA embeddings are shown for simplicity, even though the distances were calculated in 3D.

To further assess the behavior of each sub-prompt, we show the average displacement between parent and child molecules within the latent space z in Figure 5. Herein, we use arrows to indicate the average displacement in z between parent and child molecules. We consider this a bias since the collective directional change from a parent molecule to several

($\leq$ 10) children should be close to zero if the molecules are equally likely to go in any direction. Thus, the ideal result would be an arrow with nearly zero length, indicating that the modification direction is entirely random. The comparison across prompts A, B, C, and D (referred to as fine prompts) and E, F, G, and H (coarse prompts) reveals a notable distinction in movement magnitude within the latent space, with fine prompts generally leading to subtler shifts compared to coarse prompts.

A closer examination of the individual prompts within the fine and coarse categories reveals distinct patterns and tendencies attributable to specific prompts. Although prompts E, F, G, and H were all classified as coarse prompts intended to generate more significant modifications, prompt E exhibited less bias and directional tendencies compared to prompts F, G, and H. This suggests that there is a divergence in how these prompts navigate the chemical space or explore potential molecule modifications. Notably, prompts F and G appeared to be steered towards generating molecules represented in the upper right region of the latent space visualization. This indicates a potential preference or increased efficiency in fulfilling the prompt instructions within that particular area of the chemical space.

The movement patterns observed in prompts H and E reveal intriguing insights into how the LLM interprets prompts. Prompt H, involving 'significant' structural changes, results in more directional movements and biases than E, though less biased than F and G, as it explores multiple regions rather than concentrating on one area like the upper left. This aggressive, multi-directional movement can be attributed to the emphasis on substantial molecular modifications in the prompt. In contrast, prompt E, lacking specificity on the extent of changes, exhibits a more balanced exploration with less directional bias.

Furthermore, an intriguing pattern emerges when examining the origin and trajectory of molecules in these prompts. It is observed that certain molecules start from the bottom region and gradually progress upwards, dispersing in various directions. This observation aligns with the expected distribution of molecular structures in the chemical space. The bottom region tends to be populated by chain-like molecules, characterized by their linear and elongated structures. As we move upwards in the chemical space, there is a notable shift towards a higher concentration of ring-like structures. This transition from chains to rings can be attributed to the language model's exploration or exploitation of the different regions of the chemical space without any defined targets, and just based on prompts alone.

The fine prompts (A, B, C, D) and some coarse prompts (like E) exhibit an intriguing observation: the arrows representing directional changes in parent molecules often negate each other. This occurs because individual vectors, originating from parent molecules and pointing towards generated molecules, frequently point in opposing directions. Consequently, these counteracting vectors effectively cancel out, resulting in no significant collective movement within the latent space.

These observations highlight the nuanced influence of prompt engineering in steering molecular evolution in the latent space and showcase the model's ability to adapt molecular structures based on the diverse requirements of each prompt. However, to fully understand the molecular modifications observed for certain parent molecules, a deeper exploration of the latent space and the underlying modifications is necessary.

The later sections of the manuscript throw more light on these molecular modifications, such as how simple modifications like incorporating specific types of functional groups can be accomplished by asking the model to incorporate electron-withdrawing or electron-donating groups and generate new candidates, which can be crucial for tuning the electronic properties, leading to applications in drug discovery or energy-storage devices among others.

The understanding of the structural changes made by functional group additions or other transformations can provide insights into how the model interprets and responds to the prompts, which could be beneficial for using LLMs in rational molecular design and molecular optimization. While some may argue that inverse molecular design is the ultimate goal, understanding how LLMs function, perform inverse design, and comprehend molecules and chemistry is essential. To fully capitalize on the potential of LLMs in molecular design, further research is needed to better understand the design space, role of prompt engineering, and unravel the underlying mechanisms by which these models navigate the chemical space and generate molecules with (conditional generation with targets) or without (generation with no targets) desired properties.

