Pharmacophore-Guided Generative Design of Novel Drug-Like Molecules

Ekaterina Podplutova[1] **Anastasia Vepreva**[1]
Olga Konovalova[1] **Vladimir Vinogradov**[1] **Dmitrii Shkil**[2] **Andrei Dmitrenko**[1,3]

[1]Center for AI in Chemistry, ITMO University, St. Petersburg, Russia
[2]Moscow Center for Advanced Studies, Moscow, Russia
[3]D ONE AG, Zurich, Switzerland

dmitrenko@pish.itmo.ru

Abstract

The integration of artificial intelligence (AI) in early-stage drug discovery offers unprecedented opportunities for exploring chemical space and accelerating hit-to-lead optimization. However, using docking as a reward function during generative model training is computationally expensive and may yield inaccurate results. Here, we present a novel generative framework that balances pharmacophore similarity to reference compounds with structural diversity from active molecules. The framework allows users to provide custom reference sets, including FDA-approved drugs or clinical candidates, and guides the *de novo* generation of potential therapeutics. We demonstrate its applicability through a case study targeting alpha estrogen receptor modulators and antagonists for breast cancer. The generated compounds maintain high pharmacophoric fidelity to known active molecules while introducing substantial structural novelty, suggesting strong potential for functional innovation and patentability. Comprehensive evaluation of the generated molecules against common drug-like properties confirms the robustness and pharmaceutical relevance of the approach.

1 Introduction

The integration of artificial intelligence (AI) in early-stage drug discovery is transforming pharmaceutical paradigms, enabling more efficient exploration of chemical space and accelerating hit-to-lead progression [1]. Traditional method for accessing biological activity is molecular docking calculation, which predicts the binding affinity between a ligand and its target protein. However, this approach is computationally expensive [2] when performed iteratively and often yields unreliable scores. Furthermore, it often oversimplifies the complex interactions involved, leading to inaccuracies. Many scoring functions are based on linear energy combinations, which may not adequately capture the nuances of protein-ligand interactions, resulting in poor correlation with experimental binding affinities [3, 4]. Pharmacophore-guided methods offer an interpretable alternative: by emphasizing the spatial arrangement of key interaction features (hydrogen-bond donors/acceptors, aromatic and hydrophobic groups), they provide a robust proxy for biological activity across diverse scaffolds. Although pharmacophore-based similarity and latent-space models exist, few approaches jointly optimize pharmacophore fidelity, fragment diversity, and docking performance. Existing frameworks like DrugMetric use VAE-based chemical space distances for molecular generation and scaffold diversity [5, 6]. Other methods focus on generative modeling of molecular latent spaces (e.g., NP-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat.

VAE, conditional -VAE), achieving high novelty scores but often sacrificing docking fidelity or pharmacophoric consistency [7, 8, 9].

In this work, we present a framework for *de novo* molecule generation that maximizes pharmacophoric similarity to reference compounds (e.g., FDA-approved drugs) while minimizing structural similarity to improve novelty and potential patentability. We demonstrate the utility of this method through a case study targeting estrogen receptor inhibitors for breast cancer. The generated compounds show strong pharmacophoric alignment with known degraders while maintaining high structural diversity. They were further validated using docking scores and synthetic accessibility. The code and data used in this study are available at: https://anonymous.4open.science/r/NeurIPS-2025-3BF8/

2 Related works

Recent advances have proposed various frameworks for pharmacophore-aware molecular generation. Zhu et al. introduced PGMG, a graph-based generative model guided by pharmacophoric constraints, which achieved high validity, novelty, and docking scores [10]. Seo and Kim developed PharmacoNet, an automated pipeline for pharmacophore model construction and scoring, which accelerates virtual screening while retaining high accuracy [11]. Yu et al. proposed DiffPhore, a diffusion-based model that learns to generate molecules conditioned on pharmacophoric maps and can predict binding poses without explicit docking [12]. Moyano-Gómez et al. presented O-LAP, which creates cavity-filling pseudo-ligands to improve docking rescoring and account for protein-ligand shape complementarity [13]. Alakhdar et al. introduced PharmaDiff, a pharmacophore-conditioned diffusion model that generates molecules satisfying 3D feature constraints with improved docking performance [14].

