Generative Deep Learning for Drug Design: Implementation at NovAliX

Ruel Cedeno*, Yaelle Fischer, Dhoha Triki, Philippe Schambel

Department of Computational Chemistry, NovAliX, 67000 Strasbourg, France

ABSTRACT

Deep reinforcement learning (DRL) has emerged as a promising approach for generative molecular design and multiparameter optimization (MPO) in drug discovery. Here, we showcase the integration of deep learning techniques into the drug discovery platform at NovAliX. By combining and fine-tuning the leading opensource packages and proprietary softwares, we built an automated *in silico* workflow which takes our project data and structural constraints as input and optimized molecules as output. After performing extensive benchmark tests, we demonstrate that such integrated DRL-based workflow offers outstanding performance in terms of (1) ability to generate known inhibitors in preclinic and clinical trials (2) druglikeness (3) synthesizability (4) structural diversity, using CCR5 inhibitors as a test case study.

Combining the above results, we integrated our generative deep learning platform into the drug discovery pipeline within NovAliX, in synergy with virtual screening, structure-based and ligand-based drug design. Furthermore, with our internal access to a wide range of biophysics/biochemical assays, augmented synthetic chemistry, DNA-encoded library (DEL) and PROTAC development systems, our unique AI-enabled drug discovery services are tailored to the specific needs of our clients, offering a promising approach for accelerating projects from hit identification to lead optimization, enabling the discovery of new medicines.

CHEMINFORMATICS WORKFLOW AT NOVALIX DRUG DISCOVERY

Virtual Screening

Compound Libraries
- Commercial
- MedChem Designed
- DEL

Pharmacophoric Preselection
Ftrees Similarity Pharmacophoric Similarity Screening — InfiniSee

Docking & Scoring

Filtering
ML-QSAR, physchem, ADME-Tox and custom filters — StarDrop

Clustering
Chemical space analysis to maximize diversity

SeeSAR — **HYDE Score**: Hydration-Desolvation physics-based scoring of docking poses[1]

Prioritized compounds → **Synthetic Chemistry**

Generative Deep Learning

Literature Data
Affinity, Selectivity, ADME[2]

Crystal Structure
Hyde docking score

Constraints
Fragments or scaffolds

Scoring Function
$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$
Multiparametric weights

Generative (AI)
- Physchem, QED, SAScore
- Affinity
- Selectivity
- Early ADME

Optimized compounds (across multiple endpoints)
Synthesis & experimental assays
MedChem evaluation (& IP assessment)

Hit Identification → **Hit to Lead** → **Lead Optimization**

Structure & Ligand-based Drug Design

Ligandibility Analysis
- Binding pocket detection (orthosteric & allosteric site)
- Pharmacophoric modeling via docking of reference compounds
- Homology modeling

Growing & Scaffold Hopping
- 3D fragment library search based on pocket constraints
- Improvement of target efficacy & selectivity
- Hit expansion and IP generation

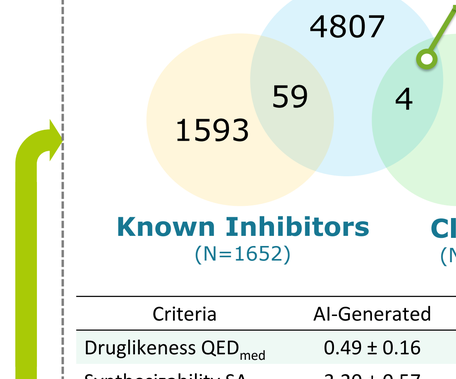
Ligand-based Drug Design
- Pharmacophoric similarity search using known actives
- Useful when target structure is unknown
- Fuzzy pharmacophores for screening billions of compounds

Searching 292,004,942,910,255 molecules

Molecular Dynamics Simulations
- Protein flexibility studies
- Binding affinity prediction & water thermodynamics
- Residence time assessment with enhanced sampling methods

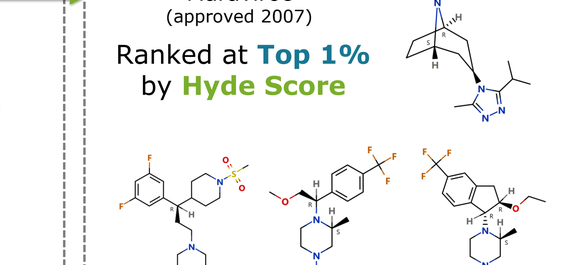
Support for Biophysics, Synthesis, DEL & PROTAC

