DARWIN: A Framework for Target Specific Diversity Constrained Natural Product like Molecule Generation

Aditi Vijil [1] **Pin-Yu Chen** [2]

Abstract

The rising global incidence of incurable diseases underscores the persistent gaps in drug discovery and development, largely rooted in the limited chemical diversity of modern drug compounds. Natural products (NP)—chemical metabolites produced by living organisms—offer a rich reservoir of diversity to address this limitation. Therefore, this study develops a novel framework, DARWIN, a genetic-algorithm based framework, that leverages the diversity of NPs and the scalability of computational techniques to propose novel Natural Product like drug candidates. While genetic algorithms have been widely used in molecule optimization, the molecules generated by them are known to lack diversity. DARWIN supports fine-grained control over molecular diversity by incorporating intermolecular similarity directly within the generation process. Since they are based on genetic algorithms, they are extremely efficient without the need for expensive pretraining on GPUs, or finetuning for targeted generation. When applied to two targets implicated in Ewing Sarcoma and Chronic Lymphocytic Leukemia, the generated molecules demonstrated improved properties relative to the DOCKSTRING baseline. Overall, DARWIN provides a novel, controllable framework for expanding the search for drug candidates beyond synthetic libraries, offering an effective method for accelerating drug discovery for currently incurable diseases.

1. Introduction

Despite decades of intensive research, many diseases remain without a cure. Of the 7,000 identified rare diseases, 95% lack an FDA-approved treatment (Fermaglich & Miller, 2023). The drug discovery pipeline itself is time-intensive and laborious; developing a single drug takes roughly 10 years and $2 billion, yet 90% of clinical drug development fails (DiMasi et al., 2016; Sun et al., 2022). This growing demand and persistent lack of effective therapeutics—despite substantial investment—underscores the urgency of exploring more diverse, less conventional chemical spaces for drug discovery. However, roughly 70% of FDA-approved small molecule drugs were based on known scaffolds (Taylor et al., 2017). This reliance on known molecular structures constrains the search for novel therapeutics; by repeatedly selecting drug candidates from a narrow pool, we restrict our search, and promising treatments remain undiscovered.

Natural products (NPs) are a vast range of specialized metabolites produced by bacteria, plants, fungi, and animals. These compounds comprise hundreds of thousands of diverse chemical structures—ranging from peptides, polyketides, saccharides, terpenes, and alkaloids—that allow organisms to thrive in specific environments (Mullowney et al., 2023). Across different organisms, natural products play versatile roles in complex interactions, serving as signals, weapons, nutrient-scavenging agents, pest protectants, and defense against environmental stressors (Al-Khayri et al., 2023; Mullowney et al., 2023). Shaped by millions of years of evolutionary pressure, natural products have been structurally optimized by nature to interact precisely with biological macromolecules and increase bioactivity, membrane-permeability, and metabolic stability. The sheer structural and chemical diversity of natural products has the potential to revitalize the current drug landscape; however, recently, natural product–based drug discovery has declined in prominence relative to synthetic approaches, largely because synthetic compound libraries vastly outnumber known natural products (Atanasov et al., 2021; Beutler, 2009).

In this paper, we propose DARWIN, a framework for targeted generation of Natural Product like molecules, based on genetic algorithms. Although genetic algorithms have been used successfully for molecule optimization (Tripp & Hernández-Lobato, 2023; Jensen, 2019), the molecules generated by them are known to lack diversity. To alleviate this problem, we explicitly incorporate intermolecular similarity as a constraint in the generation process, allowing

[1]Horace Greeley High School, Chappaqua, NY 10514, USA
[2]IBM Research, Yorktown Heights, NY 10598, USA. Correspondence to: Pin-Yu Chen <pin-yu.chen@ibm.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

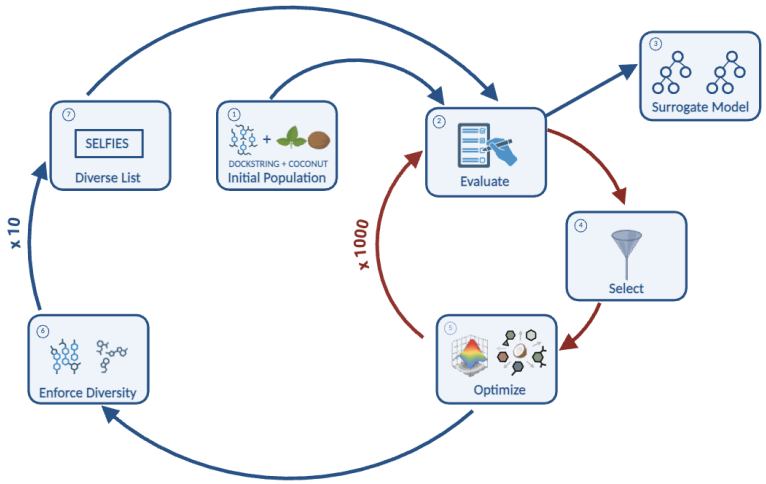

Figure 1. Architecture of DARWIN. The algorithm proceeds in 10 phases, where in each phase molecules are optimized using genetic algorithms for 1000 generations. At the end of each phase, diversity constraints are enforced so that molecules have high intermolecuar similarity.

for fine-grained control over the diversity of the generated molecules. The generation process is extremely efficient and does not require expensive pretraining, finetuning or reinforcement learning on GPUs. We demonstrate the efficacy of the method by generating natural product like molecules that can bind to two targets–IGF1R (implicated in Ewing Sarcoma) and LCK (implicated in Chronic Lymphocytic Leukemia). Our experiments show that DARWIN is able to optimize for multiple desirable properties (binding to target, molecule weight, synthetic accessibility and Natural product likeness score), while still retaining high diversity of the generated molecules. Overall, DARWIN provides a light-weight approach for target specific molecule generation of natural product like molecules, with explicit control for diversity.

