A Foundation Model for Virtual Screening of Ultra-Large Chemical Space Based on 3D Pharmacophores

Souleymane Mahaman Laouali[1,2], Alban Lepailleur[1], Dmitri Kireev[2]

[1]*CERMN1, University of Caen, Esp. de la Paix, 14000, Caen, France.*

[2]*Kireev Lab2, University of Missouri, MO 65211, Columbia, USA*.

The drug discovery process comprises multiple stages, including hit identification, hit-to-lead progression, and lead optimization. It begins with the identification of a disease-related biological target, after which scientists explore large chemical libraries to discover ligands capable of interacting with the target and exhibiting the desired biological activity, along with favorable pharmacokinetic and safety profiles. Despite decades of advancement, the overall clinical success rate remains approximately 10%, with only a small fraction of candidates achieving market approval[1]. Traditionally, hit compounds showing initial activity toward a specific target were obtained from natural products, designed by medicinal chemists, or identified through high-throughput screening [2]. However, these approaches demand substantial human and financial resources, making hit acquisition both inefficient and costly. Moreover, HTS can only evaluate a limited number of molecules, which represents a minuscule portion of the enormous chemical space estimated to exceed 10^{60} possible small molecules thereby constraining its effectiveness. A promising alternative in modern drug discovery is pharmacophore-based virtual screening, a computational process that employs statistical and machine learning algorithms to evaluate large libraries of compounds and identify potential hits efficiently. Recent successes in machine-learning-driven virtual screening[3], [4] have been fueled by the availability of massive chemical and biological datasets, which are crucial for training data-intensive predictive models. Here, we present a deep contrastive learning framework that maps 2D molecular graphs and 3D pharmacophore hypotheses into a shared latent space using dual Transformer encoders. By combining chemical feature types with their spatial relationships, the model learns biologically meaningful representations that support direct similarity-based retrieval through cosine distance. Precomputing molecular embeddings enables millisecond-scale screening of libraries containing millions of compounds, offering a practical approach for ultrafast pharmacophore-driven discovery. Overall, this work provides a scalable framework for 3D pharmacophore retrieval and highlights a general strategy for exploring ultra-large chemical spaces through learned molecular representations.

[1] A. M.-N. R. D. Discovery and undefined 2016, “Parsing clinical success rates,” *go.gale.comA MullardNature Reviews Drug Discovery, 2016•go.gale.com*, Accessed: Nov. 17, 2025. [Online].

[2] R. G. J.-D. D. T. Technologies and undefined 2006, “Hit and lead identification: Integrated technology-based approaches,” *Elsevier*, Accessed: Nov. 17, 2025. [Online].

[3] I. Wallach *et al.*, “AI is a viable alternative to high throughput screening: a 318-target study,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 7526-, Apr. 2024, doi: 10.1038/s41598-024-54655-z.

[4] F. Wong *et al.*, “Discovery of a structural class of antibiotics with explainable deep learning,” *nature.comF Wong, EJ Zheng, JA Valeri, NM Donghia, MN Anahtar, S Omori, A Li, A Cubillos-RuizNature, 2024•nature.com*, vol. 626, 2024.