To summarize, the directional shifts quantitatively captured in the latent space provide critical insights into the model's strategic approach to various prompts. These quantitative movements within the latent space illustrate the extent of exploration or exploitation achieved, highlighting the model's ability to navigate the chemical space based on the prompts provided. By quantifying these directional shifts, we demonstrate the importance of prompt engineering in leveraging LLMs for molecular design and discovery.

Guided generation

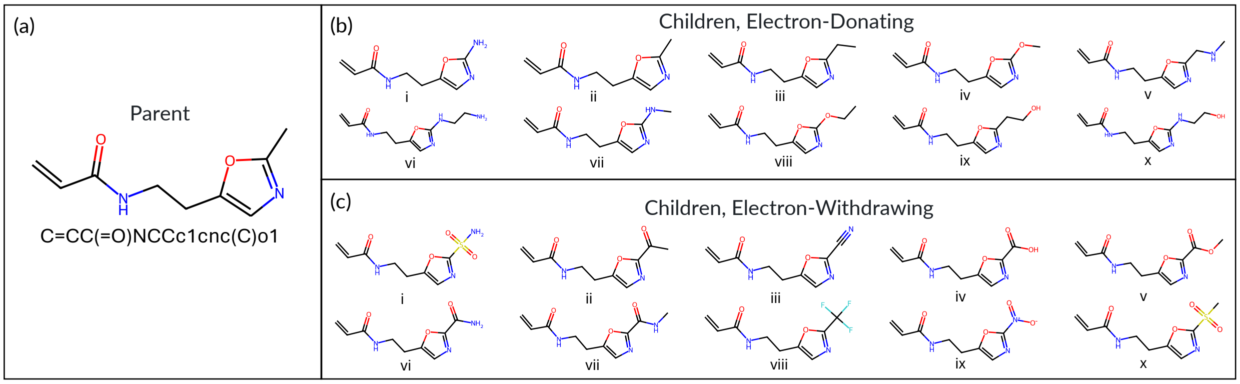

Figure 6: A representative selection of molecules generated by prompts L and P (analogous to prompt D from Figure 2).

Prompts I through L and M through P were more chemically specific than the previous prompts and targeted variations of the parent molecule with electron withdrawing groups or electron donating groups. These two sets of prompts might be useful for creating panels of molecules for evaluating inductive effects on reactivity, similar to many physical organic chemistry studies in testing the scope of a reaction. Both sets of prompts, I through L and M through P retained the parent molecular framework and installed the instructed electron withdrawing or electron donating groups in the children, without gross departures from the prompt instructions. Overall, these two sets of chemically-specific prompts behaved in a predictable manner across all 64 parent molecules.

Representative examples of molecules generated by sub-prompt D are shown in Figure 6. Similar to what was observed for Figure 6, the prompt interprets the center of the molecule as being the "backbone," and only makes changes to the methyl group on the right-hand side of the molecule. This group was exchanged for amine, methoxy, alcohol, and other similar electron-rich moieties in the case of the EDGs, as expected. For EWGs, a similar trend occurs, but the functional groups are more complex, including carbonyls, carboxylic acids, esters, nitriles, F-containing groups, and sulfonyl groups, all of which are reasonable EWG substitutions.

