ITForge: A Generative AI-Driven Drug Discovery Pipeline for Structure-Based De Novo Drug Design and Molecular Ranking with Open-Source Frameworks

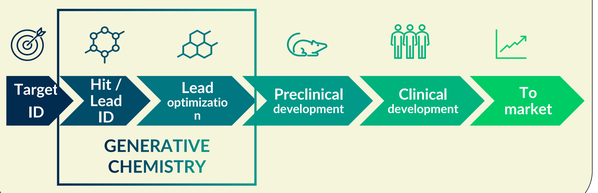
Giorgio Carbone [1], Giovanni Sandrone [2], Ugo Renato Cosentino [1], Claudio Greco [1], Grazia Rovelli [2]

[1] Department of Earth and Environmental Sciences, University of Milano Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milano (MI), Italy, g.carbone8@campus.unimib.it

[2] Italfarmaco S.p.A., New Drug Incubator, Via dei Lavoratori 54, I-20092 Cinisello Balsamo (MI), Italy

Background

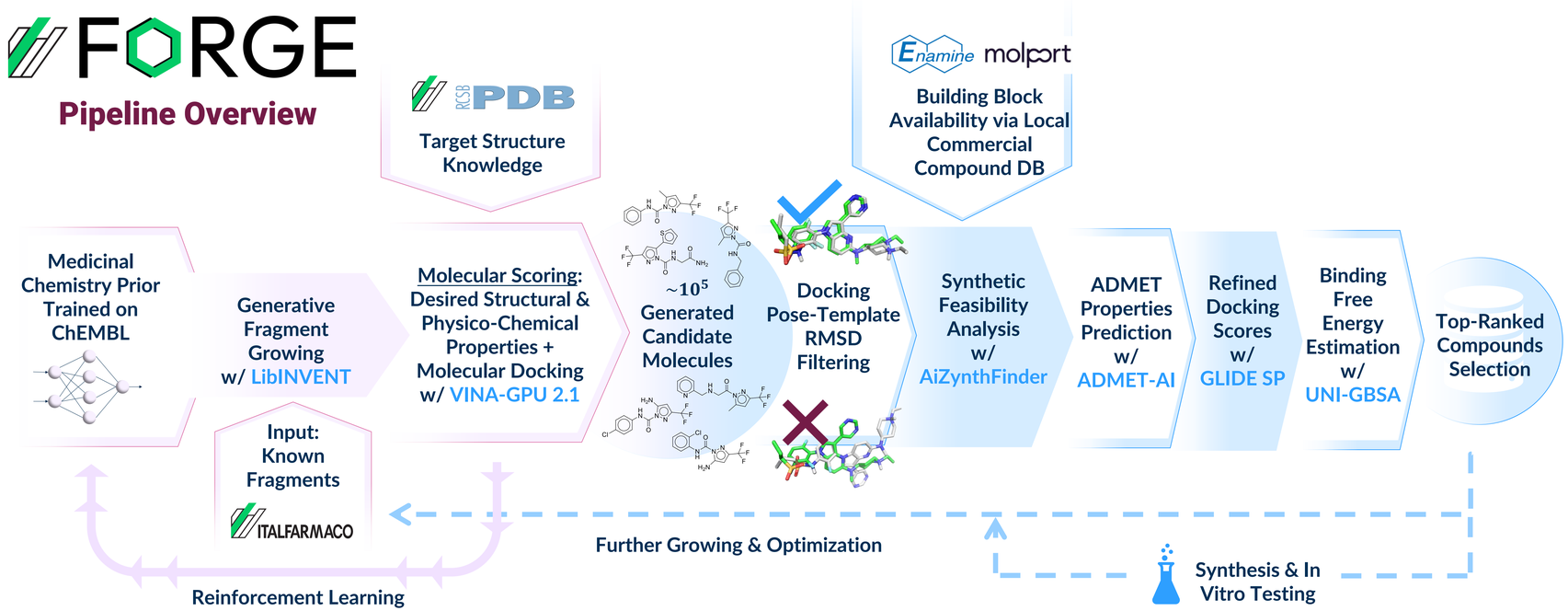
Generative Structure-Based De Novo Drug Design (SBDNDD) is a class of Computer-Aided Drug Design (CADD) methods using generative models to create novel, valid, and synthesizable ligands based on design constraints and the target protein's 3D structure. The aim is producing candidates with optimal ADME-Tox, drug-like physicochemical properties, high target affinity and selectivity, significantly accelerating early preclinical drug discovery (particularly hit/lead identification and optimization)[1].



