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 transformer-based models that have been successfully applied to various molecular design tasks. For example, C5T5-type models have been used in the design of antiviral drugs.[10] More recently, Matsukiyo et al. generated candidate molecules by exploring the latent space of a Transformer-based VAE, identifying inhibitors for the design of target proteins in therapeutic applications.[11] Additionally, MolGPT, a lightweight generative pre-trained transformer model based on a masked-attention transformer-decoder architecture, demonstrated molecular generation according to desired scaffolds while controlling multiple properties.[12] Furthermore, Tysinger et al. showed that transformers can be trained to make meaningful molecular modifications for hit expansion in bioactive molecular drug design.[13] These examples highlight the powerful utility of transformers across multiple molecular design problems.

Large Language Models (LLMs),[9] which are also based on transformer architectures initially trained on vast amounts of natural language data, have been recognized for their disruptive effect in nearly every field. Although they are not explicitly trained to be knowledgeable in chemistry, they have the advantage of being adaptable and generalizable.[14] Given their shared foundation with the aforementioned molecular design models, LLMs hold significant promise for advancing the field through their flexibility and broad applicability.

Recent studies have shown that LLMs can indeed capture and apply principles of chemistry (as represented in their training corpus) to solve complex problems, going beyond mere pattern recognition.[15,16] While we acknowledge that LLMs may not possess an understanding of chemistry similar to human experts, existing works suggest that they can potentially use chemical data to generate valid Simplified Molecular Input Line Entry System (SMILES)[17] strings (which encodes molecular structures in text form), propose meaningful molecular modifications, and guide the discovery of compounds with desired properties.

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

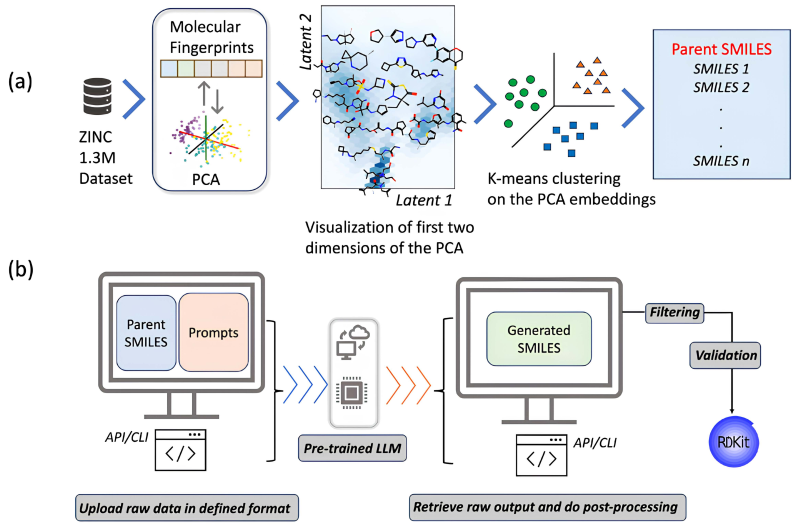


Figure 1: (a) represents the process of parent SMILES generation for the molecular modification process using Claude API. (b) represents the Claude API workflow and how unique validated (by RDKit) SMILES are obtained.

Dataset and representation learning

The dataset for this study includes approximately 1.3 million small molecules from the ZINC database.[18] 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 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[19] 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. A fingerprint size of 1024 vector and a radius of 2 atoms were used. A list of common keys was created for all the molecules, and counts-based fingerprints were generated based on the occurrence of a fragment in a molecule and its presence in the common keys. 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).

Note that the PCA embedding is used solely for selecting representative parent SMILES and for visualization purposes, not for downstream predictive tasks. Thus, a linear technique like PCA suffices. Although non-linear methods like t-distributed stochastic neighbor embedding (t-SNE)[20] and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)[21] could have been explored, PCA is straightforward, interpretable, and computationally efficient, making it well-suited for visualizing high-dimensional data in a lower-dimensional space.

We chose 64 parent molecules via K-means clustering (implemented in `scikit-learn`[22]) 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,[23] a state-of-the-art LLM. Interaction with the LLM is facilitated by the Anthropic Python SDK,[24] 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

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

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 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.

