NaviDiv: A Comprehensive Tool for Monitoring Chemical Diversity in Generative Molecular Design

Mohammed Azzouzi
LCMD, ISIC,
EPFL, Lausanne, Switzerland
Mohammed.azzouzi@epfl.ch

Thanapat Worakul
LCMD, ISIC,
EPFL, Lausanne, Switzerland
thanapat.worakul@epfl.ch

Clémence Corminboeuf
LCMD, ISIC,
EPFL, Lausanne, Switzerland
clemence.corminboeuf@epfl.ch

Abstract

The rapid progress in generative models for molecular design has led to extensive libraries of candidate molecules for biological and chemical targets. However, ensuring these molecules are diverse and representative of the broader chemical space remains challenging. Without proper tools, researchers may over-explore limited regions or miss promising candidates. This work presents NaviDiv, a comprehensive set of tools for analyzing and steering chemical diversity in generative molecular design, introducing multiple complementary metrics that capture different aspects of molecular diversity through representation distance-based, string-based, fragment-based, and scaffold-based approaches. Our package not only monitors diversity evolution but also provides adaptive diversity constraints that can be integrated into the optimization process to guide generative models toward maintaining desired levels of chemical space exploration. Through a case study on singlet fission material discovery using REINVENT4, we demonstrate how different diversity metrics evolve during reinforcement learning optimization and show that our diversity constraints can prevent model collapse while preserving property optimization performance. The package is freely available in NaviDiv GitHub repository. This initial implementation serves as a foundation for future extensions to additional molecular representations and generative architectures, addressing a critical bottleneck in automated molecular discovery.

1 Introduction

Generative molecular models represent a paradigm shift in molecular design, moving beyond fixed databases and manual construction rules to learn statistical distributions of chemical structures in high-dimensional latent spaces.[1, 2, 3, 4] These models generate realistic, chemically plausible molecules that leverage expanding chemical databases to capture broad diversity and enable the design of novel compounds with desired properties.

Multiple architectures exist for generative molecular design, differing in molecular representations and neural network structures.[5] String-based approaches treat molecules as SMILES or SELFIES sequences, using RNNs to autoregressively predict tokens and achieving over 90% validity rates for drug-like molecules.[6, 7, 8, 9] Alternative architectures include VAEs, GANs, transformers, flow-based models, and diffusion models, operating on 1D (strings), 2D (graphs), or 3D (coordinate) representations.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

For property-targeted generation, guidance strategies must steer models toward higher-performing candidates. VAEs employ gradient-based latent space optimization, while RNN models like REINVENT use reinforcement learning to update parameters via policy-driven approaches.[6, 10] However, optimization inevitably reduces molecular diversity as models drift from their training distributions. While regularization terms can maintain alignment with the original model,[11] this is often insufficient to prevent model collapse. Additional diversity-promoting scoring functions targeting overrepresented fragments or similar compounds are typically required, but their selection demands deep understanding of model behavior and careful consideration of the trade-offs between diversity maintenance and property optimization performance.

This work presents NaviDiv, a comprehensive set of tools for analyzing and steering chemical diversity in generative molecular design. We introduce multiple complementary diversity metrics spanning structural fingerprints, string-based analysis, and fragment decomposition, coupled with adaptive diversity constraints that can be integrated into the optimization process to guide molecular generation. Through a singlet fission material discovery case study, we demonstrate how different diversity metrics evolve during reinforcement learning optimization, show how our diversity constraints can prevent model collapse while maintaining property optimization performance, and provide insights into property-diversity trade-offs in molecular discovery.

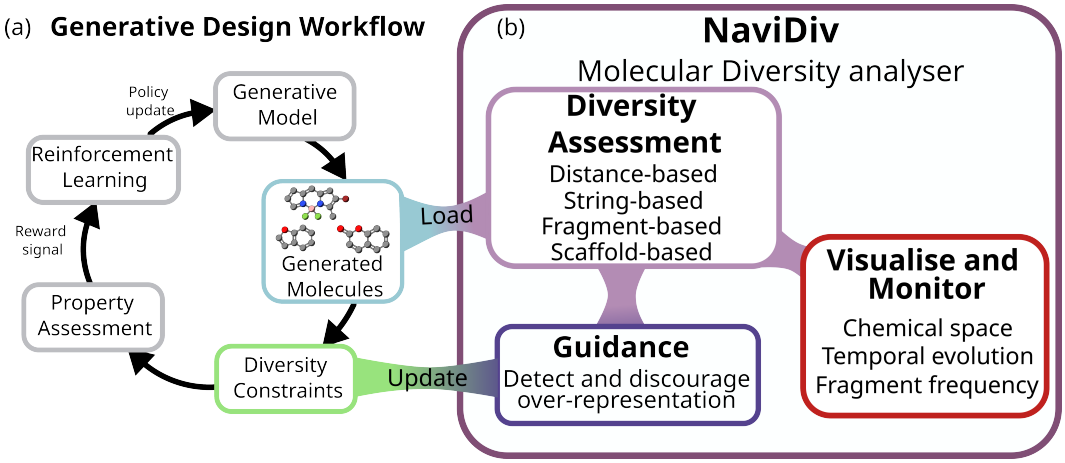

Figure 1: **NaviDiv molecular diversity tool.** (a) Representation of a generation step with reinforcement learning. (b) Overview of the molecular diversity analyser, showcasing the various metrics and visualizations provided to assess chemical diversity during the generative process. The generative model produces molecules that are analyzed using multiple diversity metrics, including representation distance-based, string-based, fragment-based, and scaffold-based approaches. The tool offers visualizations such as 2D chemical space projections and temporal evolution plots to monitor diversity changes over time. The results of the analysis can be used to update a diversity constraints function that can be integrated into the generative model's optimization process.

