# Putting Molecular Generators to Practice: Outlook and Applications of Open Source Software

Martin Šícho (Martin.Sicho@vscht.cz) – RDKit UGM – 2024-09-11

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

# De Novo Drug Design: Yesterday and Today

1700 J. Med. Chem. 1993, 36, 1700–1710

## GroupBuild: A Fragment-Based Method for *De Novo* Drug Design

Sergio H. Rotstein and Mark A. Murcko*

*Vertex Pharmaceuticals Incorporated, 40 Allston Street, Cambridge, Massachusetts 02139-4211*

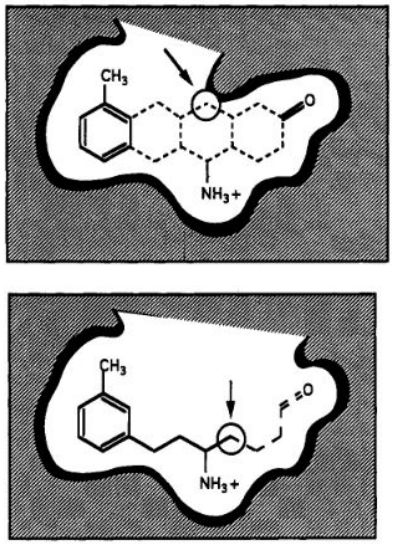
*Received February 9, 1993*

A novel method for *de novo* drug design, GroupBuild, has been developed to suggest chemically reasonable structures which fill the active sites of enzymes. The proposed molecules provide good steric and electrostatic contact with the enzyme and exist in low-energy conformations. These structures are composed entirely of individual functional groups (also known as "building blocks" or "fragments") which the program chooses from a predefined library. User-selected enzyme seed atom(s) may be used to determine the area(s) in which structure generation begins. Alternatively, GroupBuild may begin with a predocked "inhibitor core" from which fragments are grown. For each new fragment generated by the program, several thousand candidates in a variety of locations and orientations are considered. Each of these candidates is scored based on a standard molecular mechanics potential function. The selected fragment and orientation are chosen from among the highest scoring cases. Tests of the method using HIV protease, FK506 binding protein, and human carbonic anhydrase demonstrate that structures similar to known potent inhibitors may be generated with GroupBuild.

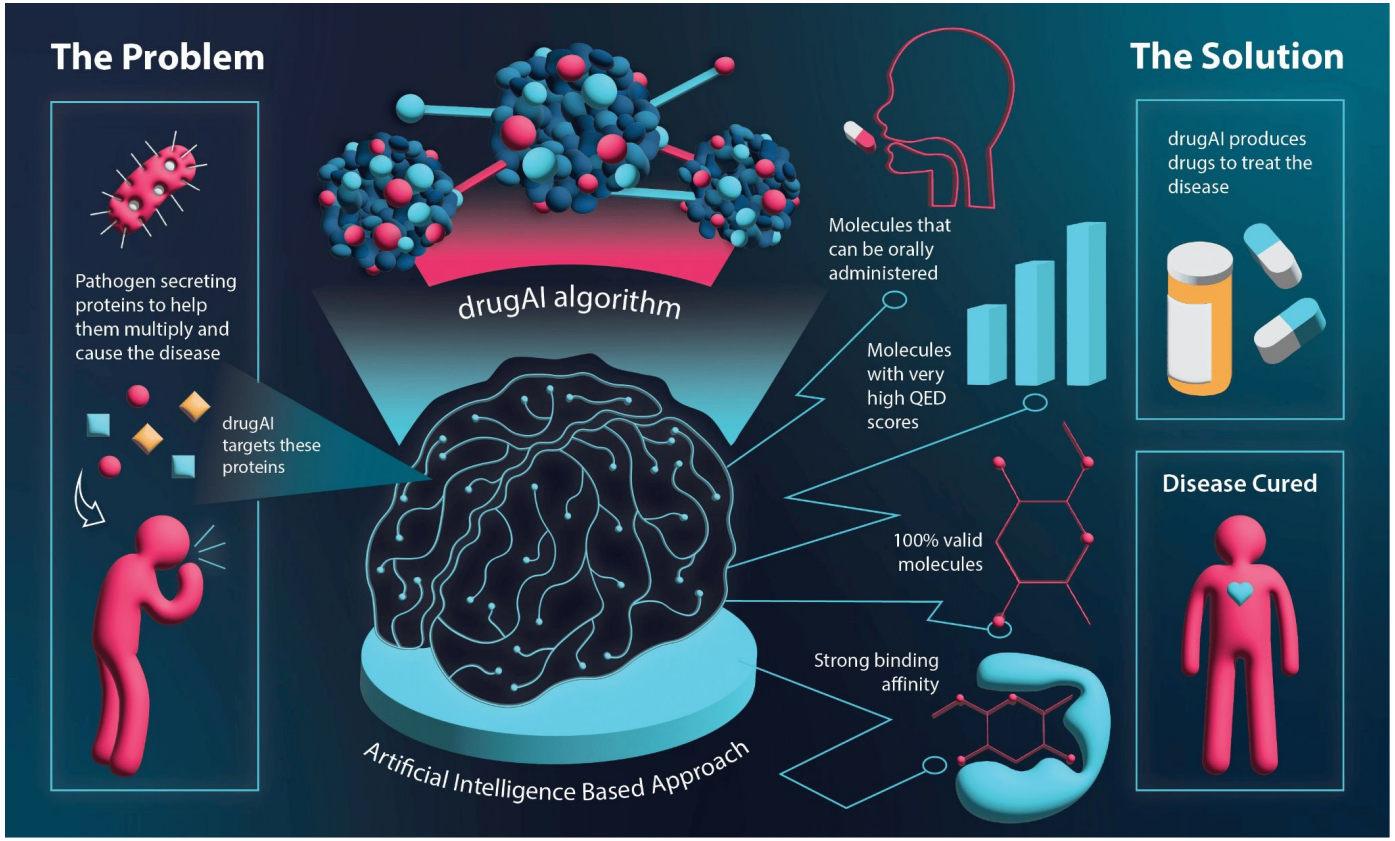
**Table I. Current Fragment Library**

| | | |
|---|---|---|
| Acid | Aldehyde | Amide |
| Amine | Benzene | Cyclohexane |
| Cyclopentane | Ethane | Ethylene |
| Hydroxy | Methoxy | Methane |
| Sulfone | Thiophene | |

combine




*Pharmaceuticals* **2024**, *17*(2), 161; https://doi.org/10.3390/ph17020161

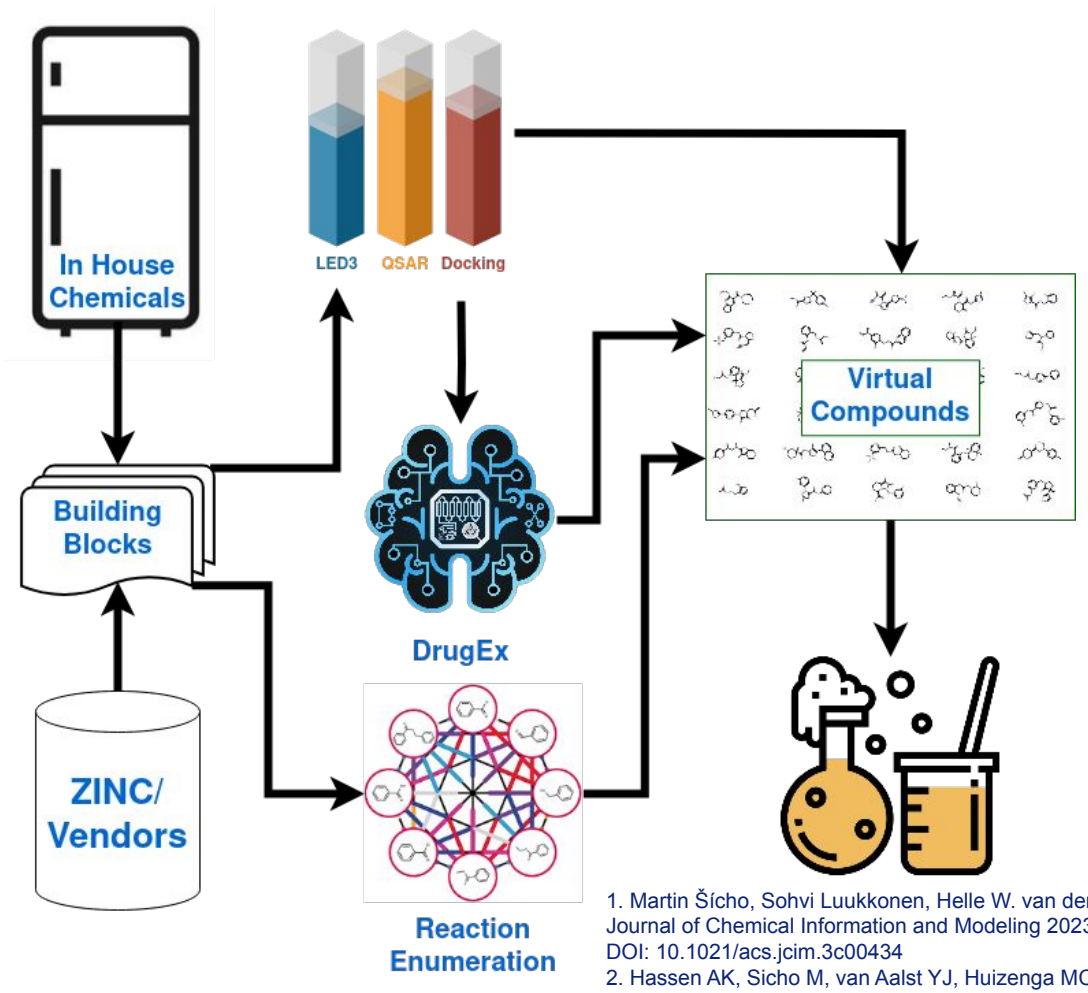


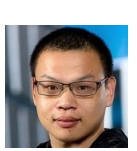

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

# The Mission: Practical De Novo Drug Design



- DrugEx (Molecular Generation)[1]
  - https://github.com/CDDLeiden/DrugEx
- LED3Score (SA Scoring)[2]
  - https://github.com/AlanHassen/led3_score
- QSPRPred (QSPR Modelling)[3]
  - https://github.com/CDDLeiden/QSPRPred
- Spock (Molecular Docking/SBDD)
  - https://github.com/CDDLeiden/spock
  - (available soon)
- GenUI (GUI)
  - https://github.com/martin-sicho/genui