Prompts for controlled molecular generation

Apart from the base prompts and prompts for guided generation, we consider three additional prompts, described in Table 3, that allow us to control the extent of molecular modification and similarity with the given parent molecules. These prompts use verbal descriptors such as “barely similar (very low Tanimoto similarity)”, “marginally similar (low Tanimoto similarity)”, and “moderately similar (moderate Tanimoto similarity)” to describe the extent of similarity compared to the parent SMILES.

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

Validation and metrics

After receiving candidate SMILES strings from the LLM, they are validated by RDKit,[25] an open-source cheminformatics toolkit. We first ensure the strings represent valid molecules through RDKit's `Chem.MolFromSmiles` function followed by `Chem.SanitizeMol`. This method checks the molecular structure's validity and adherence to standard conventions, including incorrect valency. Thereafter, the valid molecules are converted to canonical form, as SMILES are not bijective mappings. The canonical SMILES undergo further filtering, elim-

Table 3: Detailed sub-prompts used to describe how the molecular modification task should be carried out, effectively controlling the similarity of generated compounds.

Identifier	Prompt detail text
Q	Barely similar (very low Tanimoto similarity) molecules compared to the given parent molecule by altering atoms, bonds, functional groups, or making other changes to find new candidates.
R	Marginally similar (low Tanimoto similarity) molecules compared to the given parent molecule by altering atoms, bonds, functional groups, or making other changes to find new candidates.
S	Moderately similar (moderate Tanimoto similarity) molecules compared to the given parent molecule by altering atoms, bonds, functional groups, or making other changes to find new candidates.

inating 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, similar to previous works:[26,27] Tanimoto similarity, validity ratio, and chemical diversity.

We calculated the Tanimoto similarity between parents and children for each prompt. This process involves converting each SMILES string to a molecule object, generating hashed Morgan fingerprints of a 1024 vector and a radius of 2 atoms, and then computing the Tanimoto similarity between the fingerprints. The Tanimoto similarity metric quantifies the molecular transformation induced by the prompt. Thus, the Tanimoto similarity $T(c_p, c_g)$ for a given prompt is calculated using the hashed Morgan fingerprints of the parent (p) and child (c) molecules. Specifically, we generate the fingerprints for both molecules and compute their Tanimoto similarity, which quantifies the similarity based on their structural features. This similarity measure quantifies the impact of the specific prompts on the magnitude of structural change between generated molecules and their parents. Low values of Tanimoto similarity $T(c_p, c_g)$ indicate small changes to the molecule, while high values indicate significant changes from the given parent molecule.

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 (in this case, hashed Morgan 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 $T(c_p, c_g)$. 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[28,29] program. The MOPAC calculations were performed through the Python interface provided by the Atomic

Simulation Environment (ASE) package.[30] 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 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

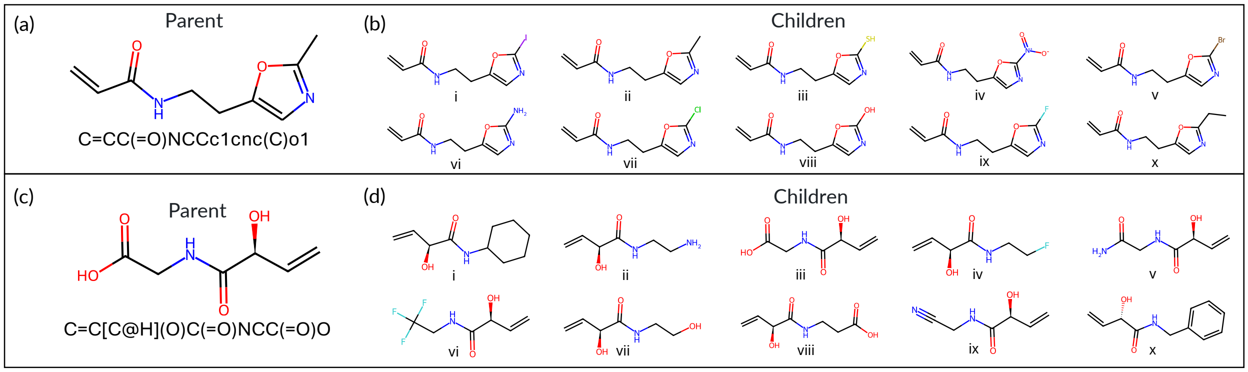


Figure 2: A representative selection of molecules generated by the prompt D ("similar molecules by changing one or two atoms or bonds to produce closely related structures") for two different parent molecules. The LLM accepts the parents (a, c) and returns 10 children (b, d) as SMILES; the molecules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

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."