2 NaviDiv Framework

NaviDiv is a comprehensive tool designed to monitor, analyze, and actively steer chemical diversity in generative molecular design workflows. The tool integrates seamlessly into existing reinforcement learning-based generation pipelines, providing real-time diversity assessment and adaptive constraint mechanisms. The use of other generative architectures (e.g., VAEs, GANs, diffusion models) or molecular representations (e.g., graphs, 3D coordinates) is not yet supported but can be implemented in future versions.

Figure 1 illustrates the core architecture of NaviDiv. In a typical generative molecular design workflow (Figure 1a), a generative model produces candidate molecules at each iteration, which are evaluated using a predefined scoring function that assesses property criteria. These scores update the model through reinforcement learning, iteratively improving molecular quality over successive generations.

NaviDiv extends this workflow by introducing a comprehensive diversity analysis layer (Figure 1b) that operates in parallel with property evaluation. The framework receives generated molecules and performs diversity assessment using multiple complementary metrics, capturing different aspects of chemical diversity. The analysis results are visualized through intuitive plots and dashboards, enabling researchers to monitor diversity trends in real-time. The framework also includes an adaptive diversity constraint module that can generate penalty functions based on the analysis results. These penalties can be integrated into the scoring function to actively guide the generative model toward maintaining desired diversity levels while optimizing for target properties. Below we present the key components of NaviDiv.

i) Assessing chemical diversity Chemical diversity is context-dependent and varies across fields.[12, 13] In Organic electronics, the focus is on π-conjugated building blocks and molecular symmetry,[14] while catalysis emphasizes ligand modifications around catalytic cores,[15, 16] and drug discovery targets scaffold and stereochemical variation.[17, 18]. Our framework implements four complementary approaches for comprehensive diversity assessment (Implementation details are provided in the appendix D.):[19, 20]

- **Representation distance-based analysis** employs molecular fingerprints (e.g., Morgan, RDKit) and similarity metrics (e.g., Tanimoto coefficient) to quantify structural diversity.
- **String-based analysis** examines SMILES or SELFIES representations using n-gram frequency distributions to capture syntactic and semantic diversity in molecular encoding.
- **Fragment-based analysis** decomposes molecules into chemically relevant substructures, tracking fragment frequency and distribution to identify overrepresented motifs and assess substructural diversity.[21]
- **Scaffold-based analysis** identifies core molecular frameworks after systematic side-chain removal, monitoring scaffold diversity and evolution to understand how generative models explore fundamental chemical architectures.[22]

ii) Visualization and Monitoring Capabilities NaviDiv provides rich visualization tools through an interactive web application built with Streamlit and a Python backend, enabling researchers to understand and interpret diversity evolution in real-time (see appendix F for screenshots). Key features include:

- **Chemical space projections**: Interactive 2D visualizations of molecular distributions using dimensionality reduction techniques (t-SNE) that reveal clustering patterns and space exploration trajectories.[23] Users can hover over or click individual points to visualize the corresponding chemical structures directly within the interface.
- **Temporal evolution plots**: Dynamic time-series analysis showing how different diversity metrics evolve throughout the optimization process, with real-time updates and interactive controls for identifying potential model collapse.
- **Fragment frequency analysis**: Interactive visualization of molecular substructure distributions enabling identification of overrepresented fragments and emerging chemical motifs. The tool displays fragment occurrence frequencies, ranks substructures by prevalence, and provides molecular structure viewing for fragments exceeding user-defined thresholds, facilitating targeted diversity constraint design.

iii) Adaptive Diversity Constraints A key innovation of NaviDiv is its ability to actively guide the generative process through adaptive diversity constraints. Based on real-time diversity analysis, the framework dynamically generates penalty functions that can be integrated into the optimization objective. These constraints operate by (Implementation details are provided in the appendix E.):

1. **Monitoring over-representation**: Identifying molecular clusters, fragments, or sequence patterns that exceed predefined frequency thresholds.
2. **Dynamic penalty generation**: Computing penalty scores for molecules that contribute to overrepresented regions of chemical space.

3 Use Case: Singlet Fission Material Discovery

To demonstrate the capabilities of our chemical diversity analysis framework, we consider the case of discovering molecules for singlet fission applications in solar cells. We employ the evaluation function established in previous work to explore the chemical space of molecules with singlet fission character, specifically assessing the difference in energy between the lowest first singly excited state and the energy of the triplet excited state.[24] (details about the scoring function can be found in the appendix F.)

We use REINVENT4 with a prior trained on an extended dataset adapted for organic electronic molecules (FORMED [25], GEOM3D [26]) and conduct reinforcement learning for 1000 iterations with 100 molecules generated per step. Note that NaviDiv's current implementation is optimized for string-based generative models using reinforcement learning, making REINVENT4 an ideal demonstration platform.

Figure 2 shows the evolution of the average molecular score alongside key diversity metrics throughout the reinforcement learning process. Details about the different diversity metrics considered in this case can be found in the appendix. As training progresses, the average score increases steadily, indicating successful optimization toward the design objective. However, this improvement comes at the cost of reduced diversity among the generated molecules.