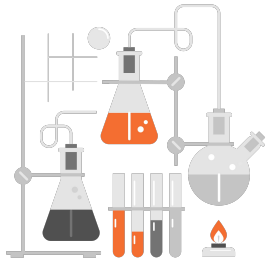



1. Martin Šícho, Sohvi Luukkonen, Helle W. van den Maagdenberg, Linde Schoenmaker, Olivier J. M. Béquignon, and Gerard J. P. van Westen
Journal of Chemical Information and Modeling 2023 63 (12), 3629-3636
DOI: 10.1021/acs.jcim.3c00434
2. Hassen AK, Sicho M, van Aalst YJ, Huizenga MCW, Reynolds DNR, Luukkonen S, et al. Generate What You Can Make: Achieving in-house synthesizability with readily available resources in de novo drug design. ChemRxiv. 2024; doi:10.26434/chemrxiv-2024-wtjt6 This content is a preprint and has not been peer-reviewed.
3. van den Maagdenberg HW, Šícho M, Alencar Araripe D, Luukkonen S, Schoenmaker L, Jespers M, et al. QSPRpred: a Flexible Open-Source Quantitative Structure-Property Relationship Modelling Tool. ChemRxiv. 2024; doi:10.26434/chemrxiv-2024-m9989 This content is a preprint and has not been peer-reviewed.

# De Novo Drug Design Case Study (Monoglyceride Lipase, MGLL)

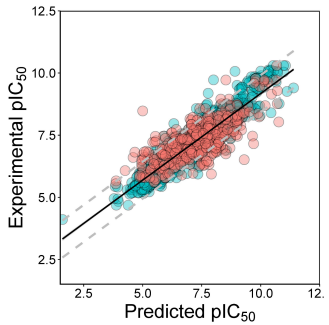
## LED3Score

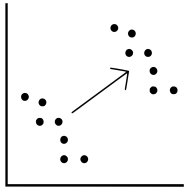
**Given a set of building blocks, can we find a synthetic route to the given compound?**

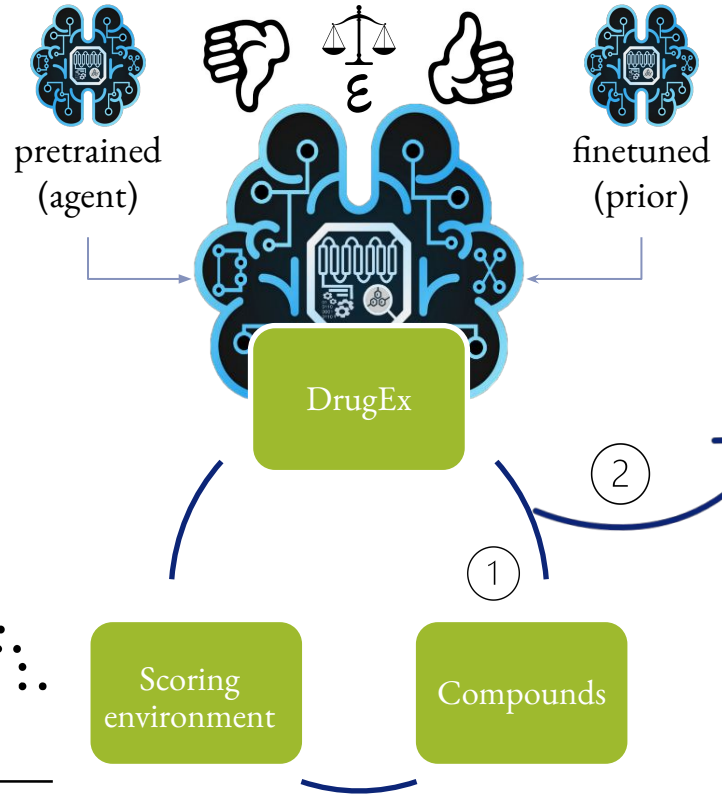


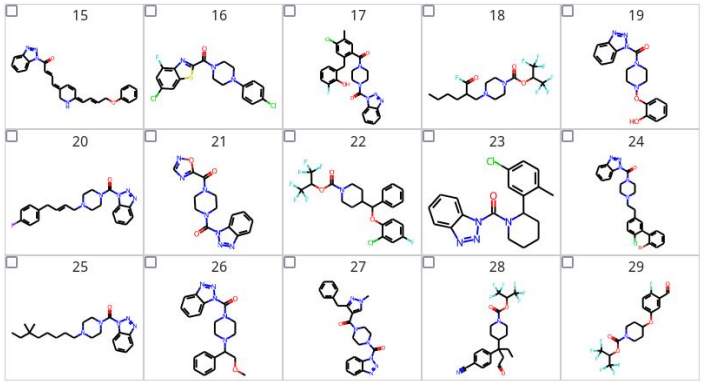
## QSAR Model

**Is the generated compound likely to bind?**

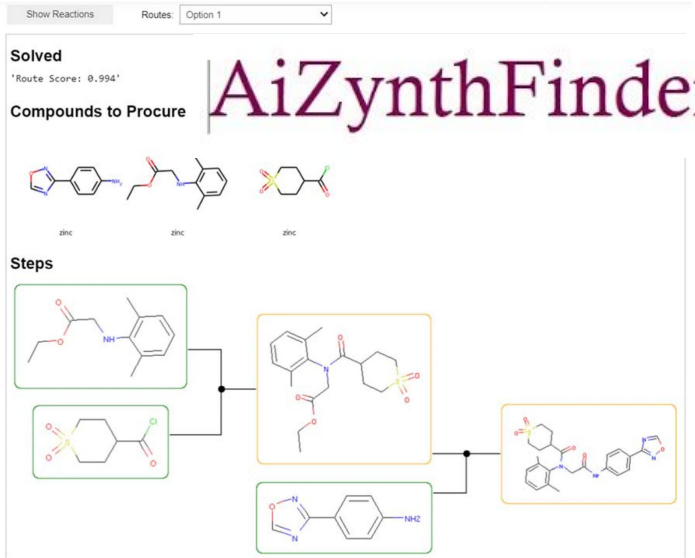





3




## Reinforcement Learning Loop

**DrugEx**: *J. Chem. Inf. Model.* 2023, 63, 12, 3629–3636
**LED3Score**: *ChemRxiv. 2024; doi:10.26434/chemrxiv-2024-wtjt6 This content is a preprint and has not been peer-reviewed.*

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

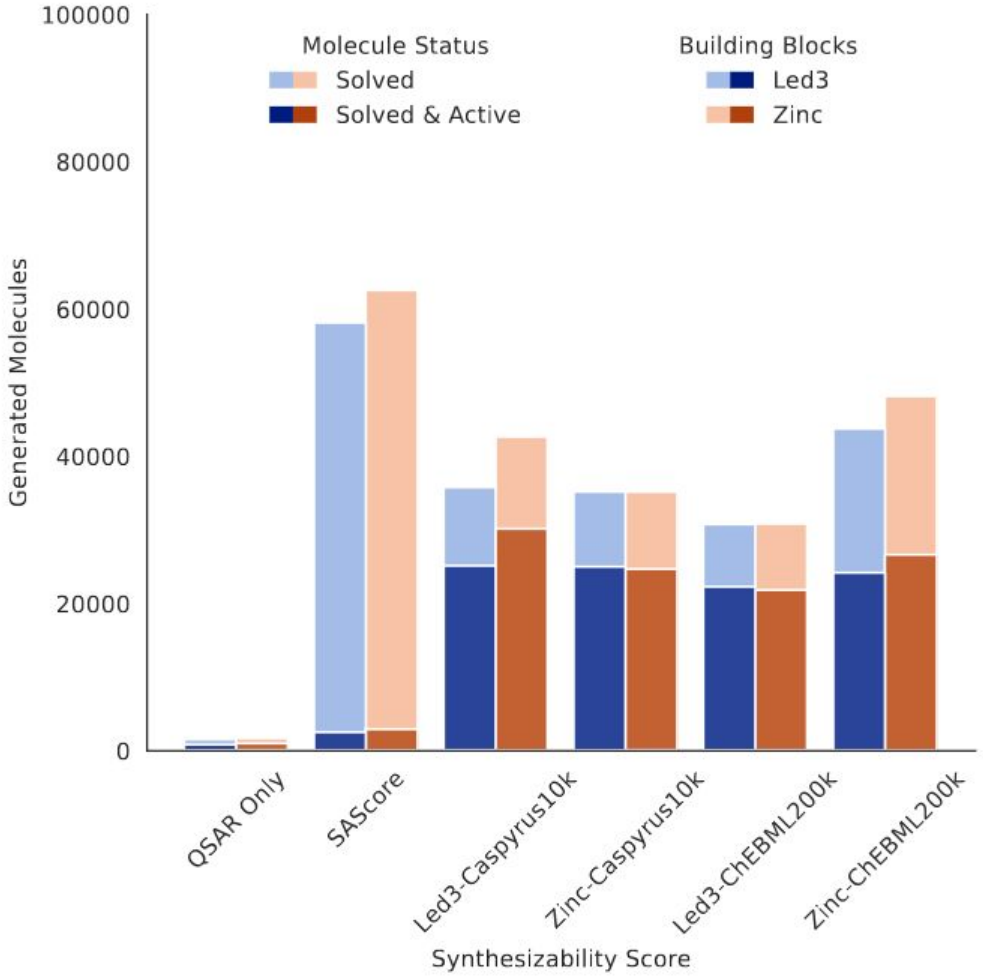
# De Novo Drug Design Case Study (Monoglyceride Lipase, MGLL)