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.[31]

Molecular fingerprint latent space

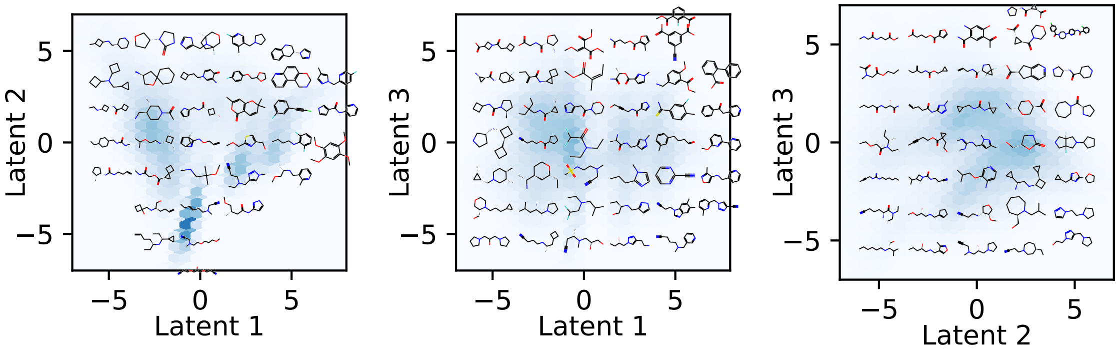


Figure 3: The latent space obtained from featurizing small molecules in the ZINC database with counts-based Morgan Fingerprints and embedding with PCA. Each panel shows a different 2D slice of the embedding, up to three components. Darker colors indicate a higher density of molecules occurring in that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

To systematically evaluate the impact of different prompts on molecule generation, we employed the following metrics (defined above): Tanimoto similarity $T(c_p, c_g)$, 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 $T(c_p, c_g)$ between the respective parent and generated molecules reveals that