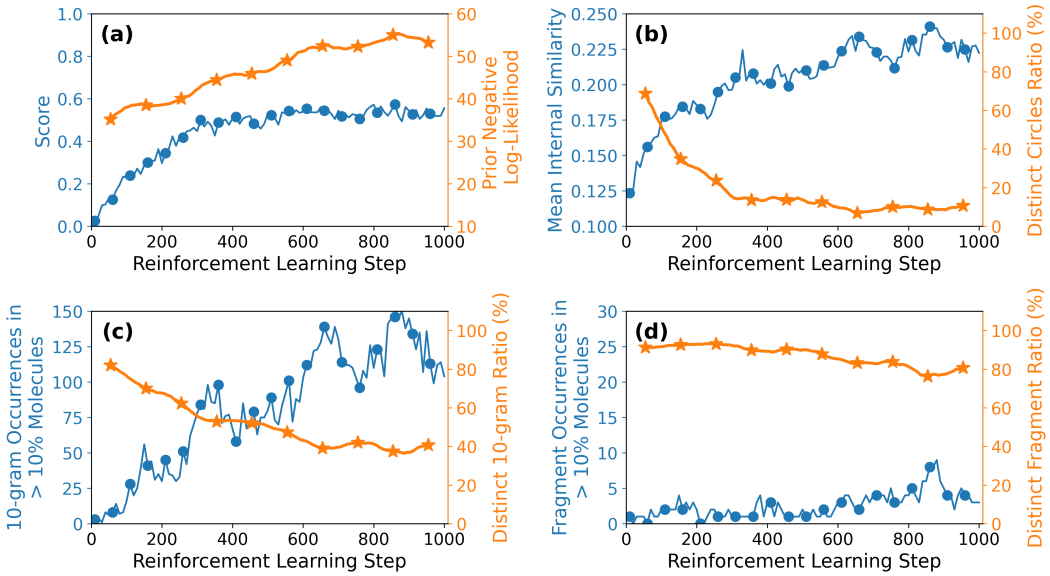

Figure 2: **Evolution of molecular optimization and diversity during reinforcement learning.** (a) Average molecular score and negative log-likelihood vs. RL steps, showing successful optimization and chemical space exploration. (b) Structural diversity via Morgan fingerprints: mean pairwise similarity (left) and number of clusters with similarity < 0.3 (right). (c) SMILES diversity via 10-grams: fraction of unique sequences (left) and overrepresented patterns $> 10\%$ (right). (d) Fragment diversity: percentage of unique fragments (left) and overrepresented fragments $> 10\%$ (right).

The analysis reveals that the extent of diversity loss varies significantly depending on the representation used:

- **Structural diversity** (evaluated using Morgan fingerprints with Tanimoto similarity) shows a strong increase in the mean internal similarity as the model focuses on structurally similar high-performing molecules. The number of distinct clusters decreases from 100 % to 10 % at the end of the 1000 iterations. This indicates a convergence toward a narrower set of molecular structures.
- **Sequence-level diversity** (measured via 10-gram analysis) exhibits more substantial reduction in the number of distinct sequences, as well as a strong increase in the number of

sequences present in more than 10 % of the molecules reaching almost 100 sequences at the end of the reinforcement learning process. This indicates convergence toward similar SMILES patterns.

- **Fragment-based diversity** shows gradual decline but maintains better diversity compared to other metrics, suggesting that while specific fragments become overrepresented, the overall fragment space remains relatively diverse.

These findings highlight the necessity of employing multiple complementary diversity metrics to fully capture the evolution of molecules generated by reinforcement learning-induced model changes. The different behavior across metrics provides insights into how the optimization process affects different aspects of chemical space exploration, enabling researchers to make informed decisions about when and how to intervene to maintain desired levels of diversity.

Next, we showcase the impact of introducing diversity constraints on the evolution of chemical diversity among molecules generated during reinforcement learning (Figure 3). We consider three different constraint regimes, each defined by specific diversity constraints applied during the reinforcement learning process. Thresholds are established based on both percentage of molecules generated in previous steps and absolute numbers, chosen to ensure effectiveness without being overly restrictive. The three constraint regimes are:

1. **Baseline (No Constraints):** Standard reinforcement learning without diversity constraints.
2. **Fragment-Based Constraints:** Avoiding overrepresented molecular fragments. Here the threshold for adding a fragment to the list of fragments to avoid is set to 5% of the molecules generated in the previous steps, or a total of 50 molecules generated that contain that fragment, and we only consider fragments that are larger than 8 non-hydrogen atoms.
3. **Combined Constraints:** Integration of similarity-based, fragment-based, and n-gram-based constraint types. Thresholds are set to 10% for similarity-based constraints, 5% for fragments (considering only fragments larger than 8 non-hydrogen atoms), and 3% for 10-grams in SMILES sequences.

Figure 3 shows results for the three constraint regimes. All regimes achieve successful optimization with steadily increasing molecular scores (Figure 3a), but at different rates: baseline (fastest), Fragment-Based (comparable), and combined constraints (slowest).

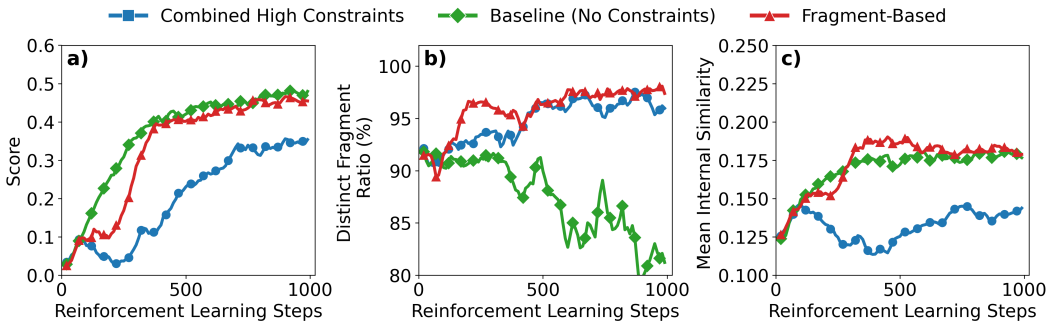

Figure 3: **Impact of diversity constraints on molecular optimization and diversity during reinforcement learning.** (a) Average molecular score vs. RL steps for three constraint regimes: Baseline (no constraints), Fragment-Based constraints, and Combined constraints. (b) Fragment diversity via percentage of unique fragments over RL steps. (c) Structural diversity via mean pairwise similarity over RL steps.