- Workflow[1]:
  a. Train DrugEx with six different SA scores as an objective:
     - **1. None** (baseline), only the QSAR model
     - **2. SAScore** by Ertl et al.[2]
     - **ML-based:**
       - **3. LED3_casp10k**
       - **4. LED3_chembl200k**
       - **5. ZINC_casp10k**
       - **6. ZINC_chembl200k (RAScore reproduction)**
  b. Generate **100,000** structures for each of the 6 cases.
  c. Solve routes for all 6 cases with AiZynthFinder.
  d. Determine desirability of the generated structures.
  e. Answer questions:
     - **Q1: How many desired compounds with solved routes can we obtain?**
     - **Q2: What is the prediction error of the ML-based scores on the generated molecules?**
     - **Q3: Can we pick and synthesize new active molecules?**


1. Hassen AK, Sicho M, van Aalst YJ, Huizenga MCW, Reynolds DNR, Luukkonen S, et al. Generate What You Can Make: Achieving in-house synthesizability with readily available resources in de novo drug design. ChemRxiv. 2024; doi:10.26434/chemrxiv-2024-wtjt6 This content is a preprint and has not been peer-reviewed. https://doi.org/10.26434/chemrxiv-2024-wtjt6

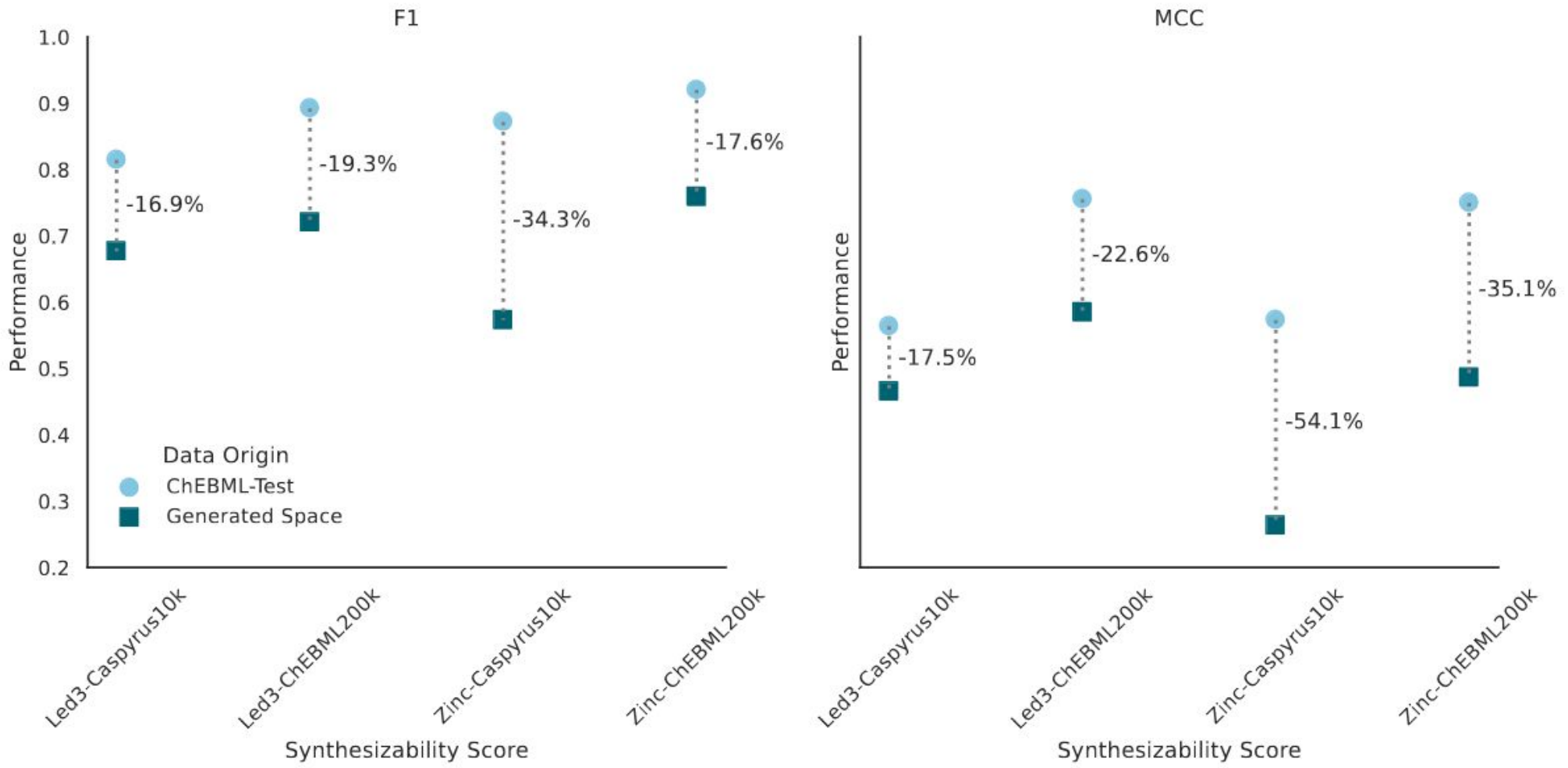
1. Ertl, P., Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1, 8 (2009). https://doi.org/10.1186/1758-2946-1-8

# Q1: Predicted Desirability (Synthesizable & Active)

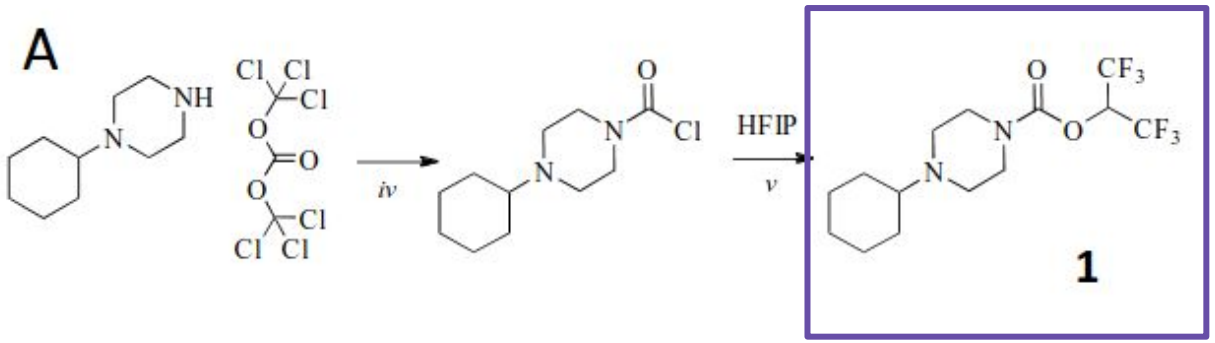


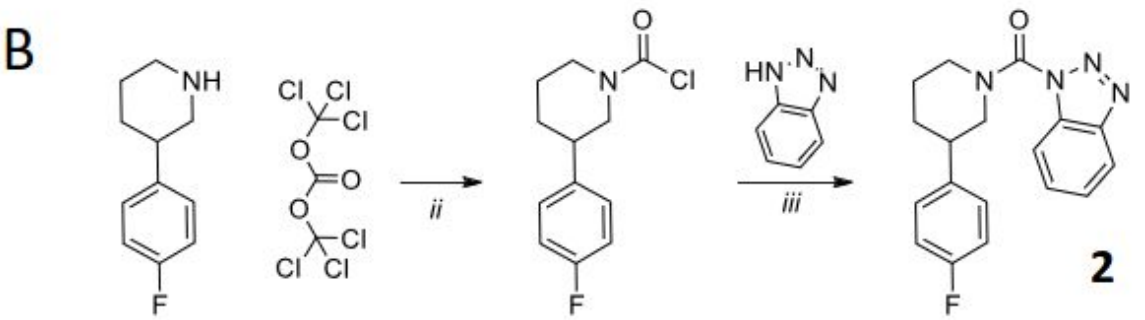

- Synthetic accessibility is important to account for
  - **QSAR baseline without SA nearly zero solved routes**
- **SAScore resulted in poor optimization of the QSAR objective**
- Building block set size does not matter much
  - ZINC and LED3 showed comparable results for all ML-based scores

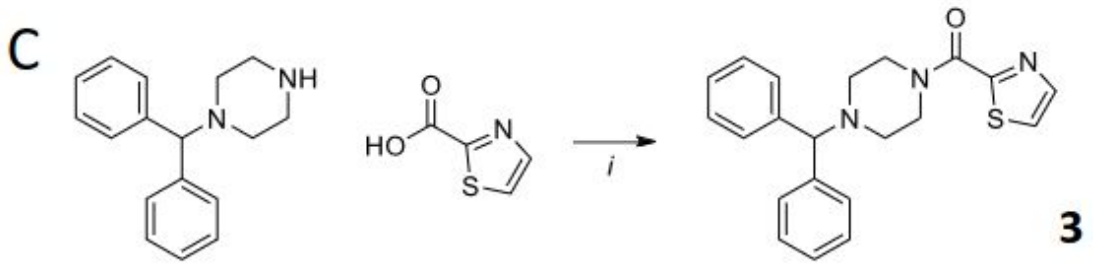
# Q2: Predictive Performance on Generated Compounds