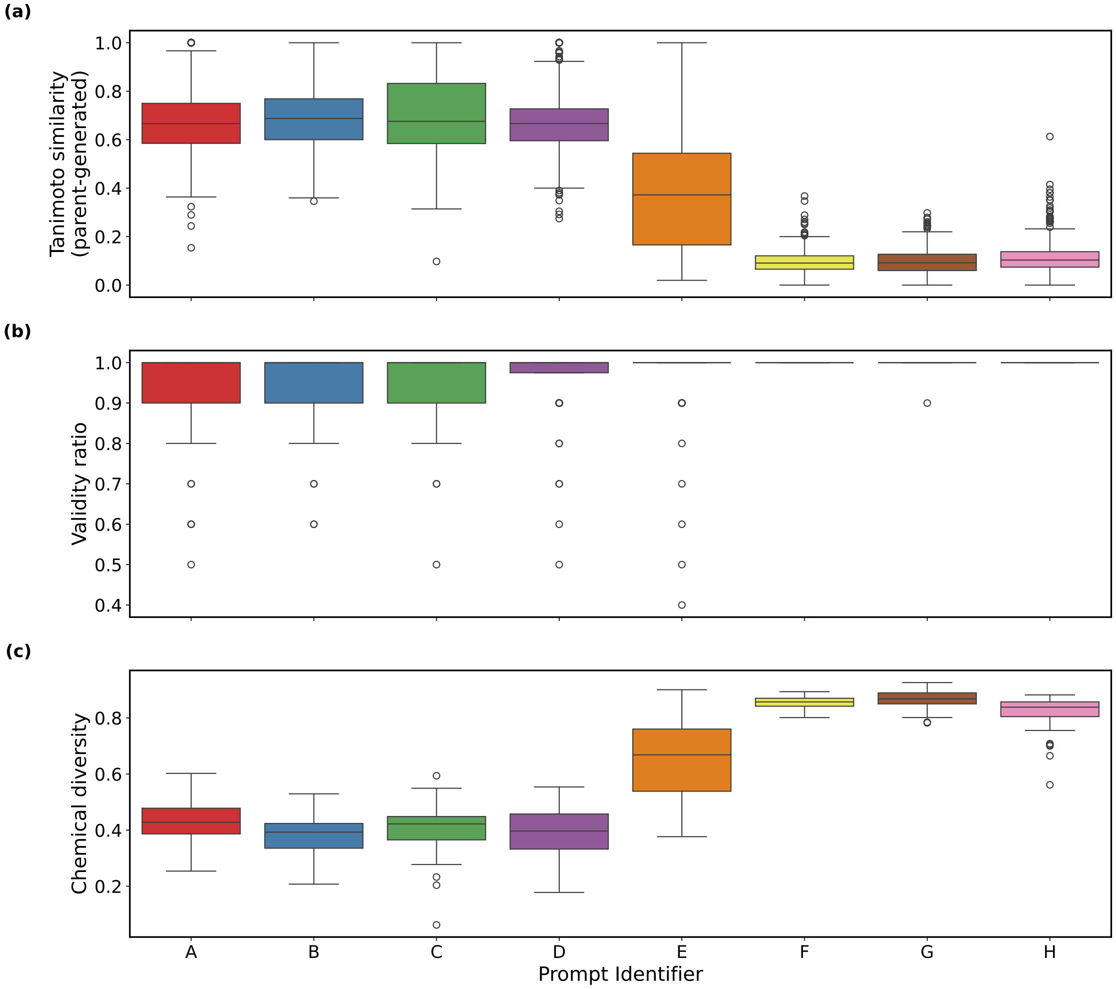


Figure 4: Metrics evaluated on each sub-prompt from Table 1 when the molecular modification task is performed on the same 64 parent molecules. (a) Tanimoto similarity $T(c_p, c_g)$, (b) validity ratio v, and (c) chemical diversity δ_c, as described in the text.

fine prompts generally result in higher $T(c_p, c_g)$ 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 $T(c_p, c_g)$ values, with prompt B showing a slightly higher median $T(c_p, c_g)$ at 0.69, compared to A, C, and D, which have median $T(c_p, c_g)$ values of 0.67, 0.68, and 0.67, respectively. Coarse prompts F-H exhibit comparable $T(c_p, c_g)$ distributions, with prompts F and G having the lowest median $T(c_p, c_g)$ at 0.09 across all prompts, and prompt H having a slightly higher median $T(c_p, c_g)$ at 0.10. However, prompt E is significantly higher at a median $T(c_p, c_g)$ of 0.37, 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 Tanimoto similarities $T(c_p, c_g)$ 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[32] for the best-performing models. Furthermore, generative models are susceptible to mode collapse, resulting in only $\tilde{2}\%$ uniquemolecules.[33] Thus, this LLM-based approach, which is itself a transformer-based architecture, consistently generated more valid and unique molecular structures even without fine-tuning. It can serve as a potential alternative to other existing transformer-based molecular generation frameworks, such as transformer architectures in conjunction with graph neural networks (GNNs),[34] conditional-based transformer architectures,[35] and others.[10,36,37]

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,[38] 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 zero-shot setting,[39] as done in our work, is inherently more challenging than in a few-shot approach[40,41] or an iterative improvement scheme.[42] In a zero-shot setting, the LLM is required to generate a response without any prior examples being given or feedback to the response generated, meaning it only gets one attempt to complete the specified task. This contrasts with a few-shot approach, where the model is provided with a few examples, or an iterative improvement scheme, where the model can refine its predictions over multiple iterations. The use of a zero-shot approach instead of a few-shot or iterative refinement significantly increases the difficulty of the molecular design problem. 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[43] that has employed similar measures to assess molecule generation tasks.

Evaluating bias

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 (≤ 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