2. Methods

2.1. Surrogate Model for Binding Affinity

A core optimization objective is the strong binding affinity to selected targets. Molecular docking provides a well-established, physics-informed approximation of protein–ligand binding by estimating binding free energy and ranking relative affinities across candidate molecules. However, docking simulations require ~15 seconds per molecule, which is often inefficient for large-scale projects. Instead, a surrogate model to predict binding free energy was built to integrate directly into the optimization framework, allowing for efficient optimization of molecules while accounting for binding affinity. The details of the surrogate model are presented in Appendix A.

2.2. DARWIN Framework

In order to generate diverse natural product-like molecules, a novel genetic algorithm-based optimization framework was developed that explicitly enforces molecular diversity. Since optimization algorithms, typically cannot optimize for population level metrics (like diversity), we use a staged approach. The algorithm starts with an initial population of molecules (in our case, 5000 molecules each taken from COCONUT (Chandrasekhar et al., 2025), a dataset of Natural Products, and DOCKSTRING). The fitness function is evaluated for all 10,000 molecules, after which the surviving molecules are selected for the next generation. This is done through uniform quantile sampling, where a fixed number of molecules from the top quantiles are sampled to prevent a small number of molecules from dominating the space. Then, a few molecules are selected as parents and undergo operations like crossover or mutation in a probabilistic fashion. After all the offspring are modified through mutation and/or crossover, they are added to the population and the fitness function is evaluated. Finally, the survivors for the next generation are picked and the process is repeated for 1,000 generations. This methodology is illustrated in Figure 1.

While this process often succeeds in optimizing the fitness function, the highest-scoring molecules tend to be minor variants of a limited set of scaffolds, thereby limiting structural diversity. To obtain finer grained control on the diversity of the generated molecules, an iterative diversity guided generation method is created. This is shown in Algorithm 1.

The algorithm works by running the genetic algorithm (GA) described above for a fixed number of cycles

Table 1. Comparison of NP-likeness scores and docking probabilities between DOCKSTRING baseline molecules and DARWIN-generated molecules for IGF1R and LCK.

	NP-LIKENESS SCORE		DOCKING PROBABILITY	
DATASET	IGF1R	LCK	IGF1R	LCK
DOCKSTRING (BASELINE)	-1.20	-1.20	0.09	0.08
DARWIN (GENERATED)	1.15	1.25	0.84	0.80

(`num_cycles` = 10). We start with an initial `diverse_list` which is empty. After each invocation of the GA algorithm, only those molecules which have a similarity less than `sim_threshold` to all molecules in the `diverse_list` are added to the `diverse_list`. Here, similarity is calculated using the Tanimoto similarity function over Morgan fingerprints.

The fitness function used for genetic algorithm includes four components. (1) Molecule Weight Score: if the molecule weight is between 100 and 600 return 1, else return 0. (2) Docking Probability: calculated using the surrogate model described above. (3) SA Score: return 0 if Synthetic Accessibility Score is > 5, else return 1. (4) Normalized NP Score: normalized NP Scored between 0 and 1. If NP Score > 2, return 1.

The final fitness fitness function is the geometric mean of these four components. After the first GA iteration, any molecule with similarity greater than `sim_threshold` to all molecules in the `diverse_list` is assigned a fitness score of 0 and is de-prioritized in selection for the next generation. Hence, diversity is accounted for at the end of each cycle, and during the optimization process, in the fitness function.

Algorithm 1: DARWIN: Iterative diversity-guided natural-product generation

Require: GA()
  Inputs: num_cycles, k (diverse molecules per cycle), $sim_threshold$, $constraints$
  $diverse_list \leftarrow \emptyset$
  for $i = 1$ num_cycles **do**
    $generated_list \leftarrow$ GA($diverse_list, sim_threshold, constraints$)
    // generated_list assumed sorted by fitness score
    $count \leftarrow 0$
    for each molecule m **in** $generated_list$ **do**
      $s \leftarrow$ MAX_SIMILARITY($diverse_list, m$)
      if $s < sim_threshold$ **then**
        $diverse_list \leftarrow diverse_list \cup \{m\}$
        $count \leftarrow count + 1$
      end if
      if $count = k$ **then**
        break
      end if
    end for
  end for
  return $diverse_list$

3. Results

3.1. Property Optimization using DARWIN

The DARWIN framework was used to generate molecules with high natural product-likeness, synthesizability, and binding affinity towards the selected targets. A sample of generated molecules for each target is shown in Figure 8 along with computed properties like molecule weight, predicted docking score, NP Score, and Synthetic Accessibility scores.