Figure 3b shows the evolution of the diversity according to the ratio of unique fragments in the generated molecules. The baseline regime shows a rapid decline in fragment diversity after 200 iterations (coinciding with the optimization plateau), dropping to around 80% by the end of the optimization. Fragment-Based and Combined constraints maintain higher fragment diversity, with Combined constraints preserving nearly 100% of unique fragments throughout the process.

Figure 3c shows the evolution of the diversity according to the mean pairwise similarity. In this case, both the baseline and Fragment-Based constraints show a rapid increase in similarity, reaching around 0.2 by the end of the optimization. Combined constraints maintain lower similarity levels, around 0.12 like initial levels, indicating better preservation of structural diversity.

The three regimes show distinct trade-offs: 1) baseline achieves rapid optimization but significant diversity loss, 2) fragment-based constraints maintain optimization speed, reduce the loss of fragment diversity, but fail to maintain high pairwise diversity, and 3) combined constraints preserve diversity at the expense of slower optimization. These results highlight the importance of diversity monitoring tools for understanding constraint impacts and adapting optimization strategies.

Combined constraints provide the best diversity preservation across all metrics, demonstrating that threshold selection is crucial for balancing optimization performance and diversity preservation. Monitoring tools are essential for providing feedback to adjust thresholds based on specific applications and desired diversity levels.

4 Limitations

While NaviDiv advances chemical diversity analysis for generative molecular design, several limitations should be acknowledged. Our evaluation focuses on a single case study (singlet fission materials with REINVENT4), and broader validation across different molecular design tasks and generative architectures would strengthen evidence for general applicability. Optimal diversity metric selection and threshold determination are context-dependent and require manual tuning, with automated parameter optimization methods remaining an open challenge. Despite being designed as model-agnostic, our implementation focuses on REINVENT4 and has limited validation with other generative architectures or molecular representations. Finally, current metrics may not capture specialized aspects of chemical diversity such as stereochemistry, conformational flexibility, or domain-specific structural motifs. These limitations suggest important directions for future development and indicate areas requiring caution when applying NaviDiv to novel domains or unprecedented scales.

5 Conclusion

We present NaviDiv, a comprehensive tool for analyzing and steering chemical diversity in generative molecular design. Our approach introduces four complementary diversity metrics—representation distance-based, string-based, fragment-based, and scaffold-based analyses—each capturing different aspects of molecular variation. The framework provides both passive monitoring capabilities through an interactive Streamlit web application and active steering through adaptive diversity constraints that can be integrated into optimization processes.

Through a singlet fission material discovery case study using REINVENT4, we demonstrate that different diversity metrics exhibit varying sensitivities to reinforcement learning optimization. While structural fingerprint-based diversity shows considerable decline, fragment-based metrics maintain relative robustness. Our constraint comparison reveals distinct trade-offs: baseline optimization achieves rapid property improvement but significant diversity loss, while combined constraints preserve diversity at the cost of slower optimization speed.

The framework addresses a critical bottleneck in automated molecular discovery by providing standardized tools for monitoring and optimizing chemical space exploration. By enabling real-time diversity assessment and adaptive constraint generation, NaviDiv empowers researchers to make informed decisions about when and how to intervene to maintain desired levels of chemical diversity during optimization campaigns.

To facilitate adoption, we make the framework freely available with stable versions archived on Zenodo (DOI: `https://zenodo.org/records/16901533`) and active development on GitHub (`https://github.com/mohammedazzouzi15/NaviDiv`). Future work will focus on three main directions: (1) extending the framework to additional molecular representations (graph-based, 3D coordinates) and generative architectures beyond reinforcement learning approaches, including VAEs, GANs, diffusion models, and transformer architectures; (2) implementing more diverse diversity

scores that are adapted to different applications; and (3) establishing sophisticated steering methods for diversity preservation.

Acknowledgments

M.A. acknowledges the Swiss National Science Foundation (SNSF) for funding through an SNSF Swiss Postdoctoral Fellowship (TMPFP2_217256).

References

[1] Edward O. Pyzer-Knapp, Jed W. Pitera, Peter W. J. Staar, Seiji Takeda, Teodoro Laino, Daniel P. Sanders, James Sexton, John R. Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, April 2022. Publisher: Nature Publishing Group.

[2] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5):e1608, 2022. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1608.

[3] Dylan M. Anstine and Olexandr Isayev. Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, April 2023. Publisher: American Chemical Society.

[4] Martin Vogt. Exploring chemical space — Generative models and their evaluation. *Artificial Intelligence in the Life Sciences*, 3:100064, December 2023.

[5] Chao Pang, Jianbo Qiao, Xiangxiang Zeng, Quan Zou, and Leyi Wei. Deep Generative Models in De Novo Drug Molecule Generation. *Journal of Chemical Information and Modeling*, 64(7):2174–2194, April 2024. Publisher: American Chemical Society.

[6] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. REINVENT 2.0: An AI Tool for De Novo Drug Design. *Journal of Chemical Information and Modeling*, 60(12):5918–5922, December 2020. Publisher: American Chemical Society.

[7] Hannes H. Loeffler, Shunzhou Wan, Marco Klähn, Agastya P. Bhati, and Peter V. Coveney. Optimal Molecular Design: Generative Active Learning Combining REINVENT with Precise Binding Free Energy Ranking Simulations. *Journal of Chemical Theory and Computation*, 20(18):8308–8328, September 2024. Publisher: American Chemical Society.