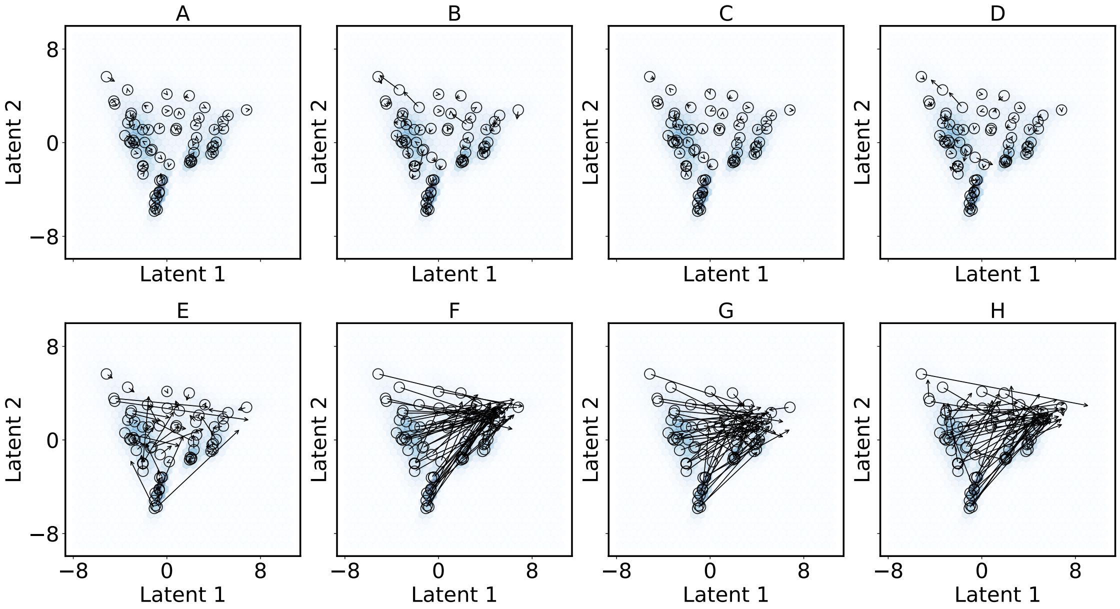


Figure 5: The average displacement in z from all 10 child molecules for each sub-prompt in Table 1. This illustrates prompt-specific bias in molecule generation tasks. Each circle indicates the parent coordinate, while the arrows indicate the average displacement between that parent and its children.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

to 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

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

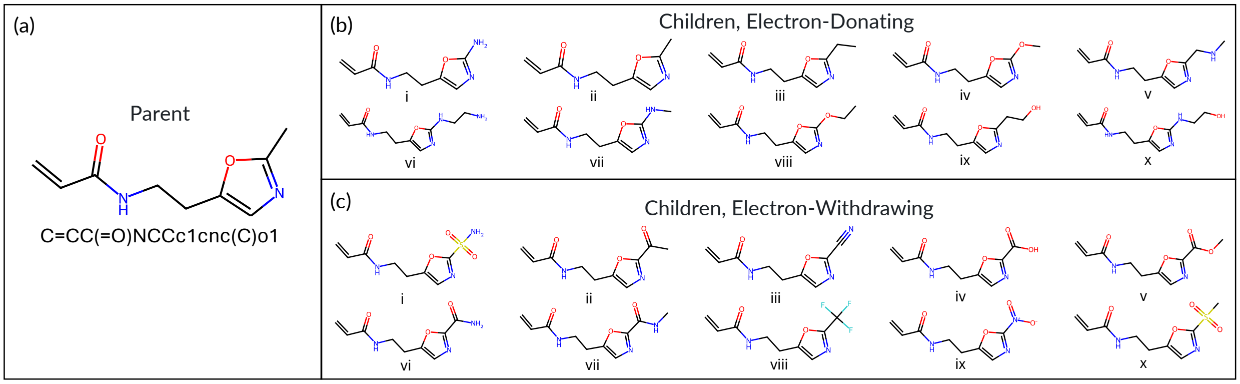


Figure 6: A representative selection of molecules generated by prompts L and P (analogous to prompt D from Figure 2).

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.

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, as shown in Figure 7. Herein, the base prompts (A-D) have an overall median change of 0.0 eV and an interquartile range (IQR) of 0.29 eV. Additionally, 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.

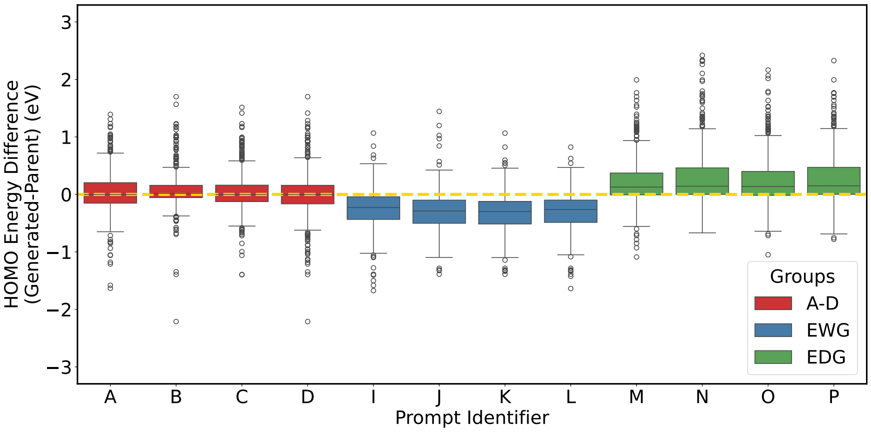


Figure 7: The figure represents the HOMO-HOMO energy differences between the parent and the generated molecules (child - parent) for the base fine prompts (A-D) as well as the EDG (I-L) and EWG (M-P) versions of the same base prompts.