Table 1 compares the NP-likeness scores and docking probabilities of DARWIN-generated molecules against a random sample from the DOCKSTRING dataset for both targets. Across both metrics, the generated molecules consistently and substantially outperform the baseline. As visualized in Fig. 2 (Appendix D), the generated molecules achieve markedly higher NP-likeness scores than the DOCKSTRING baseline for both IGF1R (Fig. 2a) and LCK (Fig. 2b). This trend is similarly reflected in Fig. 4, where generated molecules demonstrate far higher docking probabilities for IGF1R (Fig. 4a) and LCK (Fig. 4b), confirming that they exhibit stronger target binding while retaining natural product likeness (Appendix D).

Figure 5 depicts the joint distribution of NP-likeness and predicted docking probabilitues for both generated molecules and the DOCKSTRING baseline, providing a comparative view of the spread and density of compounds across the two datasets. Notably, generated molecules displayed a concentrated distribution with concurrently high NP-likeness scores and strong predicted docking probabilities, whereas the DOCKSTRING dataset had a broader spread with a larger proportion of compounds possessing low NP-likeness and weaker docking scores.

3.2. Diversity and Synthetic Accessibility

DARWIN was able to further optimize for high synthesizability, retaining a synthetic accessibility score of less than 5 for all proposed candidates. By directly accounting for

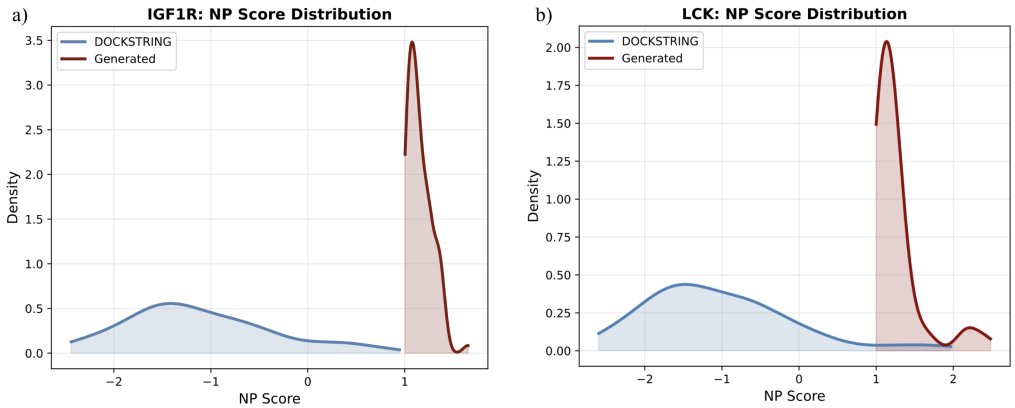

Figure 2. (a) Natural product-likeness scores for the DOCKSTRING molecules and generated molecules for IGF1R. (b) Natural product-likeness scores for the DOCKSTRING molecules and generated molecules for LCK.

diversity, generated molecules maintained a tanimoto similarity of at most 0.6 to any other generated molecule. This indicates that DARWIN effectively optimized NP-likeness, diversity, synthesizability, and predicted binding affinity, and the framework accurately preserved internal diversity while simultaneously improving multiple objectives. As a central goal of this study was to create diverse yet optimized molecules, these findings demonstrate that the framework successfully achieved this balance. Overall, the resulting generated molecules exhibit improved alignment with the biological diversity of natural products and the simultaneous optimization of drug-like properties, demonstrating DARWIN's potential in multi-objective optimization and the exploration of uncharted chemical space.

3.3. Docking Using Autodock Vina

After the best candidates were identified based on the prediction of the surrogate model and NP scores, docking was performed on the final list of 100 candidates to validate our results. The docking was performed using Autodock Vina software using the python API provided with the DOCKSTRING package. This ensures that the same setup was used for docking the molecules generated using DARWIN and the DOCKSTRING dataset. The distribution of binding free energy for the generated molecules and 100 random molecules from the DOCKSTRING dataset is shown in Figure 6 (Appendex E). The mean binding free energy for LCK was -11.132 kcal/mol (compared to -7.97 kcal/mol for 100 random molecules from DOCKSTRING), and for IGF1R was -10.280 kcal/mol (compared to -7.95 kcal/mol for 100 random molecules from DOCKSTRING). The best candidate for LCK had a score of -13.2 kcal/mol and the best candidate for IGF1R had a score of -12.1 kcal/mol. This indicates that DARWIN is able to generate novel molecules with high docking scores, while simultaneously maintaining low intermolecular similarity, high NP Score, and favorable synthetic accessibility.

The docking poses for a sample of generated candidates are given below in Figure 7 (Appendix E), indicating the optimized molecules exhibit stable binding conformations within the target protein's active site.

4. Toxicity and Retrosynthesis Prediction

The toxicity and retrosynthesis pathways for the generated molecules were analyzed and the details of this are reported in Appendix B and Appendix C respectively.

5. Conclusion

In this paper, we presented DARWIN, a framework to design diverse, novel, natural-product-like molecules further optimized for desirable drug properties. While natural evolution optimizes molecular function through selective pressure over millions of years, DARWIN can generate novel natural-product-like molecules that satisfy multiple favorable properties, including binding to any target, in a scalable manner. Unlike prior approaches that largely neglect the importance of optimizing for molecule diversity, despite its implications in treating incurable diseases, DARWIN explicitly incorporates diversity as a core design constraint. Increasingly, AI-driven molecular design is reshaping early lead identification and reducing development cost (Zhang et al., 2025). Integrating natural-product derived scaffolds within these frameworks restores the structural novelty that has declined in recent decades, addressing one of the fundamental limitations with in-silico therapeutic pipelines. DARWIN's flexibility allows it to be applied broadly, offering a pathway to address urgent global public health challenges, such as antimicrobial resistance and development of therapeutics for currently incurable diseases.