Our framework is docking-independent in training, relying exclusively on pharmacophore similarity as a proxy for biological relevance. Docking is used only for post hoc validation to benchmark the generated molecules against conventional approaches. Unlike PGMG and PharmaDiff, it balances scaffold novelty with pharmacophoric fidelity; unlike O-LAP and PharmacoNet, it avoids predefined binding sites, enabling early-stage exploration when structural data is lacking. This allows us to access diverse, patentable chemical space while preserving pharmacophoric patterns linked to activity.

3 Experiments

3.1 Overview of the proposed pipeline

We present a novel methodology for evaluating the biological activity of molecules that integrates both structural and pharmacophoric similarity assessments against a predefined set of reference compounds (Figure 1).

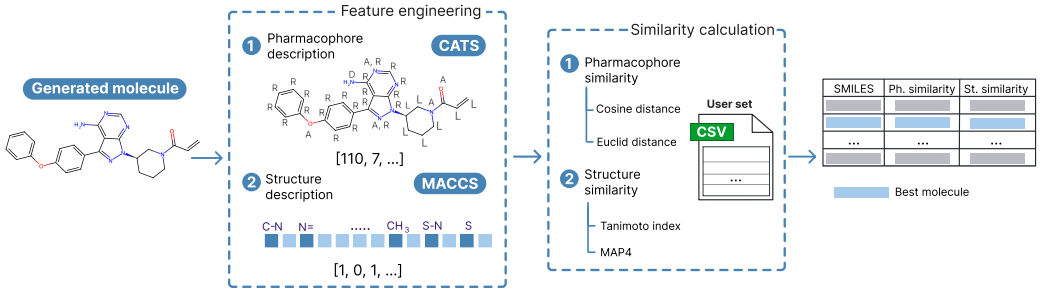


Figure 1: Schematic representation of proposed pipeline.

This approach was implemented within the reward function of the reinforcement learning (RL) model, FREED++ [15]. During each cycle of the RL process, generated molecules are encoded using two distinct molecular representations: CATS (Chemically Advanced Template Search) descriptors [16], which capture pharmacophore patterns, and MACCS (Molecular ACCess System) keys [17], which represent substructural features. To compute similarity, the resulting representations are compared to those of the molecules in a user-provided reference set. Given the distinct nature of the two representations, different similarity metrics were employed:

- Pharmacophoric similarity, derived from the continuous-valued CATS descriptors, was quantified using cosine similarity and Euclidean distance.
- Structural similarity, based on the binary MACCS fingerprints, was assessed using the Tanimoto coefficient, while MAP4 (MinHashed Atom-Pair fingerprint up to four bonds) provides a more expressive representation by combining atom-pair relationships with circular and thus shows higher scores [18].

The reward function was explicitly designed to simultaneously maximize pharmacophoric similarity and minimize structural similarity to the reference molecules. This dual-objective optimization is critical for generating novel compounds that are likely to retain the desired biological activity (guided by pharmacophore overlap) while exhibiting sufficient structural novelty to enhance their potential for patentability. (Details of the fragment vocabulary and action space definition are provided in Appendix 7.4.)

3.2 Baseline evaluation

As a reference point, we combined QED scoring with docking simulations using QVina. Docking was performed using the crystallographic structure of the alpha-estrogen receptor (PDB ID: 8AWG). The target was selected due to its central role in breast cancer pathogenesis and the availability of a high-resolution validated structure.

3.3 Reward function variants

As detailed in subsection 3.1, pharmacophore similarity was evaluated using cosine and Euclidean distances. Cosine similarity evaluates the orientation of vectors and is widely used for molecular fingerprints, while Euclidean distance captures both magnitude and direction, providing a complementary measure of dissimilarity. Structural similarity was assessed using the Tanimoto coefficient and MAP4. We tested four configurations of our reward function:

1. QED + Tanimoto + Euclidean similarity
2. QED + Tanimoto + Cosine similarity
3. QED + MAP4 + Euclidean similarity
4. QED + MAP4 + Cosine similarity

3.4 Additional profiling

Generated molecules were further evaluated with orthogonal filters. Synthetic accessibility (SA) scores estimated practical feasibility, and novelty was quantified by checking absence from ChEMBL, ZINC, and PubChem databases.

Finally, we analyzed the distributions of QED, docking scores, and molecular properties including SA, MAP4, Tanimoto, and pharmacophore similarity assessed via Euclidean and Cosine metrics subsection 7.1.