[8] Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Central Science*, 4(1):120–131, January 2018. Publisher: American Chemical Society.

[9] Francesca Grisoni, Michael Moret, Robin Lingwood, and Gisbert Schneider. Bidirectional Molecule Generation with Recurrent Neural Networks. *Journal of Chemical Information and Modeling*, 60(3):1175–1183, March 2020. Publisher: American Chemical Society.

[10] Hannes H. Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H. Mervin, and Ola Engkvist. Reinvent 4: Modern AI–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, February 2024.

[11] Vendy Fialková, Jiaxi Zhao, Kostas Papadopoulos, Ola Engkvist, Esben Jannik Bjerrum, Thierry Kogej, and Atanas Patronov. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. *Journal of Chemical Information and Modeling*, 62(9):2046–2063, May 2022. Publisher: American Chemical Society.

[12] José L Medina-Franco and Gerald M Maggiora. Molecular similarity analysis. *Chemoinformatics for drug discovery*, pages 343–399, 2013.

[13] Alexios Koutsoukas, Shardul Paricharak, Warren R. J. D. Galloway, David R. Spring, Adriaan P. IJzerman, Robert C. Glen, David Marcus, and Andreas Bender. How diverse are diversity assessment methods? a comparative analysis and benchmarking of molecular descriptor space. *Journal of Chemical Information and Modeling*, 54(1):230–242, January 2014.

[14] Antonio Facchetti. π-conjugated polymers for organic electronics and photovoltaic cell applications. *Chemistry of Materials*, 23(3):733–758, 2011.

[15] John F. Hartwig. *Organotransition Metal Chemistry: From Bonding to Catalysis*. University Science Books, Sausalito, CA, 2008.

[16] Piet W. N. M. van Leeuwen. *Homogeneous Catalysis: Understanding the Art*. Springer, 2004.

[17] Christopher A. Lipinski and Andrew L. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

[18] Michael M. Hann and Tudor I. Oprea. Pursuing the leadlikeness concept in pharmaceutical research. *Current Opinion in Chemical Biology*, 8(3):255–263, 2004.

[19] Elena Lenci and Andrea Trabocchi. Diversity-oriented synthesis and chemoinformatics: A fruitful synergy towards better chemical libraries. *European Journal of Organic Chemistry*, 2022(29):e202200575, August 2022.

[20] Nina Nikolova and Joanna Jaworska. Approaches to measure chemical similarity – a review. *QSAR & Combinatorial Science*, 22(9–10):1006–1026, December 2003.

[21] Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. Identification of diverse database subsets using property-based and fragment-based molecular descriptions. *Quantitative Structure-Activity Relationships*, 21(6):598–604, December 2002.

[22] Sarah R. Langdon, Nathan Brown, and Julian Blagg. Scaffold diversity of exemplified medicinal chemistry space. *Journal of Chemical Information and Modeling*, 51(9):2174–2185, September 2011.

[23] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[24] Thanapat Worakul, Rubén Laplaza, J. Terence Blaskovits, and Clémence Corminboeuf. Generative design of singlet fission materials by revisiting the use of a fragment-oriented database. *ChemRxiv*, March 2025.

[25] J. Terence Blaskovits, Ruben Laplaza, Sergi Vela, and Clémence Corminboeuf. Data-driven discovery of organic electronic materials enabled by hybrid top-down/bottom-up design. *Advanced Materials*, 36(2):2305602, January 2024.

[26] Simon Axelrod and Rafael Gómez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, April 2022.

[27] Franziska Kruger, Nikolaus Stiefl, and Gregory A. Landrum. rdscaffoldnetwork: The scaffold network implementation in rdkit. *Journal of Chemical Information and Modeling*, 60(7):3331–3335, July 2020.

[28] Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2024. PMID: 38147829.

A Additional Package Capabilities

Beyond tracking specific diversity metrics evolution during reinforcement learning, our framework provides comprehensive analysis and visualization tools for understanding the generative process. The tool enables:

- **interactive molecular visualization** with 2D structural representations, sorting and filtering options based on molecular properties, and side-by-side comparison of molecules from different generation steps.
- **temporal analysis**, users can monitor the evolution of specific molecular fragments, track cluster formation and dissolution patterns, observe scaffold progression, and analyze SMILES sequence pattern changes throughout optimization.

- **global dataset analysis**, generating statistical summaries of fragment distributions, detailed clustering statistics with representative structures, property-structure correlations, and comprehensive diversity trend reports with statistical significance testing.
- **comparative analysis capabilities**, include comparison against user-defined reference sets, similarity assessment with configurable thresholds, identification of novel structures dissimilar from reference compounds.

These integrated capabilities provide researchers with a comprehensive toolkit for understanding, monitoring, and optimizing chemical diversity in generative molecular design, enabling informed decision-making throughout the discovery process.

B Implementation Details

NaviDiv is implemented as a modular Python framework built on standard scientific computing libraries including RDKit for molecular manipulation, NumPy/SciPy for numerical computations, and scikit-learn for machine learning algorithms. The architecture follows object-oriented design principles with clear separation of concerns across four main modules.

B.1 Performance Optimizations

The diversity analysis framework is designed for real-time monitoring during generative model optimization, with minimal computational overhead. Performance benchmarks on a standard CPU machine (Intel Core i7, 16GB RAM) demonstrate that analyzing 100 molecules per generation step requires less than 3 seconds of computation time. For a complete optimization run of 100 steps with 100 molecules per step (10,000 molecules total), the entire diversity analysis completes in approximately 5 minutes.