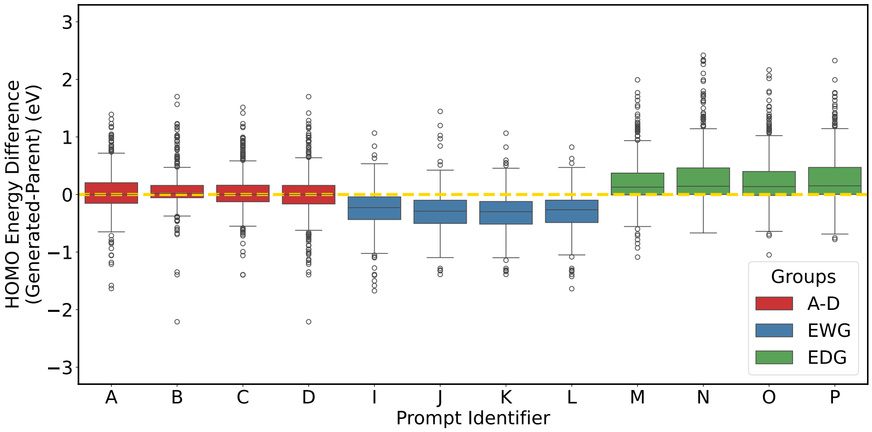

Figure 7: The figure represents the HOMO-HOMO energy differences between the parent and the generated molecules (child - parent) for the base fine prompts (A-D) as well as the EDG (I-L) and EWG (M-P) versions of the same base prompts.

We calculate the HOMO energies of the parent and child molecules to assess the degree to which the HOMO energy was modified in the child. The base prompts (A-D) have an overall median change of 0.0 eV and an interquartile range (IQR) of 0.29 eV. The HOMO energy increases 48.9% of the time and decreases 42.4% of the time, with the balance being no change. Furthermore, the different prompts all show similar behavior, with median changes all individually close to 0 eV and IQRs ranging from 0.21 eV to 0.36 eV.

The EWG and EDG prompts exhibit markedly different behaviors than the base prompts.

The EWG prompts (I-L) show a negative median change of -0.27 eV with an IQR of 0.40 eV; the HOMO energy decreases in 82.5% of cases for the EWG prompts. In contrast, the EDG prompts (M-P) have a positive median change of 0.14 eV with an IQR of 0.43 eV; the HOMO energy increases 69.7% of the time for the EDG prompts. These results are consistent with the expected impact of these functional groups on the HOMO energy of the molecules.

To summarize, the base prompts do not modify the HOMO energy level on average, while the EWG prompts tend to decrease it, and the EDG prompts increase it. This demonstrates the desired behavior in the guided prompts and shows that they are qualitatively different from the base prompts. In summary, the LLM successfully follows the natural-language instructions and generates valid molecules that achieve the desired result.

Conclusion

Our work explores the molecular design capability of large language models by making molecular modifications in the SMILES string representations of the parent molecules. We have shown that large language models like Claude 3 Opus can read, write, and make molecular modifications according to given instructions in the form of prompts, with 97% of outputs being valid molecules different from their parent molecule. By quantifying the modifications in a low-dimensional latent space, we have systematically evaluated the behavior of the large language model agent when using different prompts.

Moreover, the large language model successfully performs guided molecular generation, as shown by its ability to effectively manipulate the electronic structure of molecules using simple, natural-language prompts. This was demonstrated in the cases of electron-withdrawing group (EWG) and electron-donating group (EDG) prompts, where the model successfully lowered and raised the HOMO energy of the generated molecules relative to the parent molecules, compared to prompts that did not explicitly mention electronic structure changes.

These results showcase the model's capacity to understand and respond to specific electronic structure-related instructions, enabling targeted control over the properties of the generated molecules.

These findings open up exciting avenues for future research on molecular design. Future works should focus on developing "programming" based automatic prompt engineering methods. Such methods could help discover optimal prompts automatically instead of requiring extensive prompt engineering tailored to different applications such as drug discovery or 2D materials, enabling more efficient and targeted exploration of chemical design space. Molecular design using Large Language Models can prove to be significantly useful for accelerating the design of novel molecules with desired properties by the use of simple and natural language.

Acknowledgement

This work was supported as part of the Fast and Cooperative Ion Transport in Polymer-Based Materials (FaCT), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences at Oak Ridge National Laboratory and University of Illinois Urbana-Champaign under contract DE-AC05-00OR22725.

Supporting Information Available

The codes and raw data used in this work are available on Zenodo.[20] We also provide a web application to view the generated molecules from each prompt.