4 Results and Discussion

4.1 Overall Pharmacophore and Drug-Likeness Assessment

The evaluation of generated molecules across different reward configurations highlights the framework's ability to optimize both pharmacophoric similarity and predicted binding affinity (Table 1). The baseline molecules, generated without pharmacophore rewards, show relatively good predicted binding affinity (docking score of -8.65), complete novelty (100%), but low drug-likeness (QED of 0.30). Despite achieving more favorable docking scores, the baseline generated molecules display very low pharmacophoric similarity to established drugs, raising concerns about their biological relevance. Additionally, their synthetic accessibility remains in question (SA score of 6.28).

Introducing pharmacophore similarity and structural diversity in reward functions (Setups 1-4) led to improved molecular properties, with QED values and SA scores improving across pharmacophore-guided setups. This suggests that enforcing pharmacophoric fidelity encourages the generation of

Table 1: Evaluation of generated molecules across different reward configurations (mean ± std).

Setup	Tanimoto index (↓)	MAP4 score (↓)	Cosine similarity (↑)	Euclid similarity (↓)	QED (↑)	Docking score (↓)	SA score (↓)	Novelty (↑)
Baseline	**0.34 ± 0.05**	**0.03 ± 0.01**	0.58 ± 0.27	70.3 ± 13.03	0.30 ± 0.08	**-8.64 ± 1.03**	6.28 ± 0.64	**100**
Setup 1	**0.34 ± 0.05**	0.04 ± 0.01	**0.94 ± 0.06**	**34.80 ± 7.84**	0.33 ± 0.13	-6.49 ± 1.17	4.64 ± 0.51	**100**
Setup 2	0.36 ± 0.05	**0.03 ± 0.01**	0.83 ± 0.05	54.92 ± 8.60	**0.59 ± 0.16**	-6.71 ± 0.55	4.72 ± 0.49	99.6
Setup 3	0.35 ± 0.05	0.04 ± 0.01	**0.94 ± 0.06**	50.47 ± 10.16	0.44 ± 0.16	-7.09 ± 0.66	4.67 ± 0.45	84.5
Setup 4	0.35 ± 0.05	**0.03 ± 0.01**	0.87 ± 0.07	38.92 ± 9.37	0.34 ± 0.15	-6.47 ± 1.02	**4.61 ± 0.50**	**100**

more drug-like and synthetically accessible molecules. The impact of different similarity metrics on these property profiles is visually assessed on Figure 2. Specifically, the QED distribution (Figure 2a) for the baseline is concentrated around 0.3-0.4, while MAP4 + Cosine similarity shifts this distribution towards higher values (peak near 0.6-0.7), indicating improved drug-likeness. Similarly, the SA distribution (Figure 2c) shows a lower peak for the other methods in comparison to the baseline which has a peak at 4, suggesting improved synthetic accessibility. The docking score distribution (Figure 2b) is shifted towards less negative values for all setups compared to the baseline (peak around -8), indicating lower binding affinity. However, the average docking score of the known alpha-estrogen receptor modulators and antagonists, which served as the basis for the pharmacophore descriptors, was -6.64. This allows us to conclude that all four proposed setups are, in fact, comparable to the confirmed receptor modulators and antagonists in binding affinity, assessed by the docking score. Furthermore, cosine similarity (Figure 2f) is higher for MAP4 + Cosine similarity compared to Tanimoto + Cosine similarity which has a peak near 0.7, indicating that the MAP4 + Cosine similarity method generates structures with a higher average cosine similarity score.

MAP4 provides a rich molecular representation, encoding atom-pair relationships and leveraging MinHash to capture global topology and local motifs efficiently. Pharmacophoric and structural similarity values remain comparable across all reward setups, showing that our framework generates molecules with favorable predicted binding affinity, drug-likeness, and structural novelty.

In Figure 3, representative generated molecules and their reference analogs (one per reward setting) reproduce key pharmacophoric patterns, tri-aromatic/heteroaromatic motifs with similar linker lengths, while reshaping scaffolds. Even though docking score improvement is notable mainly in the MAP4 + cosine setup, the top molecules exhibit higher QED than reference degraders.

These results indicate that our reward functions drive convergence on biologically meaningful pharmacophoric arrangements (aromatic triads, conserved H-bond vectors, hydrophobic spacers) without collapsing to close structural analogs, balancing functional similarity and scaffold novelty.