Aim

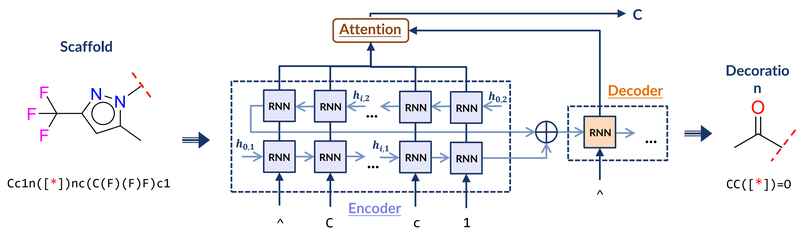
We developed ITForge: an end-to-end, target-aware, AI-driven pipeline for generative fragment growing in hit-to-lead workflows, built by integrating and optimizing open-source frameworks. ITForge addresses key SBDNDD limitations, including mode collapse, complex multi-objective optimization, poor docking scalability, and the lack of an integrated ranking system. It combines the SOTA generative model LibINVENT[2], optimized via Reinforcement Learning (RL), with an advanced post-processing module. The system balances drug-likeness and binding affinity while progressively filtering and ranking candidates using increasingly accurate scoring stages.

Pipeline Overview



Generative AI-Driven Fragment Growing with Reinforcement Learning

LibINVENT[2] is a SMILES-based deep scaffold decoration model pre-trained on ChEMBL in an unsupervised way and based on the architecture by Arús-Pous *et al* [3]:



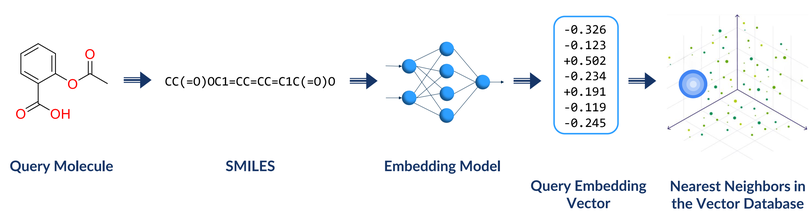
The generative process starts from a promising SMILES fragment with defined growing points and leverages RL to guide molecule generation toward high-scoring regions of chemical space based on a custom multi-objective function:

Scoring Component	Role and Purpose
Vina-GPU 2.1 score	Protein-ligand affinity
SAscore	Synthetic accessibility
Molecular weight – No. H-bond acceptors – No. H-bond donors	Drug-likeness
No. sp^3 carbon atoms – No. rotatable bonds – No. aromatic rings	Molecular flexibility

Compounds Selection: Post-Processing, Filtering, and Ranking

Goal: Refine the ranking of generated compounds by integrating progressively accurate filtering and scoring methods.

1. Docking pose-template RMSD → Filter out unrealistic high-affinity poses.
2. Synthetizability prediction with the deep retrosynthesis tool AiZynthFinder.
3. Fast search for generated molecules and predicted building blocks in an in-house vector database of commercial compounds (~ 11.5M from Enamine and Molport).



4. Prediction of 15+ ADMET properties using the SOTA ML models of ADMET-AI.
5. Refine binding affinities with Schrödinger's GLIDE SP docking protocol.
6. Estimate the binding free energy using the UNI-GBSA method.

References

[1] Pang, Chao, Jianbo Qiao, Xiangxiang Zeng, Quan Zou, and Leyi Wei. 2024. 'Deep Generative Models in De Novo Drug Molecule Generation'. Journal of Chemical Information and Modeling 64 (7): 2174–94.

[2] Fialková, Vendy, Jiaxi Zhao, Kostas Papadopoulos, Ola Engkvist, Esben Jannik Bjerrum, Thierry Kogej, and Atanas Patronov. 2022. 'LibINVENT: Reaction-Based Generative Scaffold Decoration for in Silico Library Design'. Journal of Chemical Information and Modeling 62 (9): 2046–63.

[3] Arús-Pous, Josep, Atanas Patronov, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. 2020. 'SMILES-Based Deep Generative Scaffold Decorator for de-Novo Drug Design'. Journal of Cheminformatics 12 (1): 38.