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. In conclusion, these results demonstrate that the LLM is capable of generating molecules that modify electronic structures in the precise ways it was asked. This is further confirmed by the visualization provided through our molecular viewer, which shows the specific electronic variations achieved.[31] Additionally, recent works have demonstrated the potential of few-shot prompting,[40,41] iterative refinement schemes,[42] and chain-of-thought (CoT) prompting[44] in enhancing the performance of LLMs. Therefore, it may be worthwhile to explore these approaches further, combining them with advanced prompt engineering techniques, to understand how they can improve results beyond zero-shot prompting in the context of molecular design.

Overall, 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.

Controlled molecular generation performance

Controlling the extent of molecular similarity between generated molecules and given parent molecules may be necessary in some instances. Therefore, we explored prompts Q, R, and S(Figure 8) to achieve varying levels of Tanimoto similarity. The LLM is able to effectively distinguish these levels, from the descriptive prompt keywords “barely similar (very low Tanimoto similarity)” giving the lowest median Tanimoto similarity of 0.09, “moderately similar (moderate Tanimoto similarity)” yielding the highest median of 0.63, and “marginally similar (low Tanimoto similarity)” producing an intermediate median of 0.37. The median Tanimoto similarities for the base prompts $A - D$ ranged from $0.67 - 0.69$. Hence, with the controlling prompts Q, R, and S, the LLM is able to quantify different levels of similarity to the given parent molecules based on the keywords used in the prompt.

In future work, it may be useful to explore optimization-based approaches like TextGrad,[45] which is similar to a neural network in that it can help backpropagate textual feedback from LLMs to optimize molecular structures, potentially enhancing the LLM framework’s performance and hence aiding in guided generation.

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

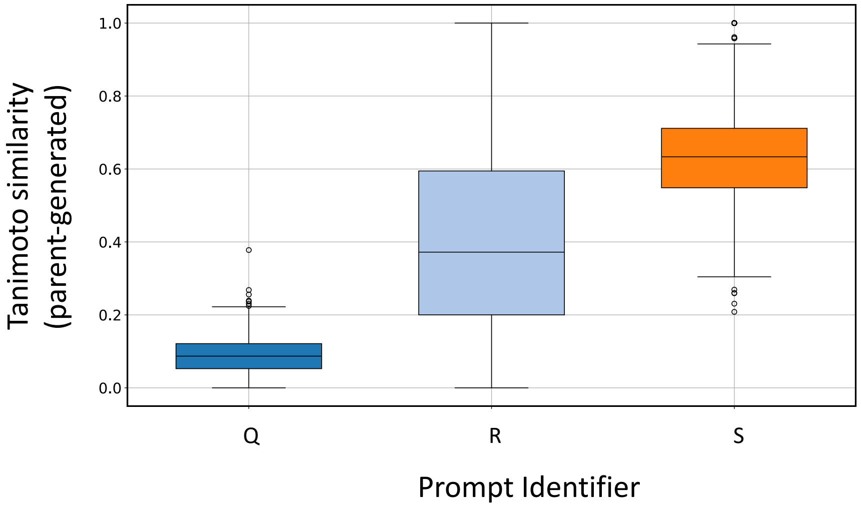


Figure 8: The figure represents the Tanimoto similarities $T(c_p, c_g)$ between the parent and the generated molecules for the prompts Q,R and S.

in a low-dimensional latent space, we have systematically evaluated the behavior of the large language model agent when using different prompts.

In addition, the large language model performs controlled molecular generation by controlling the levels of similarity to the given parent molecules, based on simple descriptive 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.

Supporting Information Available

The following contents are available in the Supporting Information: (1) Distribution of API call times for molecular generation with the Claude API across prompts A-H. (2) Median Tanimoto similarity between parent and generated molecules calculated across A-H prompts for the 64 parent SMILES structures. (3) Diversity of the generated molecules across prompts A-H, compared to the baseline diversity of the parent SMILES set. (4) Distribution of Synthetic Accessibility scores (SA Scores) for generated molecules across prompts A-H. (5) Tanimoto similarity (parent-generated) for all prompts (A-P).