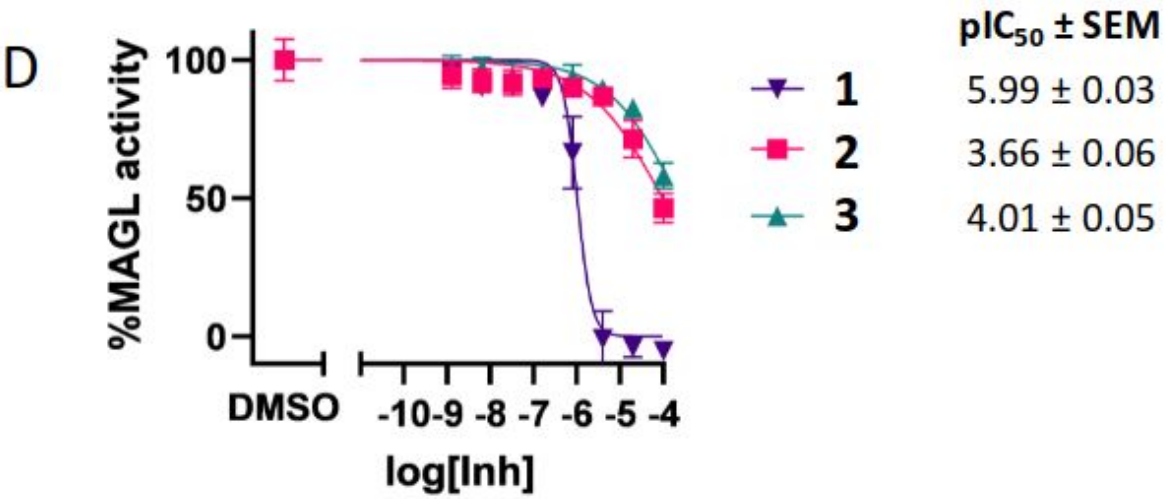


# Q3: Experimental Validation



# Chemokine Receptors (CCRs) in Cancer

| BRAIN TUMOR | |
|---|---|
| CXCR4 | PRX177561 + Bevacizumab + Sunitinib<br>POL5551 + aVEGF<br>AMD3465<br>USL311 + Lomustine*<br>AMD3100* |
| ACKR3 | X7Ab + Temozolomide |

| LUNG CANCER | |
|---|---|
| CCR4 | AF399/420/1802 |
| CXCR4 | AMD3100 + VIC-008<br>LY2510924* |

| BREAST CANCER | |
|---|---|
| CCR1 | CCX9588 + anti-PD-L1 |
| CCL2 | CNTO888 + Radiotherapy |
| CXCR2 | Reparixin + PTX |
| CXCR4 | LY2510924*<br>Balixafortide + Eribulin*<br>USL311 + Lomustine* |

| HEPATOCELLULAR CARCINOMA | |
|---|---|
| CCR2 | 747 + Sorafenib<br>RDC018 |

| PANCREATIC TUMOR | |
|---|---|
| CCR2 | PF-04136309 + nab-PTX + GEM*<br>PF-04136309 + GEM*<br>PF-04136309 + FX*<br>CCX872 + anti-PD-1<br>CCX872 + FX |
| CXCR2 | CXCR2$^{-/-}$<br>CXCR2$^{-/-}$ + anti-PD-1<br>SB225002 + RS504393 + FX<br>AZD5069 *<br>Reparixin + PTX* |
| CXCR4 | AMD3100 + anti-PD-L1 |

| COLON and GASTRIC CANCER | |
|---|---|
| CCR1 | BL5923 |
| CCR4 | AF399/420/1802 |
| CCR5 | Maraviroc + Chemotherapy* |
| CCR7 | siRNA |
| CXCR2 | Reparixin + 5-fluorouracil |
| CXCR4 | LY2510924* |

| RENAL CARCINOMA | |
|---|---|
| CCR4 | Affi 5 |

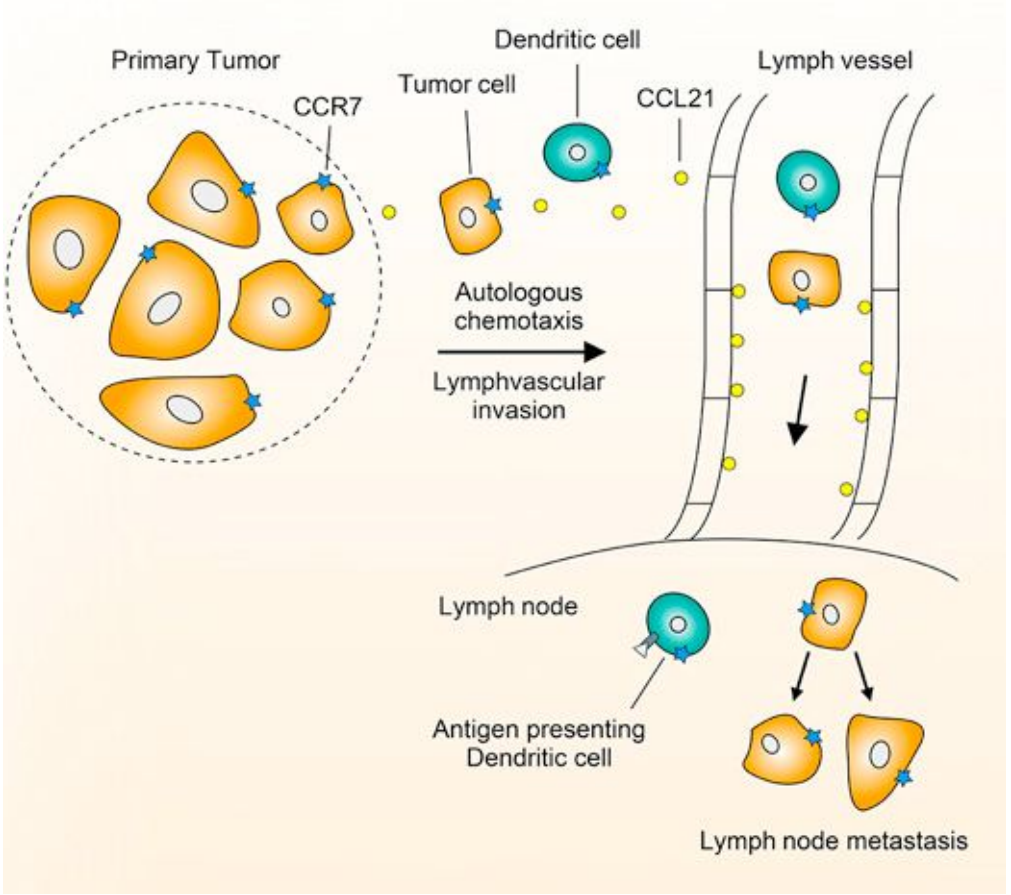
| HEMATOLOGIC MALIGNANCIES | |
|---|---|
| CCR1 | CCX721 |
| CCR4 | Mogamulizumab*<br>Anti-CCR4 CAR T-cells |
| CCR7 | MSM R707 |
| CXCR4 | AMD3100 + Ara-C<br>LY2510924*<br>PF-06747143<br>BKT140 + Rituximab<br>AMD3100<br>AMD3100 + anti-PD-1 |

| OVARIAN and PROSTATE CANCER | |
|---|---|
| CCR2 | iCCR2 |
| CCL2 | CNTO888* |
| CCR7 | siRNA |
| CXCR2 | SB225002 + Sorafenib<br>SB265610 + Docetaxel |
| CXCR4 | LY2510924*<br>AMD3100 |

| SKIN TUMOR | |
|---|---|
| CCR4 | AF399/420/1802 |
| CXCR2 | Navarixin + anti-MEK |

Chemokine receptor inhibitors in cancer. Inhibitors in preclinical models and clinical trials.

From: Mollica Poeta V. et al., Front. Immunol. 10:379. doi: 10.3389/fimmu.2019.00379




CCR7 directs cells to organs that express their ligands (CCL21 and CCL19)

From: Jaeger K. et al., Cell, 5:178, doi: 10.1016/j.cell.2019.07.028

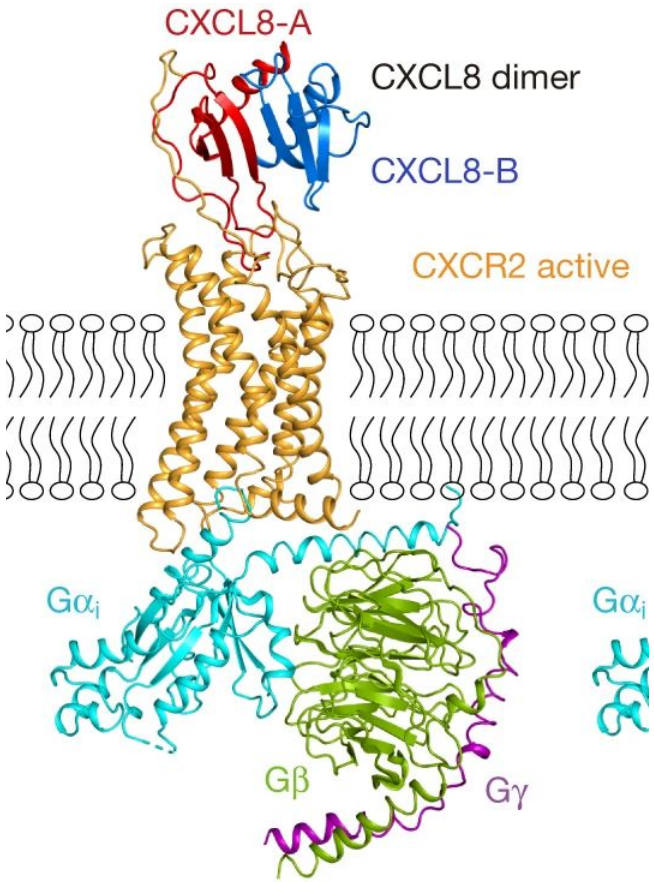
Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