References

Al-Khayri, J. M., Rashmi, R., Toppo, V., Chole, P. B., Banadka, A., Sudheer, W. N., Nagella, P., Shehata, W. F., Al-Mssallem, M. Q., Alessa, F. M., et al. Plant secondary metabolites: the weapons for biotic stress management. *Metabolites*, 13(6):716, 2023.

Asano, D., Takakusa, H., and Nakai, D. Oral absorption of middle-to-large molecules and its improvement, with a focus on new modality drugs. *Pharmaceutics*, 16(1):47, 2023.

Atanasov, A. G., Waltenberger, B., Pferschy-Wenzig, E.-M., Linder, T., Wawrosch, C., Uhrin, P., Temml, V., Wang, L., Schwaiger, S., Heiss, E. H., et al. Discovery and resupply of pharmacologically active plant-derived natural products: A review. *Biotechnology advances*, 33(8): 1582–1614, 2015.

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., and Supuran, C. T. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3): 200–216, 2021.

Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

Beutler, J. A. Natural products as a foundation for drug discovery. *Current protocols in pharmacology*, 46(1): 9–11, 2009.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., and Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5):e1608, 2022. ISSN 1759-0884. doi: 10.1002/wcms.1608. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1608. _eprint: https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1608.

Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., and Chen, H. Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, 37(1-2): 1700123, 2018.

Chandrasekhar, V., Rajan, K., Kanakam, S. R. S., Sharma, N., Weißenborn, V., Schaub, J., and Steinbeck, C. Coconut 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Research*, 53(D1):D634–D643, 2025.

Clemons, P. A., Bodycombe, N. E., Carrinski, H. A., Wilson, J. A., Shamji, A. F., Wagner, B. K., Koehler, A. N., and Schreiber, S. L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proceedings of the National Academy of Sciences*, 107(44):18787–18792, 2010.

Dai, H., Tian, Y., Dai, B., Skiena, S., and Song, L. Syntax-directed variational autoencoder for structured data. *arXiv:1802.08786*, 2018.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47:20–33, 2016.

Domingo-Fernández, D., Gadiya, Y., Preto, A. J., Krettler, C. A., Mubeen, S., Allen, A., Healey, D., and Colluru, V. Natural products have increased rates of clinical trial success throughout the drug development process. *Journal of Natural Products*, 87(7):1844–1851, 2024.

Fermaglich, L. J. and Miller, K. L. A comprehensive study of the rare diseases and conditions targeted by orphan drug designations and approvals over the forty years of the orphan drug act. *Orphanet journal of rare diseases*, 18(1):163, 2023.

García-Ortegón, M., Simm, G. N., Tripp, A. J., Hernández-Lobato, J. M., Bender, A., and Bacallado, S. Dockstring: easy molecular docking yields better benchmarks for ligand design. *Journal of chemical information and modeling*, 62(15):3486–3502, 2022.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Gottipati, S. K., Sattarov, B., Niu, S., Pathak, Y., Wei, H., Liu, S., Thomas, K. M. J., Blackburn, S., Coley, C. W., Tang, J., Chandar, S., and Bengio, Y. Learning to navigate the synthetically accessible chemical space using reinforcement learning, 2020. URL https://arxiv.org/abs/2004.12485.

Grigalunas, M., Burhop, A., Zinken, S., Pahl, A., Gally, J.-M., Wild, N., Mantel, Y., Sievers, S., Foley, D. J., Scheel, R., et al. Natural product fragment combination to performance-diverse pseudo-natural products. *Nature Communications*, 12(1):1883, 2021.

Gupta, A., Müller, A. T., Huisman, B. J., Fuchs, J. A., Schneider, P., and Schneider, G. Generative recurrent networks for de novo drug design. *Molecular Informatics*, 37(1-2):1700111, 2018.

Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv:1802.04364*, 2018.

Kim, H., Kim, M., Choi, S., and Park, J. Genetic-guided gflownets for sample efficient molecular optimization, 2024. URL https://arxiv.org/abs/2402.05961.

Li, Y., Vinyals, O., Dyer, C., Pascanu, R., and Battaglia, P. Learning deep generative models of graphs. *arXiv:1803.03324*, 2018a.

Li, Y., Zhang, L., and Liu, Z. Multi-objective de novo drug design with conditional graph generative model. *Journal of Cheminformatics*, 10(1):33, 2018b.

Ligon, B. L. Penicillin: its discovery and early development. In *Seminars in pediatric infectious diseases*, volume 15, pp. 52–57. Elsevier, 2004.

Lim, J., Ryu, S., Kim, J. W., and Kim, W. Y. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, 10(1):31, 2018.

Liu, Y., Cai, M., Zhao, Y., Hu, Z., Wu, P., and Kong, D.-X. Time-dependent comparison of the structural variations of natural products and synthetic compounds. *International Journal of Molecular Sciences*, 25(21):11475, 2024.