4.2 Methodological limitations

This study has several methodological limitations. While the generated molecules show high pharmacophoric similarity to known degraders, their docking scores and QED remain moderate. The use of a limited set of pharmacophore descriptors may also restrict scaffold diversity. Future work will extend the approach with richer pharmacophore representations, additional similarity metrics, and diverse generative models to improve both biological relevance and chemical novelty, ultimately enabling synthesis and biological validation.

5 Conclusion and Future work

We proposed a pharmacophore-guided generative approach for designing potentially active and selective molecules using a reinforcement learning model. Pharmacophoric similarity was evaluated with CATS descriptors using Euclidean and cosine metrics, while structural novelty was encouraged by minimizing similarity based on MACCS descriptors using the classical Tanimoto coefficient, as well as the recently proposed MAP4 metric. In a case study targeting estrogen receptor inhibitors for breast cancer, the generated compounds showed high pharmacophoric similarity to known actives and low structural similarity, suggesting strong novelty and patentability. All molecules also met basic drug-like criteria, supporting the method's potential for further development and experimental validation.

6 Acknowledgment

This work supported by the Ministry of Economic Development of the Russian Federation (IGK 000000C313925P4C0002), agreement No139-15-2025-010.

References

[1] Pandiella A. Privat C. Ocana, A. Integrating artificial intelligence in drug discovery and early drug development: a transformative approach. *Biomarker Research*, 13(1):1–11, 2025.

[2] P. H. M. Torres, A. C. R. Sodero, P. Jofily, and F. P. Silva-Jr. Key topics in molecular docking for drug design. *Int J Mol Sci*, 20(18):4574, September 2019.

[3] X. Zheng et al. Challenges in scoring functions for molecular docking. *Journal of Computational Chemistry*, 43:1200–1215, 2022.

[4] H. Yu et al. Deep learning approaches for protein-ligand binding prediction. *Bioinformatics*, 39:500–512, 2023.

[5] B Li, Y Zhang, J Li, L Zhang, Z Li, L Wang, X Zhang, and Y Li. Drugmetric: quantitative drug-likeness scoring based on chemical space distance. *Bioinformatics*, 40(1):1–9, 2024.

[6] K Lee, J Kim, S Lee, H Lee, and S Lee. Drug-likeness scoring based on unsupervised learning. *Nature Communications*, 12(1):1–9, 2021.

[7] B Wu, K Zheng, and L Zhang. Cross-adversarial learning for molecular generation in drug design. *Frontiers in Pharmacology*, 12:827606, 2022.

[8] Y Gadiya, A Kumar, and S Singh. Pemt: a patent enrichment tool for drug discovery. *Bioinformatics*, 39(1):btac716, 2023.

[9] Deepak Paul, Gaurav Sanap, Shreyas Shenoy, Dev Kalyane, Kailash Kalia, and Rakesh K. Tekade. Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1):80–93, 2021.

[10] L. Zhu et al. Pgmg: Pharmacophore-guided molecule generation using graph neural networks. *Journal of Medicinal Chemistry*, 66:1234–1248, 2023.

[11] K. Seo and H. Kim. Pharmaconet: automated pharmacophore modeling and scoring. *Bioinformatics*, 40:2100–2115, 2024.

[12] H. Yu et al. Diffphore: Diffusion-based pharmacophore-guided lead discovery. *Nature Machine Intelligence*, 7:150–162, 2025.

[13] P. Moyano-Gómez et al. O-lap: Cavity-filling pseudo-ligands for docking rescoring. *Journal of Chemical Information and Modeling*, 64:2500–2512, 2024.

[14] R. Alakhdar et al. Pharmadiff: Pharmacophore-conditioned diffusion models for drug discovery. *ACS Central Science*, 11:1000–1012, 2025.

[15] Alexander Telepov, Artem Tsypin, Kuzma Khrabrov, Sergey Yakukhnov, Pavel Strashnov, Petr Zhilyaev, Egor Rumiantsev, Daniel Ezhov, Manvel Avetisian, Olga Popova, and Artur Kadurin. Freed++: Improving rl agents for fragment-based molecule generation by thorough reproduction, 2024.