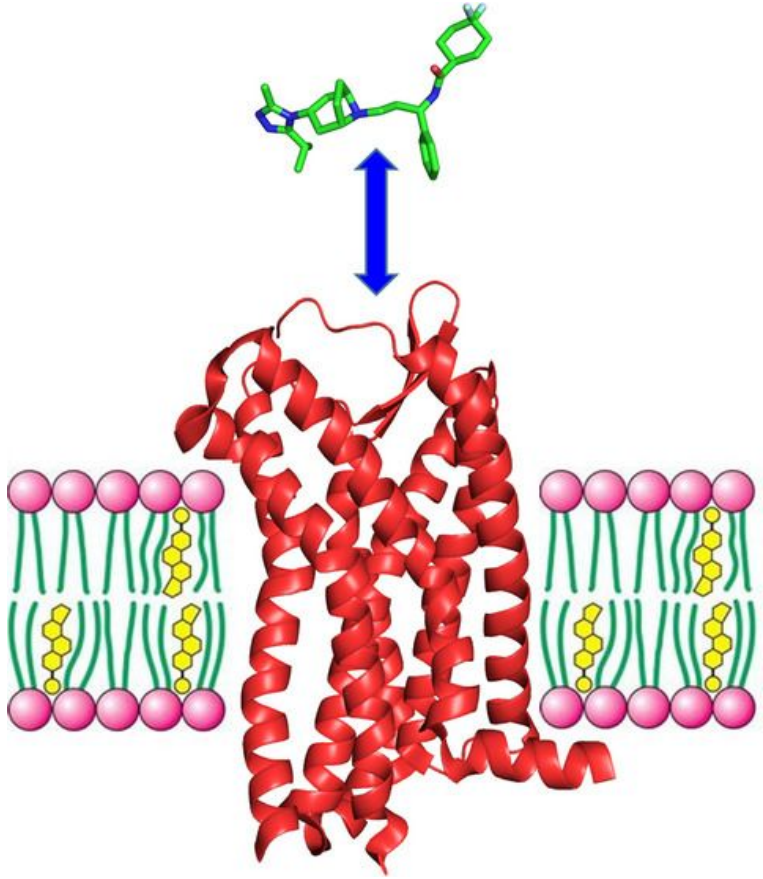
# Activation and Deactivation of CCRs

**Activation**



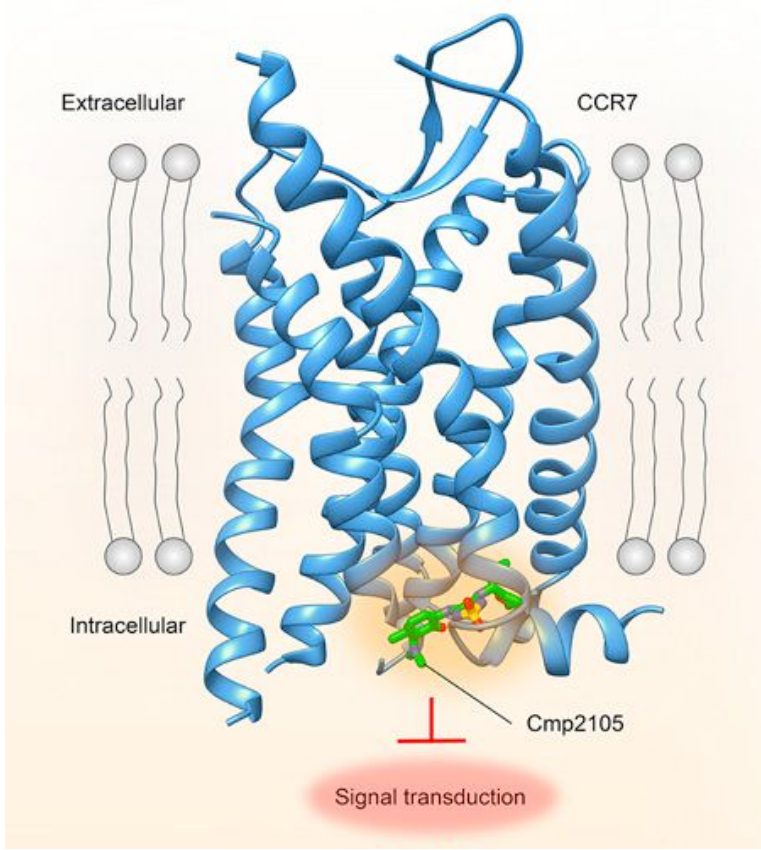

Endogenous activation via chemokine.
From: Liu, K. et al., Nature 585:126-135,
doi: 10.1038/s41586-020-2492-5

**Deactivation - Orthosteric/Extracelular Allosteric**



Maraviroc, extracellular allosteric antagonist of CCR5.
From: Calmet, P. et al., FEBS J, 287:2367-2385.
doi: 10.1111/febs.15145

**Deactivation - Intracellular Allosteric**




Cmp2105, intracellular allosteric antagonist of CCR7.
From: Jaeger K. et al., Cell, 5:178, doi: 10.1016/j.cell.2019.07.028

Universiteit Leiden

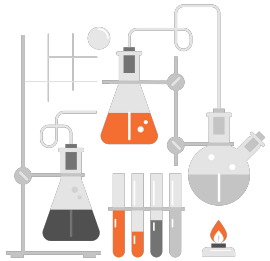
UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

# Molecular Docking in De Novo Drug Design Towards Intracellular Allosteric Ligands of CCR2

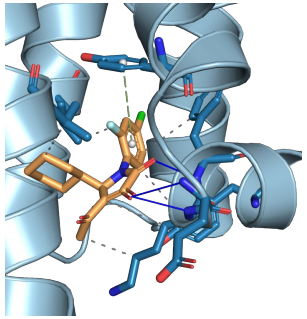
## ZINCScore

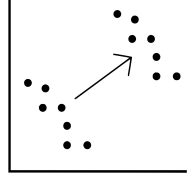
**Given a set of building blocks, can we find a synthetic route to the given compound?**

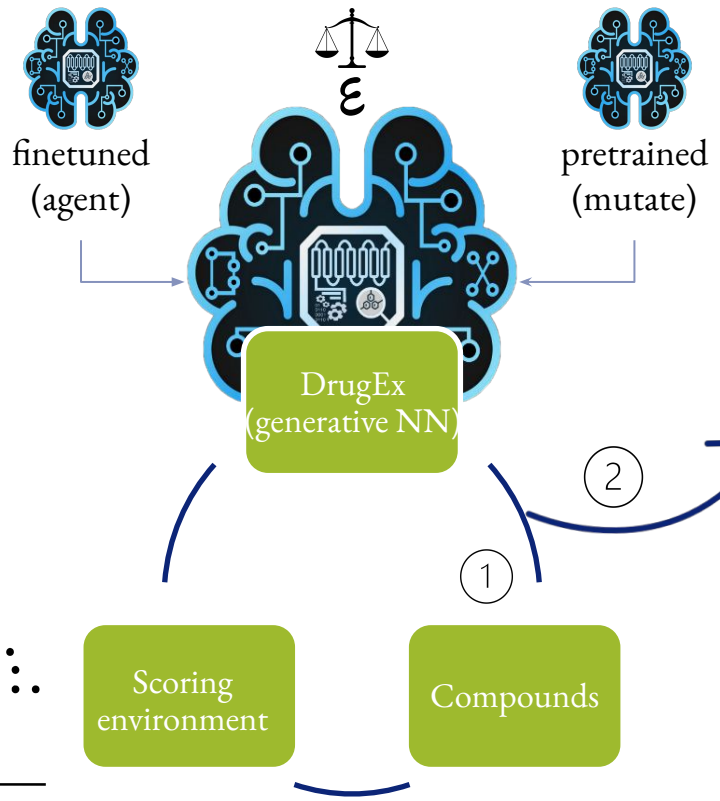


## Molecular Docking

**Score on interactions the ligand can make**

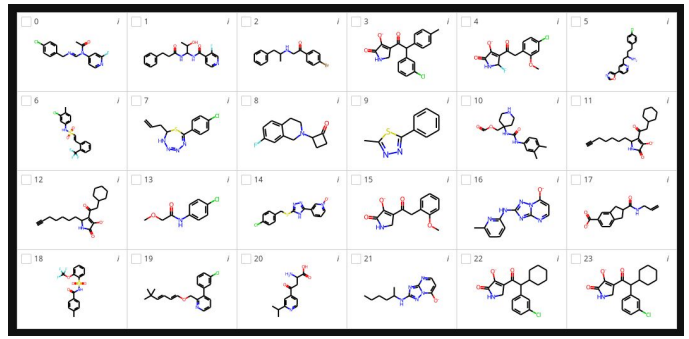




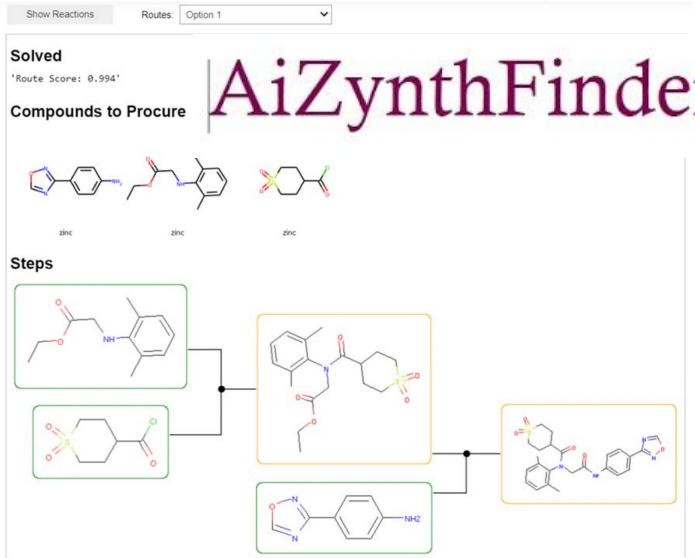



## Reinforcement Learning Loop

**Sicho et al.** *Optimizing Molecular Interactions in De Novo Drug Design: Structure-Based Generation of Intracellular Allosteric Ligands for CCR2 with Transformers, Reinforcement Learning and Docking*, 2025, Unpublished.