Loeffler, H. H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L. H., and Engkvist, O. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

Luker, T., Alcaraz, L., Chohan, K. K., Blomberg, N., Brown, D. S., Butlin, R. J., Elebring, T., Griffin, A. M., Guile, S., St-Gallay, S., et al. Strategies to improve in vivo toxicology outcomes for basic candidate drug molecules. *Bioorganic & medicinal chemistry letters*, 21(19):5673–5679, 2011.

Ma, N., Zhang, Z., Liao, F., Jiang, T., and Tu, Y. The birth of artemisinin. *Pharmacology & therapeutics*, 216:107658, 2020.

Mullowney, M. W., Duncan, K. R., Elsayed, S. S., Garg, N., van der Hooft, J. J., Martin, N. I., Meijer, D., Terlouw, B. R., Biermann, F., Blin, K., et al. Artificial intelligence for natural product drug discovery. *Nature Reviews Drug Discovery*, 22(11):895–916, 2023.

Newman, D. J. and Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. *Journal of natural products*, 79(3):629–661, 2016.

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

Patridge, E., Gareiss, P., Kinch, M. S., and Hoyer, D. An analysis of fda-approved drugs: natural products and their derivatives. *Drug discovery today*, 21(2):204–207, 2016.

Ritchie, T. J. and Macdonald, S. J. F. The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? *Drug Discovery Today*, 14(21):1011–1020, November 2009. ISSN 1359-6446. doi: 10.1016/j.drudis.2009.07.014. URL https://www.sciencedirect.com/science/article/pii/S1359644609002785.

Samanta, B., Abir, D., Jana, G., Chattaraj, P. K., Ganguly, N., and Rodriguez, M. G. NeVAE: a deep generative model for molecular graphs. In *AAAI*, volume 33, pp. 1110–1117, 2019.

Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4 (1):120–131, 2017.

Stratton, C. F., Newman, D. J., and Tan, D. S. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorganic & Medicinal Chemistry Letters*, 25(21):4802–4807, November 2015. ISSN 1464-3405. doi: 10.1016/j.bmcl.2015.07.014.

Sun, D., Gao, W., Hu, H., and Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica. B*, 12(7):3049–3062, July 2022. ISSN 2211-3835. doi: 10.1016/j.apsb.2022.02.002. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9293739/.

Taylor, R. D., MacCoss, M., and Lawson, A. D. G. Combining Molecular Scaffolds from FDA Approved Drugs: Application to Drug Discovery. *Journal of Medicinal Chemistry*, 60(5):1638–1647, March 2017. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.6b01367. URL https://doi.org/10.1021/acs.jmedchem.6b01367.

Tripp, A. and Hernández-Lobato, J. M. Genetic algorithms are strong baselines for molecule generation. *arXiv preprint arXiv:2310.09267*, 2023.

Waksman, S. A. Streptomycin: background, isolation, properties, and utilization. *Science*, 118(3062):259–266, 1953.

Waltenberger, B., Mocan, A., Šmejkal, K., Heiss, E. H., and Atanasov, A. G. Natural products to counteract the epidemic of cardiovascular and metabolic disorders. *Molecules*, 21(6):807, 2016.

Wang, E., Schmidgall, S., Jaeger, P. F., Zhang, F., Pilgrim, R., Matias, Y., Barral, J. K., Fleet, D. J., and Azizi, S. Txgemma: Efficient and agentic llms for therapeutics. *CoRR*, abs/2504.06196, 2025a. doi: 10.48550/ARXIV.2504.06196. URL https://doi.org/10.48550/arXiv.2504.06196.

Wang, L., Song, C., Liu, Z., Rong, Y., Liu, Q., Wu, S., and Wang, L. Diffusion Models for Molecules: A Survey of Methods and Tasks, February 2025b. URL http://arxiv.org/abs/2502.09511. arXiv:2502.09511 [cs].

Zhang, K., Yang, X., Wang, Y., Yu, Y., Huang, N., Li, G., Li, X., Wu, J. C., and Yang, S. Artificial intelligence in drug development. *Nature Medicine*, 31(1): 45–59, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03434-4. URL https://www.nature.com/articles/s41591-024-03434-4.

Zhang, M.-Q. and Wilkinson, B. Drug discovery beyond the 'rule-of-five'. *Current Opinion in Biotechnology*, 18 (6):478–488, December 2007. ISSN 0958-1669. doi: 10.1016/j.copbio.2007.10.005.

A. Details of the Surrogate Model

To build the surrogate model, docking data for 260,000 molecules were obtained from the DOCKSTRING dataset (García-Ortegón et al., 2022), which contains docking scores and poses for 58 medically relevant targets. A classification and regression model were trained for our chosen targets (LCK and IGF1R), using an 80-20 train-test split. Molecular representations were created using Morgan fingerprints (2048 bit, radius of 2) and a random forest model was chosen after experimenting with a variety of classifiers using scikit-learn. The classification model was trained with 500 estimators, whereas the regression model used 100 estimators. This approach allowed for efficient prediction of binding affinity without running docking simulations for each generated molecule, allowing efficient integration with DARWIN.

The peformance of the classification and regression models are given in Table 2 below.

Table 2. Classification and Regression Performance for the Surrogate Model trained on IGF1R and LCK targets trained on the DOCKSTRING dataset.