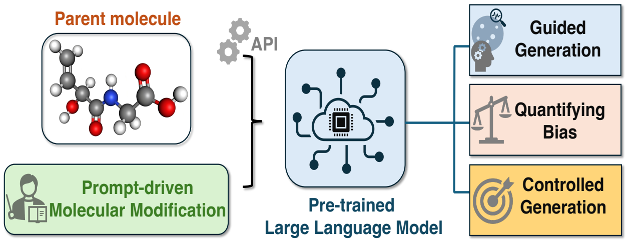
Acknowledgement

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.

Data and Software Availability Statement

The complete raw data and codes used in this work for analysis and visualization are publicly available on Zenodo.[31]

Table Of Contents Graphic



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 generative 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) Mao, J.; Wang, J.; Zeb, A.; Cho, K.-H.; Jin, H.; Kim, J.; Lee, O.; Wang, Y.; No, K. T. Transformer-based molecular generative model for antiviral drug design. *Journal of chemical information and modeling* **2023**, *64*, 2733–2745.

(11) Matsukiyo, Y.; Yamanaka, C.; Yamanishi, Y. De novo generation of chemical structures of inhibitor and activator candidates for therapeutic target proteins by a transformer-based variational autoencoder and bayesian optimization. *Journal of Chemical Information and Modeling* **2023**, *64*, 2345–2355.

(12) Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling* **2021**, *62*, 2064–2076.

(13) Tysinger, E. P.; Rai, B. K.; Sinitskiy, A. V. Can We Quickly Learn to “Translate” Bioactive Molecules with Transformer Models? *Journal of Chemical Information and Modeling* **2023**, *63*, 1734–1744.

(14) Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; others Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**,

(15) White, A. D.; Hocky, G. M.; Gandhi, H. A.; Ansari, M.; Cox, S.; Wellawatte, G. P.; Sasmal, S.; Yang, Z.; Liu, K.; Singh, Y.; others Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery* **2023**, *2*, 368–376.

(16) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, 1–11.

(17) 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.

(18) 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.

(19) 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.

(20) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*.

(21) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**,

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

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

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

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

(26) Brown, N.; Fiscato, M.; Segler, M. H.; Vaucher, A. C. GuacaMol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling* **2019**, *59*, 1096–1108.

(27) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; others Molecular sets (MOSES): a benchmarking platform for molecular generation models. *Frontiers in pharmacology* **2020**, *11*, 565644.

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

(29) 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.

(30) 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.

(31) 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`.

(32) 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.

(33) 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*.

(34) Buehler, M. J. Generative pretrained autoregressive transformer graph neural network applied to the analysis and discovery of novel proteins. *Journal of Applied Physics* **2023**, *134*.

(35) Dobberstein, N.; Maass, A.; Hamaekers, J. Llamol: a dynamic multi-conditional generative transformer for de novo molecular design. *Journal of Cheminformatics* **2024**, *16*.

(36) Wu, F.; Radev, D.; Li, S. Z. Molformer: Motif-based transformer on 3d heterogeneous molecular graphs. Proceedings of the AAAI Conference on Artificial Intelligence. 2023; pp 5312–5320.

(37) Bran, A. M.; Schwaller, P. Transformers and large language models for chemistry and drug discovery. *arXiv preprint arXiv:2310.06083* **2023**,

(38) 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.

(39) Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* **2022**, *35*, 22199–22213.

(40) Liu, X.; Guo, Y.; Li, H.; Liu, J.; Huang, S.; Ke, B.; Lv, J. DrugLLM: Open Large Language Model for Few-shot Molecule Generation. *arXiv preprint arXiv:2405.06690* **2024**,

(41) Liu, Y.; Ding, S.; Zhou, S.; Fan, W.; Tan, Q. MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction. *arXiv preprint arXiv:2406.12950* **2024**,

(42) Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; others Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* **2024**, *36*.

(43) 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.

(44) Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; Zhai, X. Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence* **2024**, *6*, 100213.

(45) Yuksekgonul, M.; Bianchi, F.; Boen, J.; Liu, S.; Huang, Z.; Guestrin, C.; Zou, J. TextGrad: Automatic" Differentiation" via Text. *arXiv preprint arXiv:2406.07496* **2024**,