3



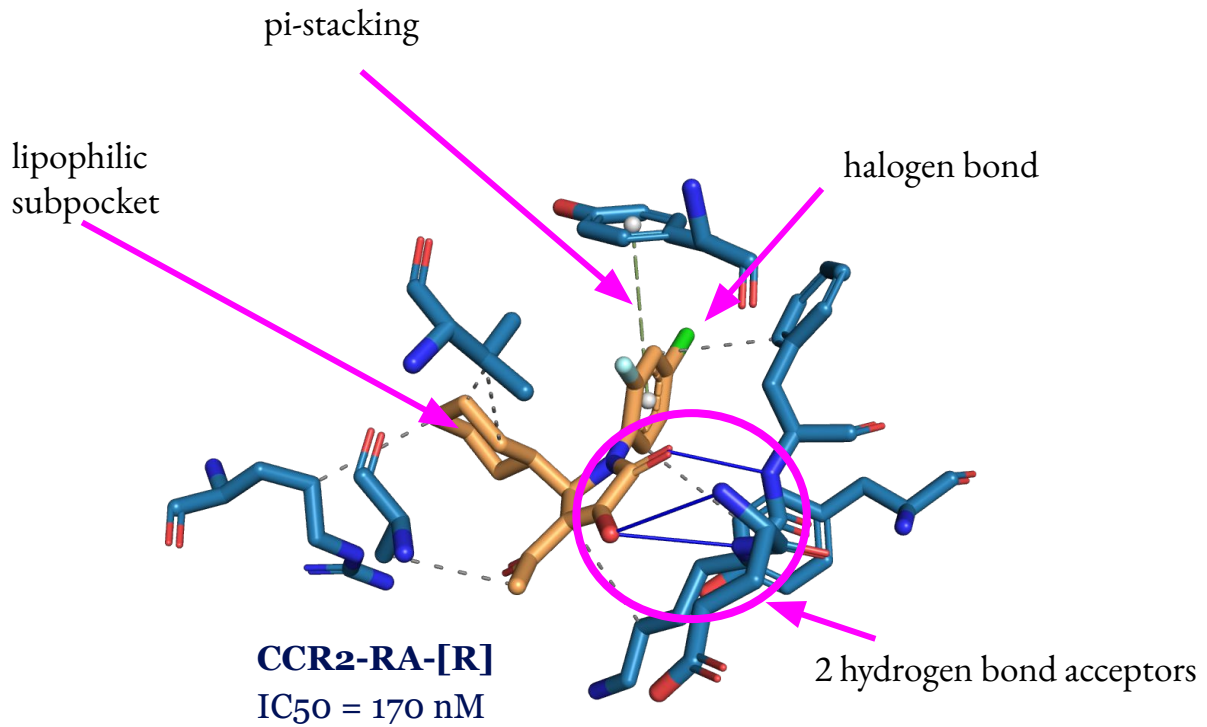

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

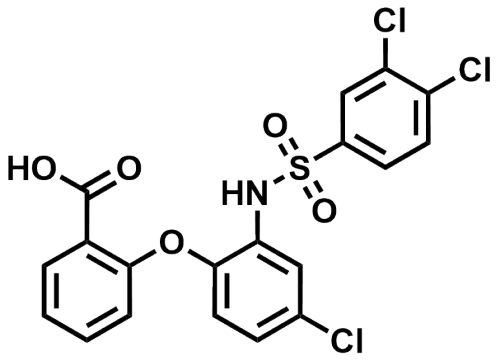
LACDR

# The Binding Pocket

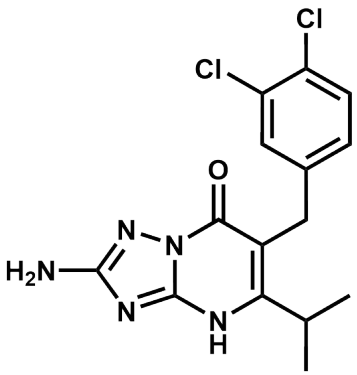
- **One crystal structure with an intracellular allosteric ligand (CCR2-RA-[R]):**
  - https://www.rcsb.org/structure/5T1A
  - downsides:
    - slightly lower resolution (2.81 Å)
    - some residues incomplete
    - mutations of some residues
  - upsides
    - important residues in the binding site are complete and have a meaningful orientation towards the ligand
    - most of the questionable residues are not directly in the binding site
- **Usable for docking after cleanup and some repairs**
  - add incomplete residues
  - reverse mutations close enough to the binding site with a plausible rotamer of the wild type amino acid



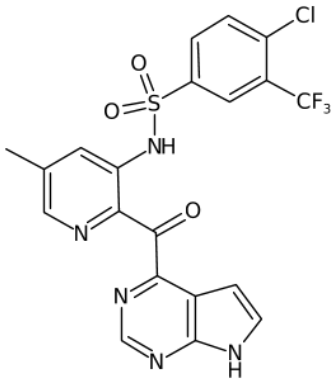
# Known Allosteric CCR2 Ligands



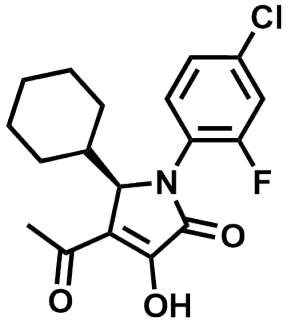
**SD-24**
Ki = 3.2 nM



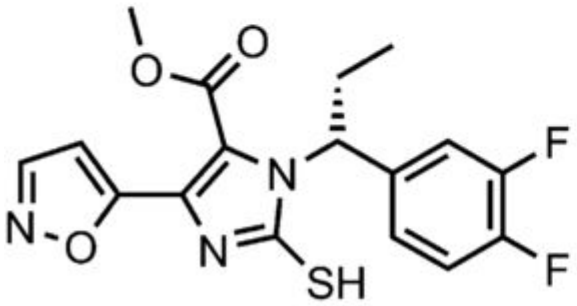
**Compound 39**
Ki = 1.6 nM



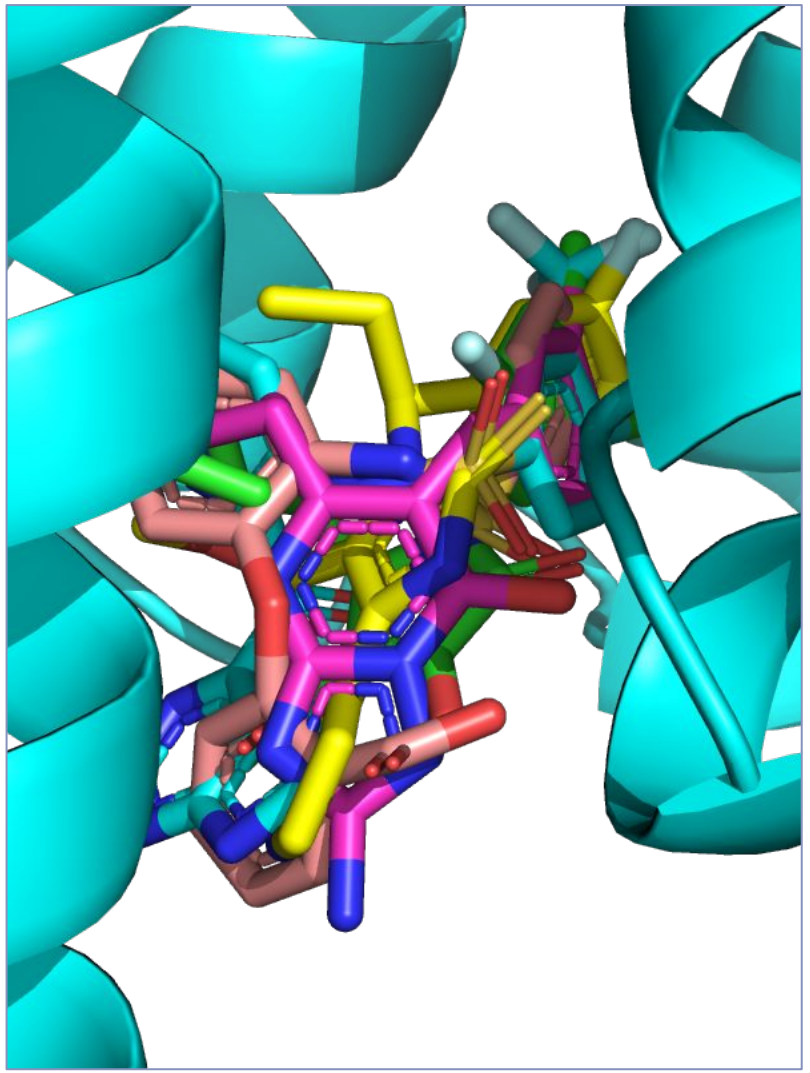
**CCX140**
IC50 = ???



**CCR2-RA-[R]**
IC50 = 12 nM



**JNJ-27141491**
IC50 = 13 nM



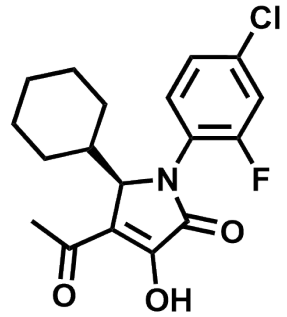
Universiteit Leiden

* data from ChEMBL

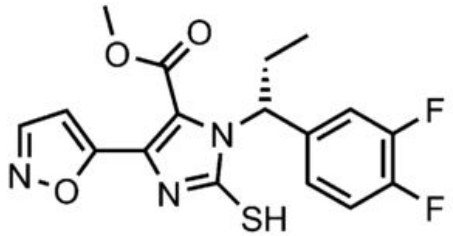
UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

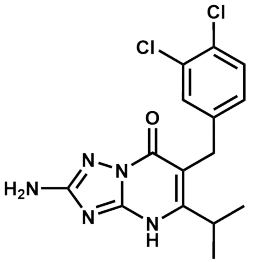
# Docking of Known Ligands (AutoDock Vina)



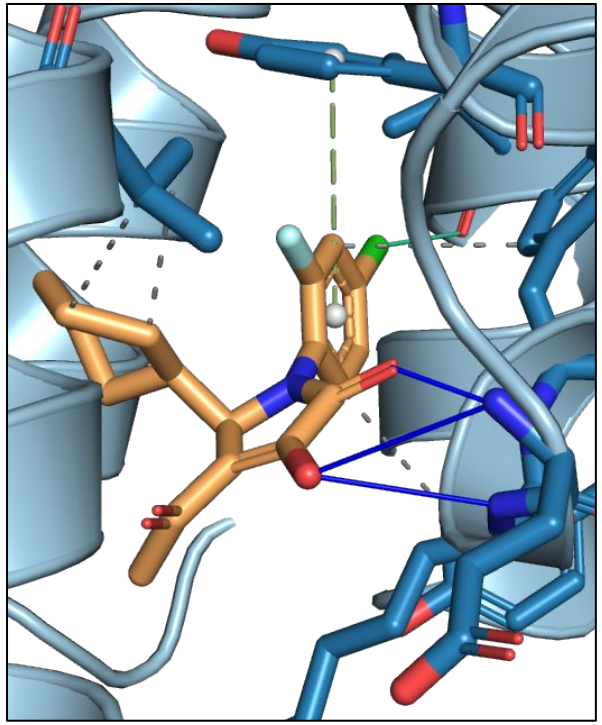

CCR2-RA-[R]