	CLASSIFICATION		REGRESSION		
TARGET	ACCURACY ↑	AUC↑	RMSE ↓	MAE ↓	R^2
IGF1R	0.918	0.900	0.546	0.429	0.636
LCK	0.922	0.941	0.188	0.374	0.735

B. Toxicity Analysis

Toxicity is one of the main reasons for the failure of drug candidates. In order to ensure that the generated molecules do not exhibit clinical toxicity, the 100 final candidates with high predicted binding affinity to LCK and IGF1R with favorable SA scores were evaluated for clinical toxicity using TxGemma (Wang et al., 2025a) model. TxGemma is a state of the art, specialized model developed for predicting a large number of therapeutic properties of molecules. The 2B version of the model was loaded from Huggingface hub and toxicity was calculated using the 'ClinTox' endpoint, denoting Clinical Toxicity. Out of the 100 molecules each for LCK and IGF1R, only 2 molecules for each target were predicted to be toxic, indicating that the majority of molecules created by DARWIN are likely to be non-toxic. The use of predictive models enables medicinal chemists to identify potential toxicities early in the drug development process, improving efficiency and accelerating the overall drug discovery process.

C. Retrosynthesis Prediction

Synthesis of small molecules is a challenging task, especially when the molecules are generated de novo and do not come from easily synthesizable libraries. IBM RXN is an AI-powered platform that uses neural network models trained on large reaction datasets like patents to provide retrosynthesis pathways for synthesizing molecules. While these models may overlook some practical laboratory conditions such as reaction constraints and reagent availability, they are still used by chemists as a starting point in planning efficient laboratory workflows. To aid chemists in the synthesis of the generated molecules, retrosynthesis prediction was performed on the generated molecules using IBM RXN. The retrosynthesis pathway for a sample molecule targeting IGF1R is shown in Figure 3.

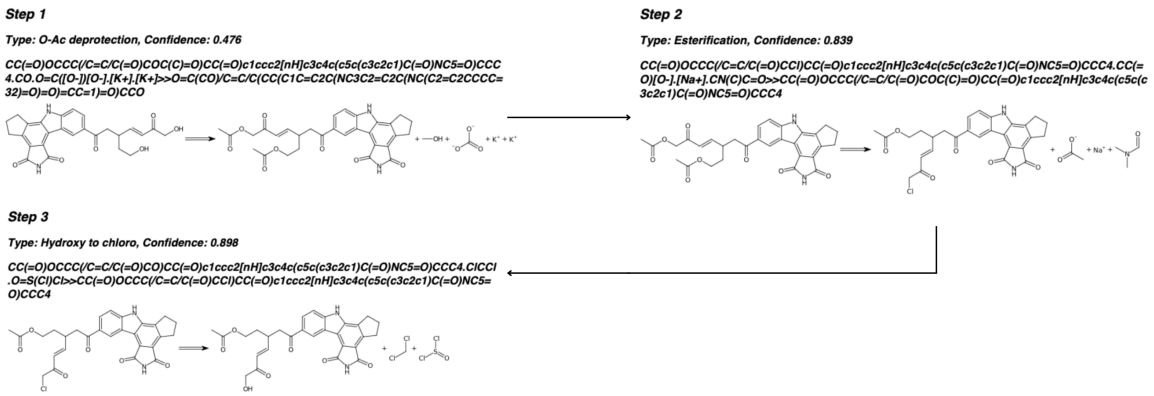

Figure 3. Synthesis Pathway for a sample molecule targeting IGF1R created using IBM RXN.

D. Docking Probability Distribution and Joint Distribution of NP Likeness and Docking Probability

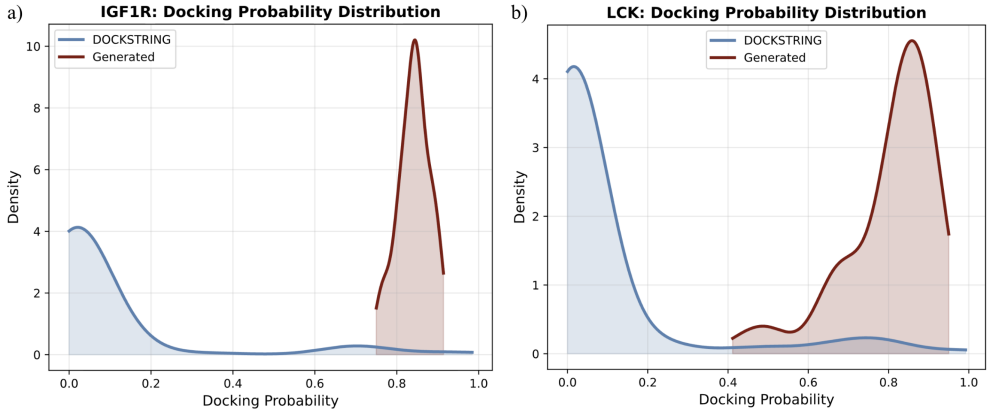

Figure 4. (a) Docking probability for the DOCKSTRING molecules and generated molecules for IGF1R. (b) Docking probability for the DOCKSTRING molecules and generated molecules for LCK.