This computational efficiency enables seamless integration into existing generative workflows without significantly impacting overall runtime. The lightweight nature of the analysis allows for real-time diversity monitoring, where penalty functions can be computed and applied during the generation process itself. This capability is particularly valuable for implementing adaptive diversity constraints that respond to the evolving chemical space exploration patterns, enabling researchers to guide the generation toward maintaining desired diversity levels throughout the optimization process.

B.2 Extensibility and Modularity

The modular design enables easy extension through abstract base classes and plugin architecture. New diversity metrics can be integrated by implementing the `DiversityMetric` interface, while custom visualization components inherit from `BaseVisualizer`. Configuration management uses Hydra framework for hierarchical configuration files, enabling reproducible experimental setups.

C Diversity Metrics

C.0.1 Representation Distance

A variety of molecular representations can be used to assess and visualize chemical diversity, particularly those based on molecular descriptors or pre-established fingerprints such as Morgan fingerprints and RDKit fingerprints. These representations encode molecular structures into numerical vectors, which can then be projected into a lower-dimensional space (e.g., 2D) using dimensionality reduction techniques such as t-SNE or PCA. This allows for intuitive visualization of the chemical space and the evolution of generated molecules over time.

To quantify diversity, similarity metrics such as the **Tanimoto coefficient** or **Euclidean distance** can be applied to these fingerprint vectors. These metrics enable the monitoring of diversity trends, for example, by tracking the mean pairwise distance between molecules or by applying clustering algorithms to observe changes in the number and distribution of molecular clusters. This approach is highly versatile and can be tailored to specific applications by selecting or designing molecular representations that emphasize relevant chemical features.

C.0.2 String-Based Metrics

Another widely used approach for representing molecules is through string-based formats, such as SMILES (Simplified Molecular Input Line Entry System). These representations are particularly common in deep generative models, where molecules are treated as sequences of characters. In this context, chemical diversity can be assessed through semantic or syntactic analysis of the strings. One common method involves analyzing the frequency and distribution of n-grams—subsequences of characters within the SMILES strings. Unlike structural fragments, n-grams do not necessarily correspond to chemically meaningful substructures but can still capture patterns in how molecules are encoded. This type of analysis provides a lightweight and flexible way to monitor diversity in generative models, especially when structural decoding is computationally expensive or unavailable.

C.0.3 Fragment-Based Metrics

Chemical diversity can also be assessed through the analysis of molecular fragments (substructures obtained by systematically decomposing molecules). This approach becomes particularly relevant for larger molecules, where recurring subunits may dominate the chemical space. By collecting and cataloguing the fragments present across a dataset, one can evaluate their frequency of occurrence and identify overrepresented motifs. Furthermore, fragment-level analysis can be extended to correlate the presence of specific fragments with molecular properties or performance metrics, offering insights into how certain substructures influence the overall behaviour or score of a molecule. This method provides a granular view of diversity and is especially useful for guiding fragment-based design strategies.

Molecular fragments can also be compared through simplified representations that abstract away detailed chemical information. For example, fragments can be converted into wireframe models that omit bond order, or atom types can be replaced with generic placeholders. In some cases, both simplifications are applied simultaneously. These abstraction techniques allow for a more generalized comparison of molecular patterns, which can be particularly useful when focusing on topological or connectivity-based features rather than specific chemical identities.

D Implementation Details of Diversity Metrics

D.1 Representation Distance-Based Metrics

This approach uses molecular representations such as structural fingerprints and distance metrics to quantify similarity or dissimilarity between compounds based on their overall structure.

D.1.1 Morgan Fingerprints and Tanimoto Similarity

Morgan fingerprints are computed using RDKit with the following parameters:

- Radius: 3 (equivalent to ECFP6)
- Number of bits: 2048
- Use features: False
- Use chirality: True

The Tanimoto coefficient between two fingerprints A and B is calculated as:

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

To evaluate the diversity of a set of generated molecules, we define similarity measures between molecular structures using molecular fingerprints and apply the Tanimoto similarity to quantify pairwise molecular distances. Based on this, we calculate metrics such as the internal diversity (IntDiv), which captures the average dissimilarity within the set.

An alternative approach involves computing the "number of circles" (similar to Renz et al., 2024). In this method, molecules are sequentially selected from the set, and any other molecule within a predefined distance threshold is discarded. This process is repeated until no molecules remain, and

the number of selected molecules provides an estimate of diversity. This technique is conceptually related to the leader clustering algorithm.

D.1.2 Clustering Algorithm

Molecular clustering is performed using the following algorithm:

Algorithm 1 Molecular Clustering for Diversity Assessment

Input: Set of molecules M, similarity threshold τ
Output: Number of unique clusters $n_{clusters}$
Initialize empty cluster list $C = \{\}$
for each molecule m_i in M **do**
    $assigned = False$
    for each cluster c_j in C **do**
        if $T(m_i, representative(c_j)) > \tau$ **then**
            Add m_i to cluster c_j
            $assigned = True$
            break
        end if
    end for
    if $assigned == False$ **then**
        Create new cluster with m_i as representative
        Add new cluster to C
    end if
end for
return $|C|$

D.1.3 Fragmentation Algorithm

Molecular fragmentation is performed using the following implementation based on the rdScaffoldNetwork implementation in RDKit [27]:

Algorithm 2 Fragmentation Algorithm

Input: Input molecule mol, minimum number of atoms threshold min_num_atoms
Output: List of fragment SMILES strings $fragments$
Initialize scaffold network parameters
$fragments \leftarrow \{MolToSmiles(mol)\}$ ▷ Initialize with original molecule
$net \leftarrow buildScaffoldNetwork([mol], params)$
for each $scaffold$ in $net.nodes$ **do**
    $scaffold\_mol \leftarrow MolFromSmiles(scaffold)$
    if $scaffold\_mol.GetNumAtoms() \geq min\_num\_atoms$ **then**
        $fragments.add(scaffold)$
    end if
    $removed\_frags \leftarrow removeSubstructure(mol, scaffold)$
    for each $frag$ in $removed\_frags$ **do**
        if $frag.GetNumAtoms() \geq min\_num\_atoms$ **then**
            $fragments.add(MolToSmiles(frag))$
        end if
    end for
end for
Filter out invalid SMILES:
$fragments \leftarrow \{smi \in fragments : MolFromSmiles(smi) \neq null\}$
return $fragments$ as list

D.2 String-Based Metrics Implementation

Since the generative models used are language-based—typically relying on SMILES (Simplified Molecular Input Line Entry System) representations—we can quantify molecular diversity through string-based metrics. This approach analyzes the frequency of recurring substrings, commonly referred to as n-grams.