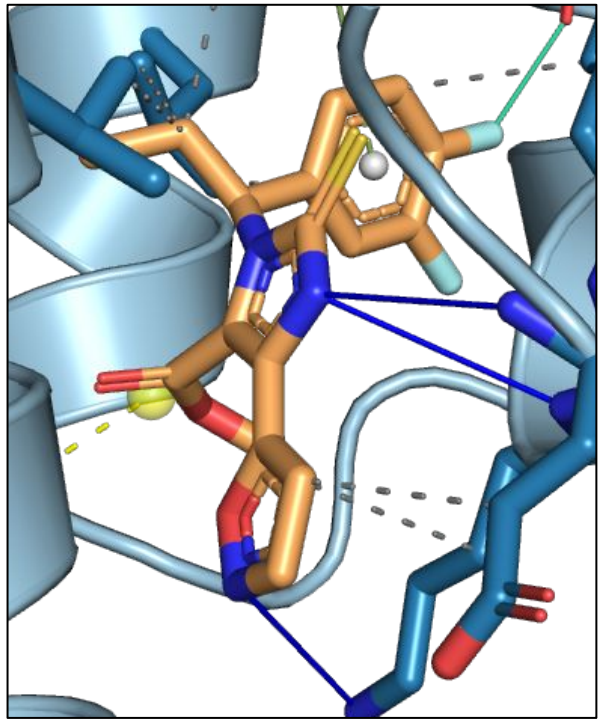


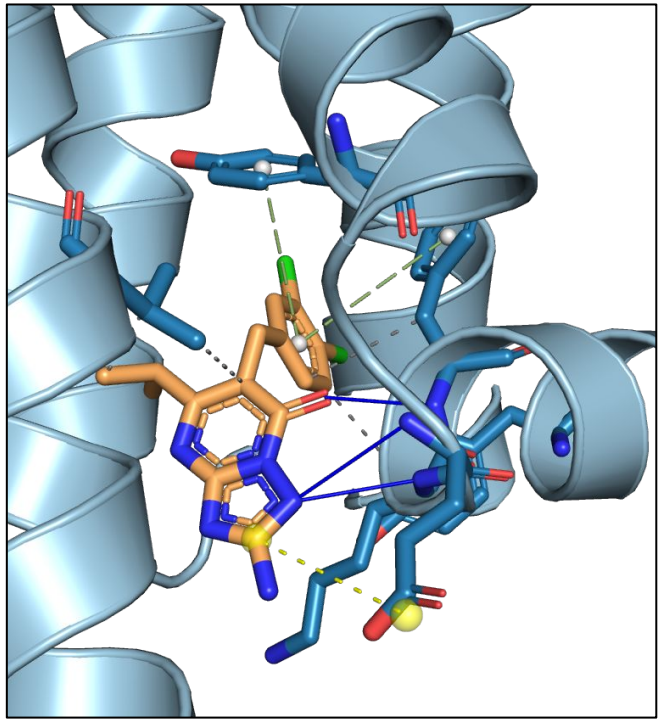

JNJ-27141491




Compound 39

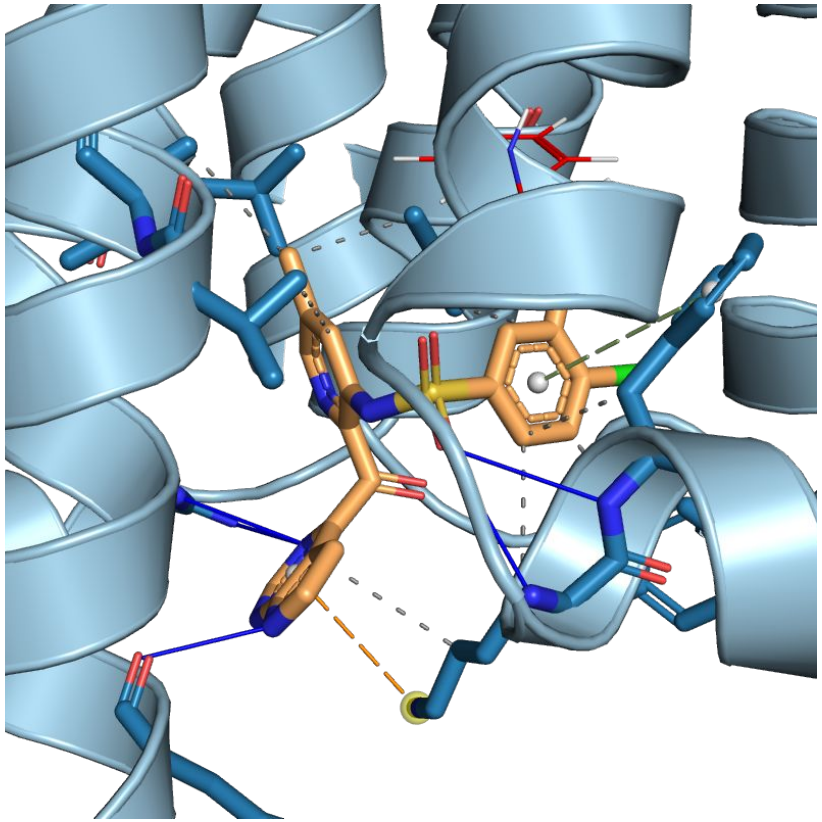


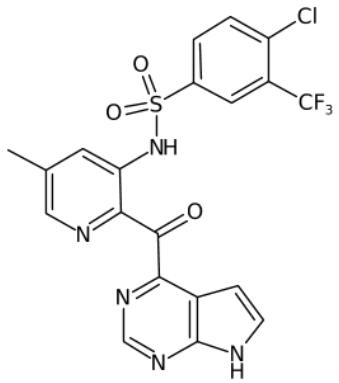




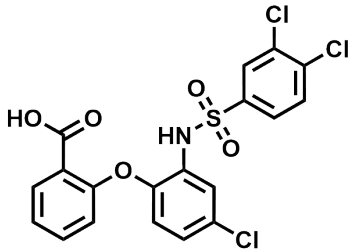
Interactions: https://github.com/pharmai/plip

# Docking of Known Ligands (AutoDock Vina)

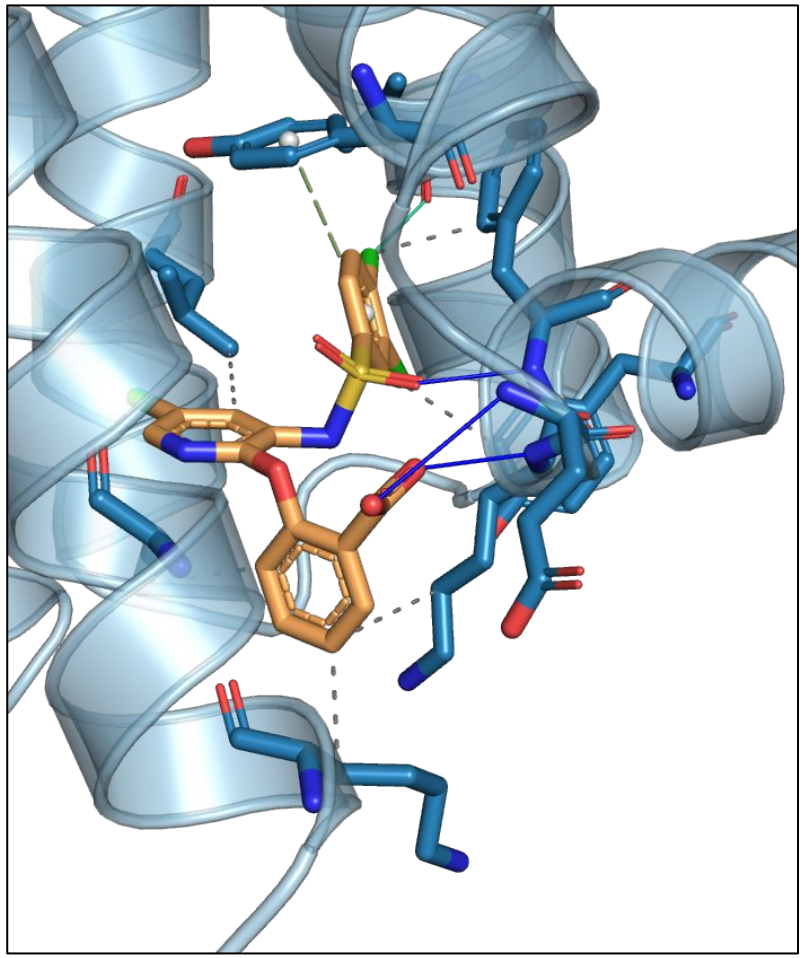





CCX140

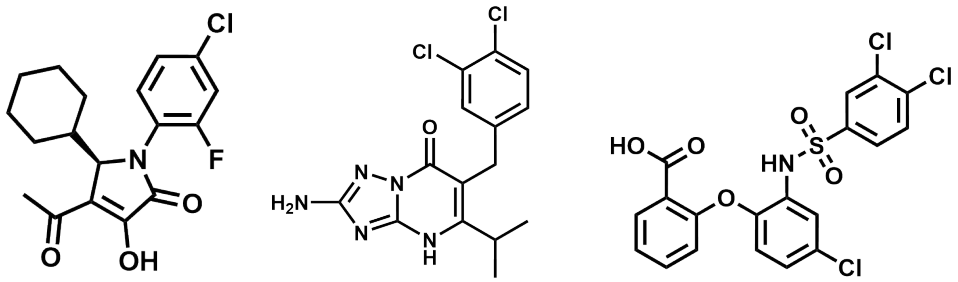



SD-24



Interactions: https://github.com/pharmai/plip

# Scoring the Interactions



- Determined by **5 most active compounds from each scaffold group**
- Interactions manually divided into groups:
  - **Required**
    - Occur in all scaffold groups
  - **Essential**
    - Believed to be important for increased activity
    - pi-stacking interactions with key residues
  - **Important**
    - Known parts of the binding site that many of the high affinity/potency ligands exploit, but each different way
    - especially lipohilic interactions
  - **Interesting/New**
    - interactions that the top ligands have, but not all of them + potentially interesting residues to hit

| | |
|---|---|
| _hbondd_LYS_311_A | 15.0 |
| _hydroph_LEU_81_A | 13.0 |
| _hbondd_PHE_312_A | 13.0 |
| _hbondd_GLU_310_A | 12.0 |
| _halogenbond_VAL_63_A | 12.0 |
| _hydroph_PHE_312_A | 12.0 |
| _hydroph_VAL_244_A | 11.0 |
| _hydroph_TYR_315_A | 11.0 |
| _pistack_TYR_305_A | 11.0 |
| _hydroph_LYS_311_A | 9.0 |
| _hydroph_ALA_241_A | 7.0 |
| _hydroph_LYS_237_A | 5.0 |
| _hydroph_TYR_305_A | 4.0 |
| _pistack_PHE_312_A | 4.0 |
| _hbonda_LYS_237_A | 4.0 |
| _hydroph_LEU_67_A | 4.0 |
| _hydroph_ARG_138_A | 3.0 |
| _hbondd_ARG_138_A | 3.0 |
| _hydroph_VAL_63_A | 2.0 |
| _saltbridge_GLU_310_A | 2.0 |