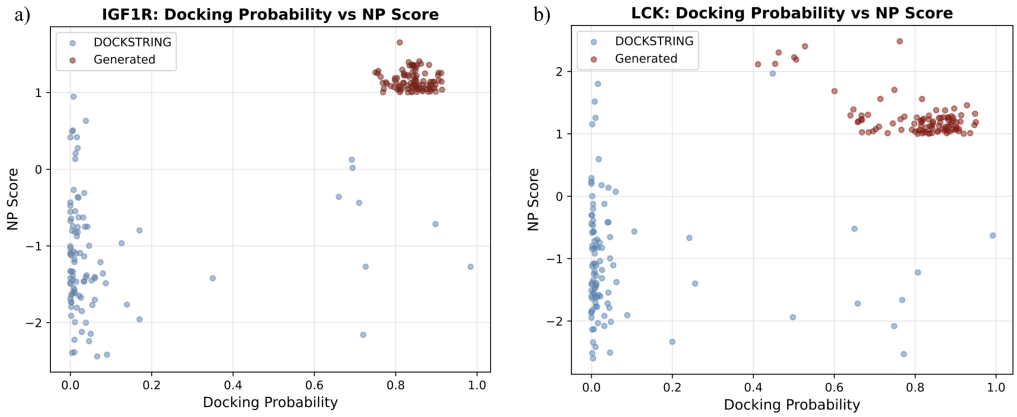

Figure 5. Distribution of NP Likeness and Docking Probability for LCK and IGF1R (a) Natural product-likeness scores and docking probability for the DOCKSTRING dataset and generated molecules for IGF1R. (b) Natural product-likeness scores and docking probability for the DOCKSTRING dataset and generated molecules for LCK.

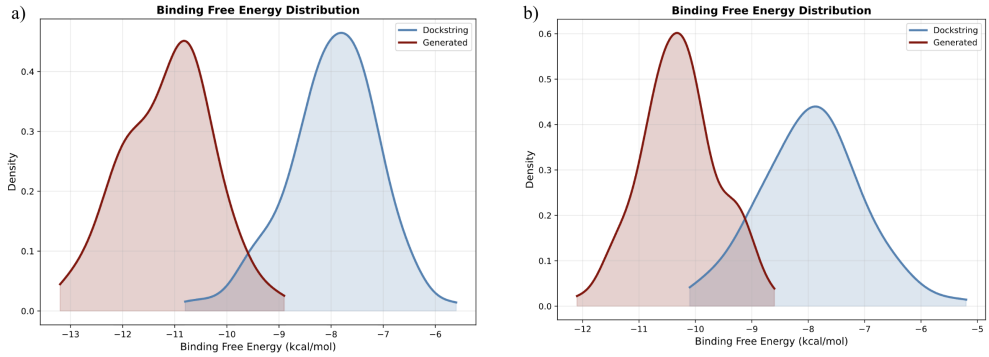

Figure 6. Binding Free Energy distribution for the generated molecules vs molecules from DOCKSTRING dataset for (a) LCK and (b) IGF1R.

E. Docking Poses

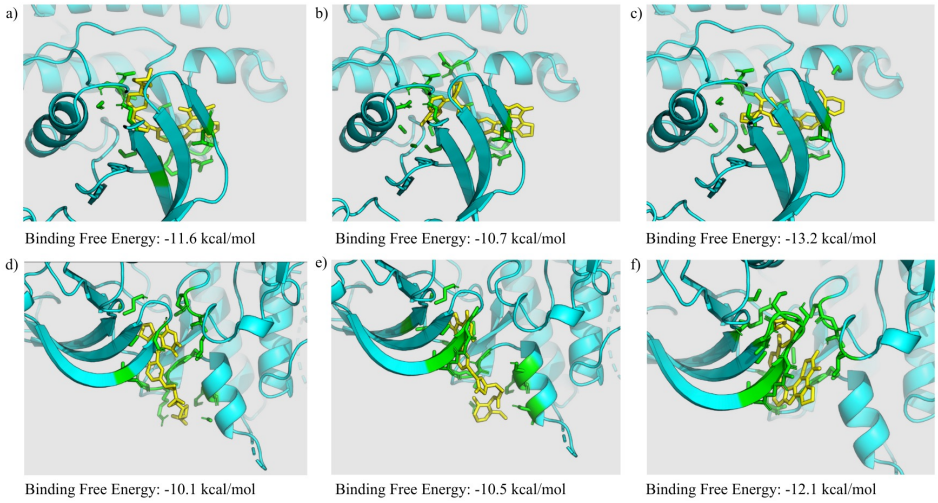

Figure 7. Docking poses for the some generated molecules. The structure and properties of these molecules are in Figure 8. Molecules a-c were for the target LCK and molecules d-f were for the target IGF1R.