D.2.1 N-gram Analysis

An n-gram refers to a contiguous sequence of n characters or tokens in SMILES strings. By examining how often specific n-grams occur across generated molecules, we can assess redundancy or variation in the output. High frequency of certain n-grams indicates limited diversity and over-representation of particular structural motifs, while uniform distribution suggests broader exploration of chemical space.

E Diversity Constraint Algorithms

A promising approach to promoting chemical diversity involves penalizing molecules when they exhibit excessive similarity, share common backbones, or contain overrepresented molecular fragments. These penalties are directly tied to desired chemical diversity and can be quantified through penalty scores.

Defining these penalties within the fitness function requires thorough analysis of generated molecules. Based on findings, users can adjust penalization functions accordingly. This section provides detailed algorithmic procedures for the diversity-aware constraint functions mentioned in the main text.

E.1 Similarity-Based Constraints

Similarity-based constraints focus on penalizing molecules that are too similar to previously generated compounds, thereby encouraging exploration of new chemical space.

Algorithm 3 Similarity-Based Diversity Constraint

Input: Generated molecules M_{new}, molecules to avoid M_{avoid}, threshold τ_{sim}
Output: Penalized scores for M_{new}
for each molecule m_i in M_{new} **do**
    $penalty = 1.0$ ▷ Default score multiplier
    for each molecule m_j in M_{avoid} **do**
        if $T(m_i, m_j) > \tau_{sim}$ **then**
            $penalty = 0.0$ ▷ Complete penalty
            break
        end if
    end for
    Apply penalty to score of m_i
end for
Update M_{avoid} with cluster representatives from M_{new} if cluster size $> threshold$

E.2 Fragment-Based Constraints

Fragment-based constraints target overrepresented molecular fragments to maintain substructural diversity during optimization.

Algorithm 4 Fragment-Based Diversity Constraint

Input: Generated molecules M_{new}, fragment frequency dict F_{freq}, threshold τ_{frag}
Output: Penalized scores for M_{new}
for each molecule m_i in M_{new} **do**
    $F_i = fragment(m_i)$ ▷ Extract fragments
    $penalty = 1.0$
    for each fragment f in F_i **do**
        if $F_{freq}[f] > \tau_{frag}$ **then**
            $penalty = 0.0$
            break
        end if
    end for
    Apply penalty to score of m_i
    Update F_{freq} with fragments from F_i
end for

E.3 N-Gram-Based Constraints

N-gram-based constraints monitor SMILES sequence patterns to prevent convergence toward similar string representations.

Algorithm 5 Fragment-Based Diversity Constraint

Input: Generated molecules M_{new}, fragment frequency dict F_{freq}, threshold τ_{frag}
Output: Penalized scores for M_{new}
for each molecule m_i in M_{new} **do**
    $F_i = fragment(m_i)$ ▷ Extract fragments
    $penalty = 1.0$
    for each fragment f in F_i **do**
        if $F_{freq}[f] > \tau_{frag}$ **then**
            $penalty = 0.0$
            break
        end if
    end for
    Apply penalty to score of m_i
    Update F_{freq} with fragments from F_i
end for

E.4 Implementation details for the case study

For the singlet fission case study, we employed the REINVENT4 framework with a prior model trained on an extended dataset tailored for organic electronic molecules. The dataset combines the FORMED dataset [25] and the GEOM3D dataset [26], ensuring a diverse representation of relevant chemical structures. The reinforcement learning process was conducted for 1000 iterations, generating 100 molecules per step. This setup provided a comprehensive dataset for analyzing the evolution of chemical diversity during the optimization process.

We run the reinforcement learning with the three different constraint regimes as described in the main text. For each regime we run the exploration for 5 independent runs with different random seeds. The thresholds for the constraints were established based on both percentage of molecules generated in previous steps and absolute numbers, chosen to ensure effectiveness without being overly restrictive. Specifically, for the Fragment-Based constraints, the threshold for adding a fragment to the list of fragments to avoid was set to 5% of the molecules generated in the previous steps, or a total of 50 molecules generated that contain that fragment. For the Combined constraints, thresholds were set to 10% for similarity-based constraints, 5% for fragments (considering only fragments larger than 8 non-hydrogen atoms), and 3% for 10-grams in SMILES sequences.

The results in figure 3 of the main text show the average results over the 5 independent runs. This approach provides a robust assessment of the impact of diversity constraints on molecular optimization

and diversity preservation during reinforcement learning. The difference between the different runs was small and the trends observed were consistent across all runs.

F Singlet Fission Evaluation Function

This section details the singlet fission evaluation function used in the case study.

F.1 Energy-Based Scoring

The singlet fission scoring function evaluates molecules based on the energy difference between the first singlet excited state (S_1) and twice the triplet state energy ($2T_1$):

$$\Delta E_{SF} = E(S_1) - 2 \cdot E(T_1) \tag{2}$$

For optimal singlet fission, ΔE_{SF} should be close to zero or slightly negative.