Linking of fragments[5] from soaking or SPR screen

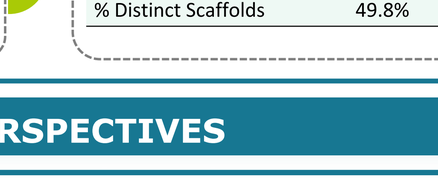


Hit Fragments K_d > 270µM — Linked Fragments K_d = 0.3 µM

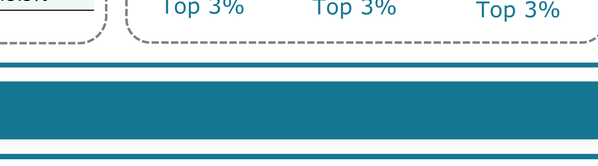
DEL library design and hit finding



Validation of Crystallographic Structures



PROTAC design and data analysis



TEST CASE STUDY USING GENERATIVE DEEP LEARNING

Design of CCR5 Inhibitors and Maraviroc Rediscovery

CCR5 Literature Data
ChEMBL
Known Inhibitors (1652 with 8 in preclinic)

CCR5 Cocrystal
4mbs — SeeSAR
Hyde Docking Score[1]

Structural Constraints
- Scaffolds
- Unwanted substructures

Scoring Function
RDKit, LightGBM, scikit-learn, CDDD, BAYER
ML-QSAR pIC_{50} model

AstraZeneca REINVENT 4
Generative Molecular Design[4]

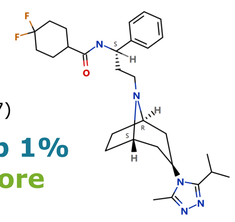
Stage 1	
✓ Physchem	480 < MW < 650
✓ Druglikeness	3.5 < logP < 6.5
✓ Synthesizability	0.4 < Fsp^3 < 0.75
	0.30 < QED < 0.60
	2 < SAScore < 4
Stage 2	
✓ Affinity (Predicted pIC_{50} > 8)	
Stage 3	
✓ Docking Score (Hyde pK_i > 8)	

AI-generated (N=4870)

Known Inhibitors (N=1652) — **Clinic** (N=8)

Criteria	AI-Generated	CHEMBL
Druglikeness QED_{med}	0.49 ± 0.16	0.38 ± 0.14
Synthesizability SA_{med}	3.20 ± 0.57	3.67 ± 0.72
% Distinct Scaffolds	49.8%	43.5%

Pfizer
Maraviroc (approved 2007)
Ranked at **Top 1%** by **Hyde Score**



AZD5672 Top 3% — Vicriviroc Top 3% — INCB009471 Top 3%

CONCLUSIONS & PERSPECTIVES

- ✓ We described the implementation of deep learning for molecular design and multiparameter optimization as part of the cheminformatics tools at NovAliX.
- ✓ By leveraging a unique combination of leading opensource packages and proprietary softwares, we built an *in silico* workflow which demonstrated excellent performance.
- ✓ This AI-integrated workflow can be coupled with diverse research services at NovAliX which are customized to the requirements of our clients.

REFERENCES

1. Schneider, N.; Lange, G.; Hindle, S.; Klein, R.; Rarey, M. A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes: methods behind the HYDE scoring function. Journal of Computer-Aided Molecular Design, 2012, 27, 15–29.
2. Fang, C.; Wang, Y.; Grater, R.; Kapadnis, S.; Black, et al. Prospective Validation of Machine Learning Algorithms for Absorption, Distribution, Metabolism, and Excretion Prediction: An Industrial Perspective. Journal of Chemical Information and Modeling, 2023, 63, 3263–3274.
3. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chemical Science, 2019, 10, 1692–1701.
4. Loeffler, H. H.; He, J.; Tibo, A.; Janet, J. P.; Voronov, A.; Mervin, L. H.; et al. Reinvent 4: Modern AI–driven generative molecule design. Journal of Cheminformatics, 2024, 16.
5. Guo, J.; Knuth, F.; Margreitter, C.; Janet, J. P.; Papadopoulos, K.; Engkvist, O.; et al. Link-INVENT: generative linker design with reinforcement learning. Digital Discovery, 2023, 2, 392–408