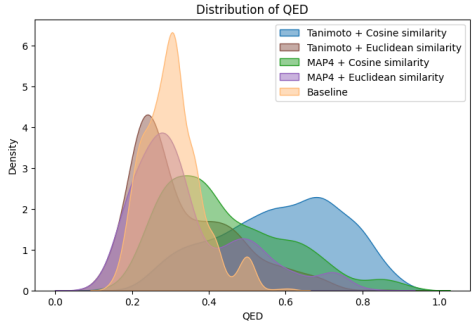
[16] Michael Reutlinger, Christian P. Koch, Daniel Reker, Nickolay Todoroff, Petra Schneider, Tiago Rodrigues, and Gisbert Schneider. Chemically advanced template search (cats) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Molecular Informatics*, 32(2):133–138, February 2013.

[17] Jingbo Yang, Yiyang Cai, Kairui Zhao, Hongbo Xie, and Xiujie Chen. Concepts and applications of chemical fingerprint for hit and lead screening. *Drug Discovery Today*, 27(11):103356, November 2022.

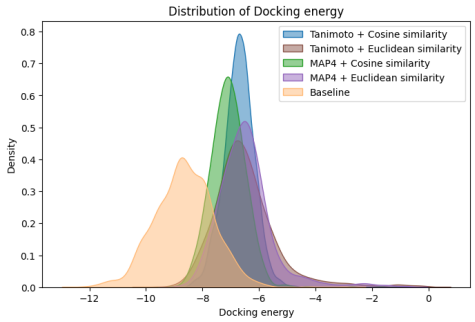
[18] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, 2020.

7 Technical Appendices and Supplementary Material

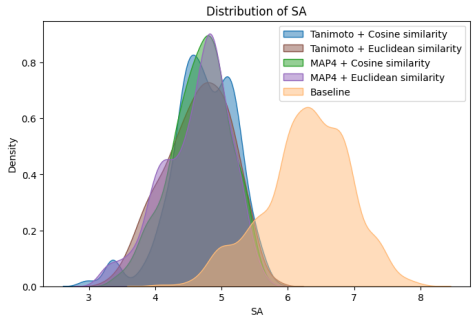
7.1 Distribution of key properties evaluated in experiments



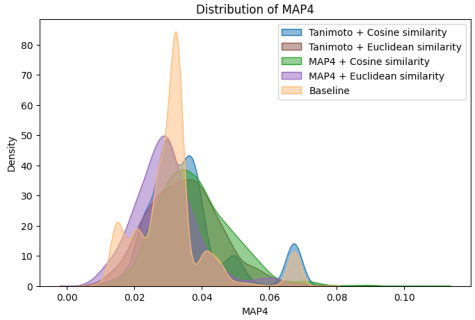
(a) QED



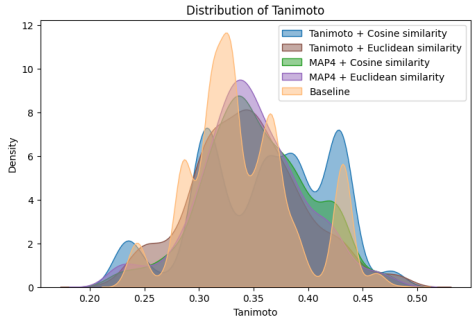
(b) Docking energy



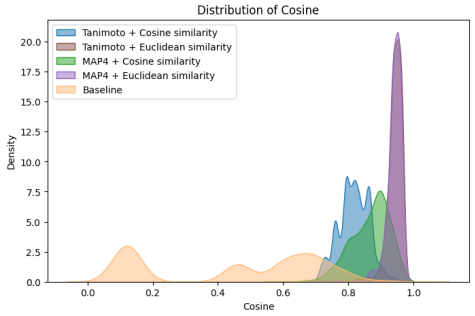
(c) SA



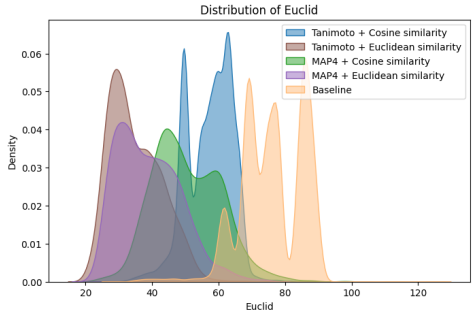
(d) MAP4



(e) Tanimoto



(f) Cosine



(g) Euclid

Figure 2: Distributions of key properties evaluated in experiments.