References

(1) Yang, Z.; Ye, W.; Lei, X.; Schweigert, D.; Kwon, H.-K.; Khajeh, A. De novo design of polymer electrolytes with high conductivity using gpt-based and diffusion-based gener-

ative models. *arXiv preprint arXiv:2312.06470* **2023**,

(2) Hu, M.; Tan, Q.; Knibbe, R.; Xu, M.; Jiang, B.; Wang, S.; Li, X.; Zhang, M.-X. Recent applications of machine learning in alloy design: A review. *Materials Science and Engineering: R: Reports* **2023**, *155*, 100746.

(3) Ryu, B.; Wang, L.; Pu, H.; Chan, M. K.; Chen, J. Understanding, discovery, and synthesis of 2D materials enabled by machine learning. *Chemical Society Reviews* **2022**, *51*, 1899–1925.

(4) Tong, X.; Liu, X.; Tan, X.; Li, X.; Jiang, J.; Xiong, Z.; Xu, T.; Jiang, H.; Qiao, N.; Zheng, M. Generative models for de novo drug design. *Journal of Medicinal Chemistry* **2021**, *64*, 14011–14027.

(5) Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discovery Today* **2021**, *26*, 2707–2715.

(6) Wang, M.; Wang, Z.; Sun, H.; Wang, J.; Shen, C.; Weng, G.; Chai, X.; Li, H.; Cao, D.; Hou, T. Deep learning approaches for de novo drug design: An overview. *Current Opinion in Structural Biology* **2022**, *72*, 135–144.

(7) Bhadwal, A. S.; Kumar, K.; Kumar, N. GenSMILES: An enhanced validity conscious representation for inverse design of molecules. *Knowledge-Based Systems* **2023**, *268*, 110429.

(8) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; others Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, *152*, 60–69.

(9) Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; Badhwar, S.; Bocarsly, J. D.; Bran, A. M.; Bringuier, S.; Brinson, L. C.; Choudhary, K.; Circi, D.; others 14 examples of how

LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* **2023**, *2*, 1233–1250.

(10) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1603.

(11) Tingle, B. I.; Tang, K. G.; Castanon, M.; Gutierrez, J. J.; Khurelbaatar, M.; Dandarchuluun, C.; Moroz, Y. S.; Irwin, J. J. ZINC-22– A free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of Chemical Information and Modeling* **2023**, *63*, 1166–1176.

(12) Zhong, S.; Guan, X. Count-based morgan fingerprint: A more efficient and interpretable molecular representation in developing machine learning-based predictive regression models for water contaminants' activities and properties. *Environmental Science & Technology* **2023**, *57*, 18193–18202.

(13) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(14) Anthropic Introducing the next generation of Claude. `https://www.anthropic.com/news/claude-3-family`, 2024; Accessed: 19 April 2024.

(15) Anthropic Anthropic Python API library. `https://github.com/anthropics/anthropic-sdk-python`, 2024; Accessed: 19 April 2024.

(16) RDKit Community RDKit: Open-source cheminformatics. 2023; `https://www.rdkit.org`, Version 2023.09.5.

(17) Stewart, J. J. P. AMS 2024.1 MOPAC: MOPAC Engine based on the MOPAC2016 source code. 2016; `http://OpenMOPAC.net`.

(18) Stewart, J. J. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of molecular modeling* **2013**, *19*, 1–32.

(19) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; others The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29*, 273002.

(20) Debjyoti Bhattacharya; Harrison Cassady; Michael Hickner; Wesley Reinhart Dataset for "Large Language Models as molecular design engines". 2024; `https://doi.org/10.5281/zenodo.11110873`.

(21) Macedo, B.; Ribeiro Vaz, I.; Taveira Gomes, T. MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design. *Scientific Reports* **2024**, *14*, 1212.

(22) Mercado, R.; Rastemo, T.; Lindelöf, E.; Klambauer, G.; Engkvist, O.; Chen, H.; Bjerrum, E. J. Practical notes on building molecular graph generative models. *Applied AI Letters* **2020**, *1*.

(23) Guo, T.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N.; Wiest, O.; Zhang, X.; others What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* **2023**, *36*, 59662–59688.

(24) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1608.