F.2 Machine Learning Model Details

The evaluation uses a pre-trained Graph Neural Network (GNN) using Chemprop [28] to predict S_1 and T_1 energies. The model was trained on the Formed dataset [25], which contains a diverse set of organic molecules with computed excited state energies. The GNN architecture includes message-passing layers to capture molecular graph information, followed by fully connected layers for energy prediction. Details of the training procedure, hyperparameters, and performance metrics can be found in the paper by Worakul et al. [24].

F.3 Additional Molecular Filters

The evaluation function includes additional sorting functions to filter out molecules that do not meet specific criteria:

- Molecular weight: 300-800 Da
- Synthetic accessibility score: ≤ 3

G Screenshots of the Web Application

Figures 4 and 5 show screenshots of the web application interface for molecular design and diversity analysis, respectively. The first figure illustrates the main interface where users can input molecules, configure analysis parameters, and visualize generated structures. This first screenshot highlights the per step metric evolution plots, the chemical space visualization, and the molecular structure viewer. The second figure showcases the fragment focused analysis results, including fragment frequency distributions and overrepresented substructures.

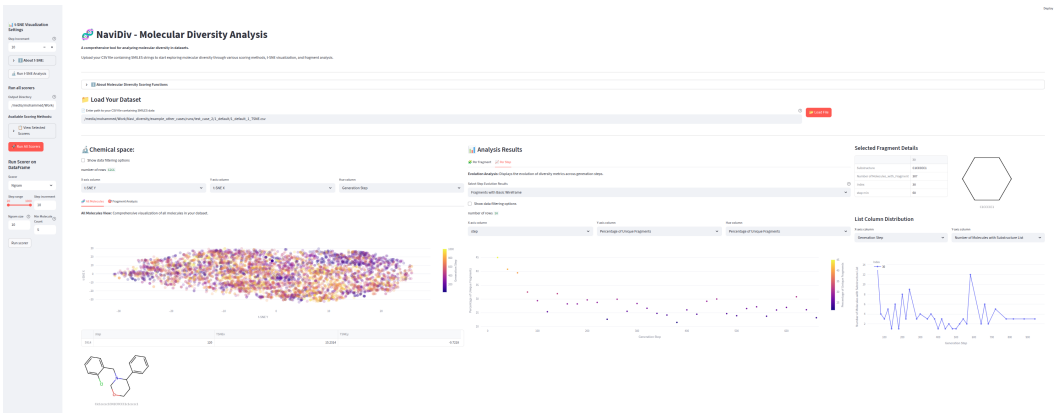

Figure 4: Screenshot of the molecular design interface.

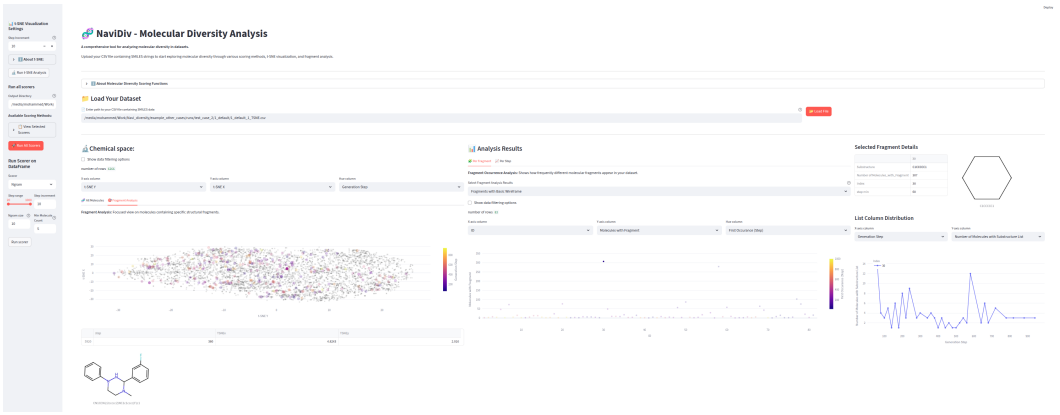

Figure 5: Screenshot of the diversity analysis results.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the contribution as a comprehensive framework for analyzing chemical diversity in generative molecular design with multiple complementary metrics, demonstrated through a singlet fission case study.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations section acknowledges the focus on a single case study, the need for broader validation, context-dependent metric selection, limited validation with other architectures, and potential gaps in capturing specialized aspects of chemical diversity.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results requiring formal proofs. It presents an analysis framework and empirical observations.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides experimental details in the main text and appendix, including the use of REINVENT4, dataset information (FORMED, GEOM3D), and 1000 iterations with 100 molecules per step.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides computational details in the main text and the code to reproduce the experiments is available in the link provided in the paper. The code includes instructions for running the experiments and reproducing the results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies key experimental details: REINVENT4 with FORMED+GEOM3D training data, 1000 RL iterations, 100 molecules per step, and singlet fission evaluation function. Additional implementation details are provided in the codebase.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides detailed information about the computational resources used for the experiments and discussed in the computational performance section.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves computational analysis of molecular structures without human subjects, privacy concerns, or potential for harmful applications. It contributes positively to scientific knowledge.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The conclusion discusses positive impacts for materials discovery, energy applications, and sustainability. The work is a tools contribution with minimal negative impact potential, focusing on improving research methodology.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites REINVENT4, FORMED, and GEOM3D datasets. However, specific license information should be added for complete compliance.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a chemical diversity analysis framework with documentation in the main text and appendix. Complete documentation will be provided with the code release.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core methodology does not involve LLMs as a component. LLMs may have been used for writing assistance but not for the scientific methodology itself.