F. Sample Generated Molecules

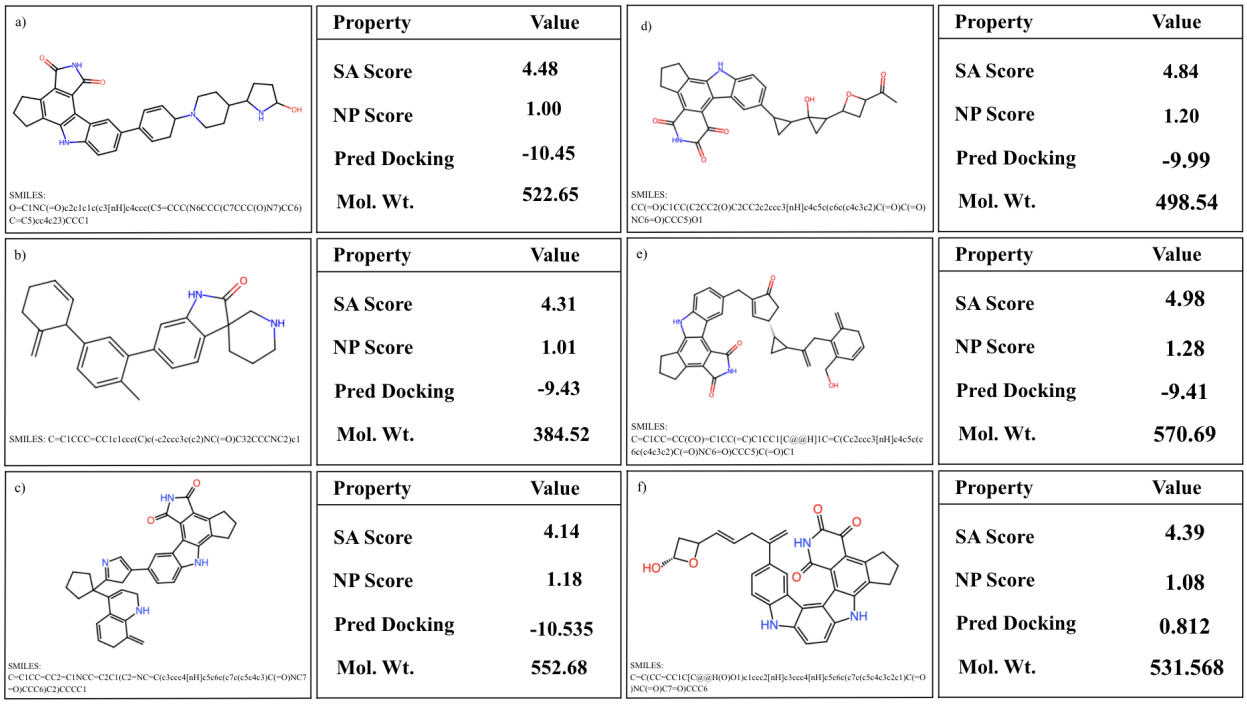

Figure 8. Visualizations of sample molecules generated by DARWIN. Molecules a-c were for the target LCK and molecules d-f were for target IGF1R.

G. Related Work

Natural Products Traditionally, many of our earliest drugs like Penicillin (Ligon, 2004), Streptomycin (Waksman, 1953) and Artemisinin (Ma et al., 2020) were derived from Natural Products. Current studies further highlight natural products' success in drug development, playing key roles in oncology (Atanasov et al., 2015), cardiovascular disease (e.g. statins), and multiple sclerosis (e.g. fingolimod) (Waltenberger et al., 2016). An analyses of FDA-approved new molecular entities indicate that roughly one-third are natural products or their analogues (Patridge et al., 2016), with this figure increasing to nearly half in oncology (Newman & Cragg, 2016). Natural products also show increased rates of clinical trial success throughout the drug development process (Domingo-Fernández et al., 2024). Frequently, natural products further enable the development of orally bioavailable drugs that violate Lipinski's Rule of Five, thereby enhancing molecular diversity (Zhang & Wilkinson, 2007). Cyclosporin, for example, achieves 30% oral bioavailability, demonstrating permeability despite Rule of Five violations (Asano et al., 2023).

Structurally, natural products exhibit remarkable diversity unmatched by synthetic compounds. A study comparing the structural and physicochemical features of natural product based drugs and synthetic drugs found that NP derived drugs displayed greater chemical diversity, occupy larger regions of chemical space, have improved binding selectivity, lower hydrophobicity, higher stereochemical content, and greater molecular complexity (Stratton et al., 2015; Liu et al., 2024; Atanasov et al., 2021; Clemons et al., 2010). Greater complexity and stereochemical content are correlated with decreased preclinical toxicity (Luker et al., 2011; Ritchie & Macdonald, 2009), suggesting that drugs derived from natural products may possess inherently safer pharmacological properties. Notably, natural products often interact with biological targets in unique ways not anticipated by synthetic libraries, revealing previously unknown biological pathways and novel mechanisms of action (Atanasov et al., 2021). Combining natural product fragments in different combinations and arrangements can provide a chemically diverse class of pseudo-natural products, but fusion often requires extensive domain knowledge (Grigalunas et al., 2021). In this paper, we propose to generate NP-like molecules, to exploit the favorable properties of natural products.

Generative Modeling Recently, generative modeling (Gupta et al., 2018; Dai et al., 2018; Gottipati et al., 2020; Samanta

et al., 2019; Li et al., 2018a; Lim et al., 2018; Li et al., 2018b) has emerged as a powerful paradigm for molecular design, enabling the de novo generation of chemically valid molecules. Several approaches have been proposed including Variational Autoencoders (Gómez-Bombarelli et al., 2018; Jin et al., 2018), Reinforcement Learning (Olivecrona et al., 2017; Blaschke et al., 2018; Loeffler et al., 2024; Segler et al., 2017), GFlownets (Kim et al., 2024; Bengio et al., 2021) and diffusion models (Wang et al., 2025b). See (Bilodeau et al., 2022) for a detailed review. However, in all these methods, diversity is not explicitly modeled, and they hope to achieve it by pretraining on a diverse dataset. However, as the number of conditions to be optimized increases, diversity is often compromised as the models tend to generate variations of a small number of scaffolds. In this work, we explicitly model for diversity, allowing for precise control of intermolecular similarity of the generated molecules.