7.2 Reference molecule set

The reference set used in this study comprises 53 estrogen receptor antagonists and degraders. The full list of SMILES strings is available in the supplementary GitHub repository: `https://anonymous.4open.science/r/NeurIPS-2025-3BF8/`.

7.3 Reward weighting and normalization.

The total reward was computed as a weighted sum of property-specific objectives. We empirically assigned weights of 1 to the *QED* and docking-related terms, and 2 to all other components (pharmacophore and structural similarity metrics such as *CATS*, *MAP4*, and *Tanimoto*). Each term was scaled to the $[0, 1]$ range using min–max normalization within each batch to balance the magnitude of different objectives. The final reward function was defined as:

$$R = \sum_i w_i r_i,$$

where w_i is the assigned weight and r_i is the normalized property-specific score. This configuration provided stable convergence and balanced optimization between drug-likeness, pharmacophoric fidelity, and structural novelty.

7.4 Fragment vocabulary and action space

The FREED++ action space was constructed from a hybrid fragment vocabulary. Radical substituents and attachment rules were inherited from the original FREED++ implementation to preserve valid connection patterns and valency constraints. Starting scaffolds were obtained by parsing the ChemDiv database and decomposing molecules into Murcko scaffolds to ensure structural diversity. Representative, non-redundant cores were selected based on scaffold uniqueness without additional filtering for synthetic accessibility or physicochemical properties. All generated molecules were subsequently evaluated *post hoc* for QED, SA, novelty, and docking scores as described in Section 3.4.

7.5 Best generated molecules

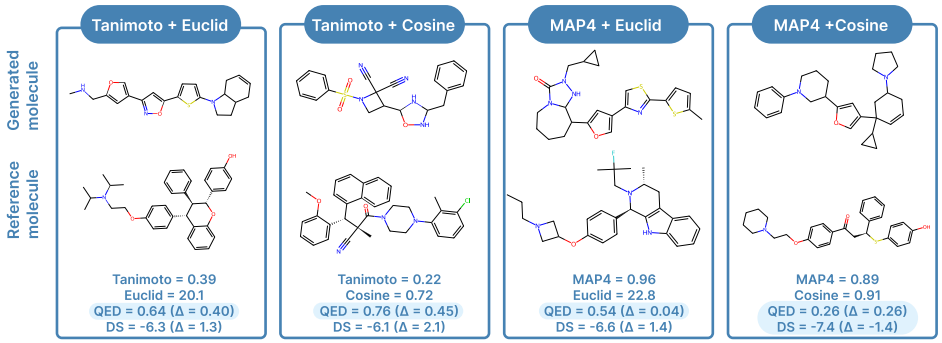


Figure 3: Best generated molecules and their pharmacophore analogue.

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 consistently highlight the key methodological contribution pharmacophore-guided generative design and frame it within the context of existing challenges in drug discovery.

   Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations section explicitly acknowledges methodological constraints, including reliance on docking as a proxy for biological activity, the potential bias introduced by pharmacophore similarity metrics, and the restricted scope imposed by reference sets. The absence of wet-lab validation is also stated, reinforcing transparency.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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 complete disclosure of experimental settings, including dataset composition, choice of reference molecules, pharmacophore descriptors, similarity metrics, model architecture, and training parameters. These details are sufficient to reproduce the reported results and support the main conclusions, independent of access to code or supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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 authors release both the trained models and generation pipeline with detailed usage instructions. Data splits, input reference molecules, and evaluation metrics are included in the GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: Training and evaluation settings are reported in GitHub.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: [Yes]

   Justification: The results for all experimental setups are now accompanied by standard deviations for each metric, providing clear quantitative measures of variability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: Calculations were performed on a server with an NVIDIA A6000 GPU (20 GB RAM).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 study complies with ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

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 discussion covers positive impacts, such as accelerating drug discovery and enabling patentable molecule design, as well as risks, including potential misuse of generative methods to design harmful compounds.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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 clearly credits the sources of all external assets, including molecular datasets, similarity metrics and filtering tools. References to the original publications are provided, and the use of these resources complies with their respective licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

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 newly introduced assets are described in detail within the paper and supplementary material. Clear documentation and usage instructions are provided alongside the released code repository, ensuring that other researchers can easily understand, reproduce, and extend the work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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 method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.