SCORE = $W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$

all active ligands (pchembl >= 6.5)

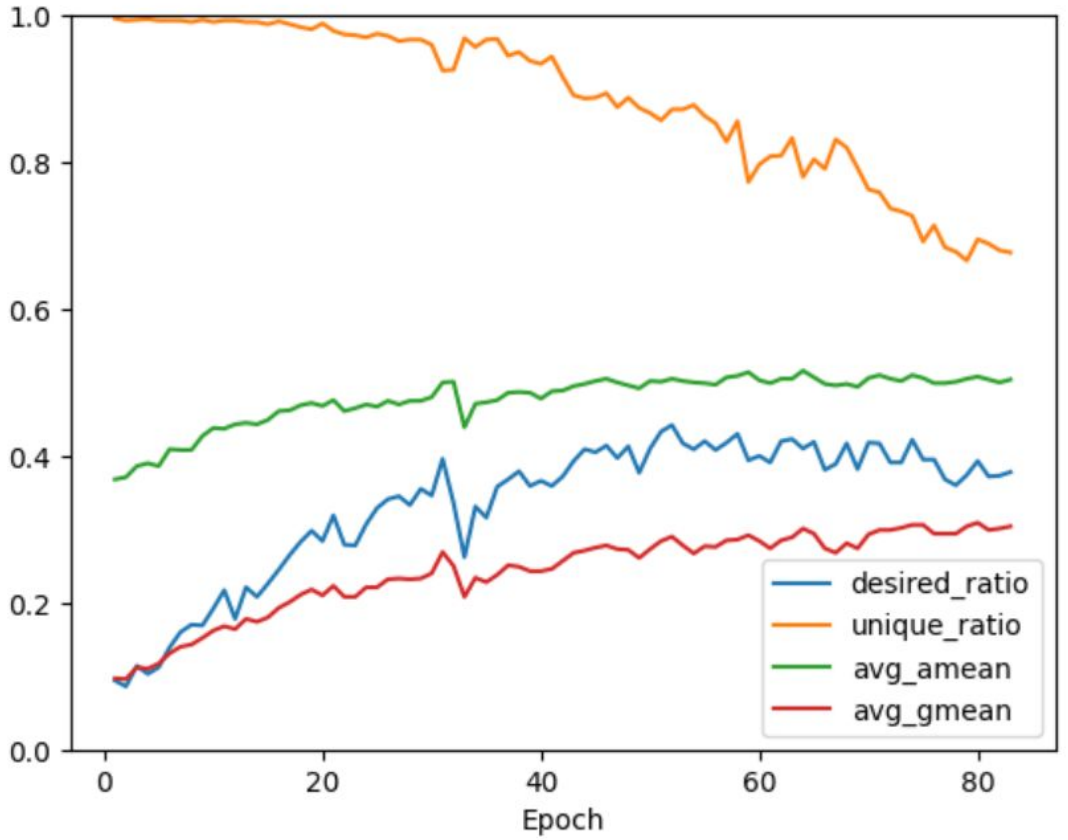
| | |
|---|---|
| _hbondd_LYS_311_A | 179.0 |
| _hydroph_PHE_312_A | 156.0 |
| _hydroph_LEU_81_A | 145.0 |
| _hbondd_PHE_312_A | 142.0 |
| _pistack_TYR_305_A | 142.0 |
| _hydroph_TYR_315_A | 136.0 |
| _hydroph_VAL_244_A | 125.0 |
| _hydroph_LYS_311_A | 123.0 |
| _hydroph_ALA_241_A | 94.0 |
| _hbondd_GLU_310_A | 93.0 |
| _hbondd_ARG_138_A | 83.0 |
| _hydroph_LEU_67_A | 75.0 |
| _pistack_PHE_312_A | 61.0 |
| _halogenbond_VAL_63_A | 59.0 |
| _hydroph_VAL_63_A | 55.0 |
| _hydroph_LYS_237_A | 54.0 |
| _hydroph_THR_77_A | 53.0 |
| _hydroph_TYR_305_A | 48.0 |
| _hydroph_ARG_138_A | 36.0 |
| _hydroph_LEU_134_A | 27.0 |
| _hbonda_LYS_237_A | 26.0 |
| _hbondd_ARG_238_A | 26.0 |
| _pication_LYS_311_A | 23.0 |
| _hydroph_ILE_245_A | 22.0 |
| _pication_ARG_138_A | 20.0 |
| _saltbridge_GLU_310_A | 19.0 |

# Reinforcement Learning

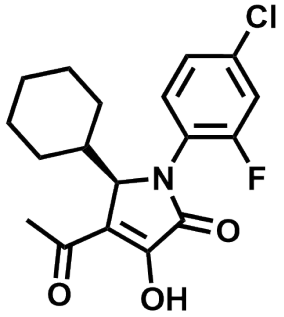
**Generative Workflow**

- generate 10,000 molecules and score them
- keep ligands with:
  - **ZINCScorer > 0.75**
  - Required and Essential IFPScore component = 1 (**hydrogen bonds with conserved residues + pi-stacking**)
  - => **96 structures**

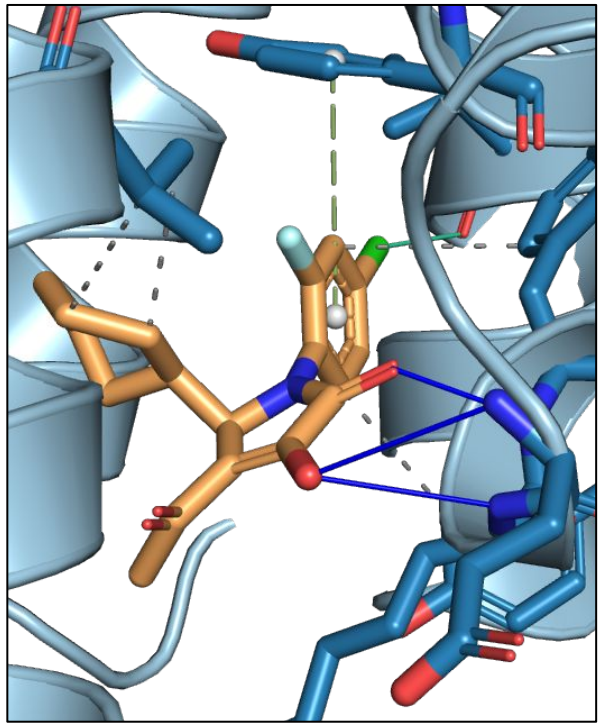
**=> after manual prioritization: synthesis of 4 distinct scaffolds (2 easy, 2 hard)**

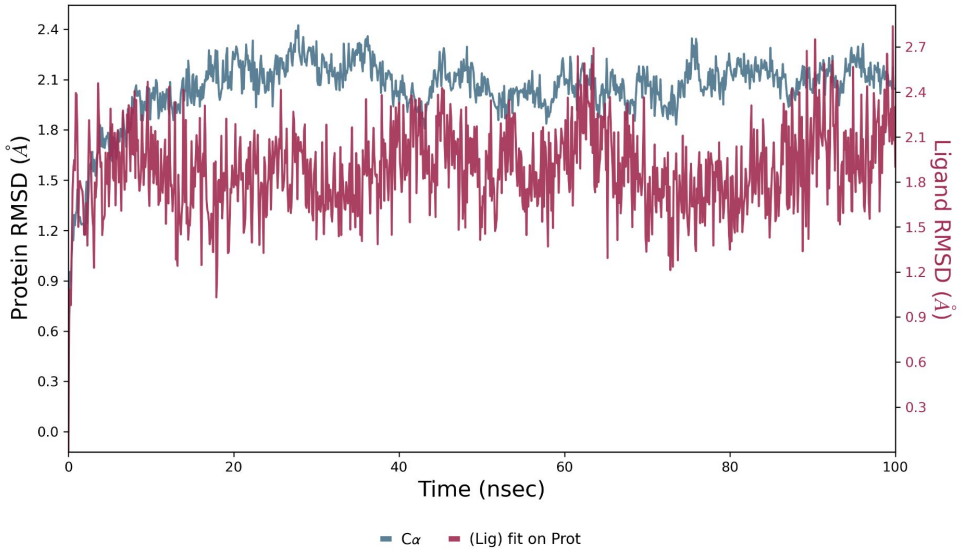


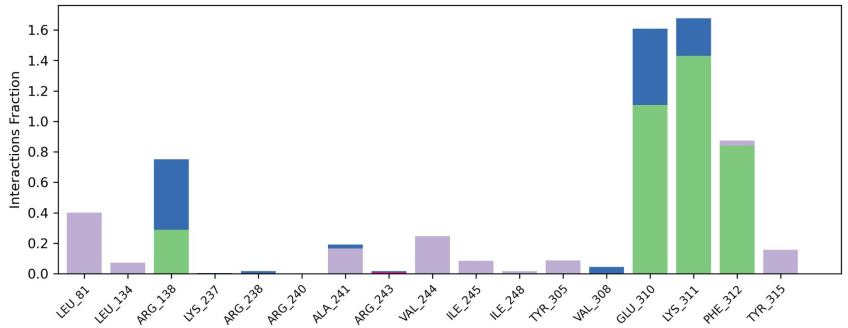
# Molecular Dynamics with CCR2-RA-[R]



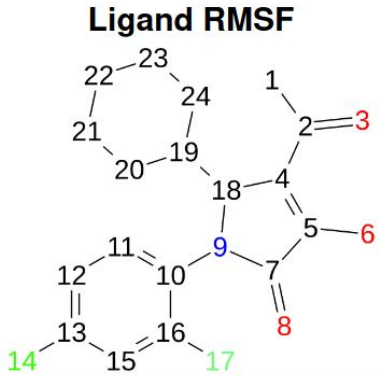

**CCR2-RA-[R]**

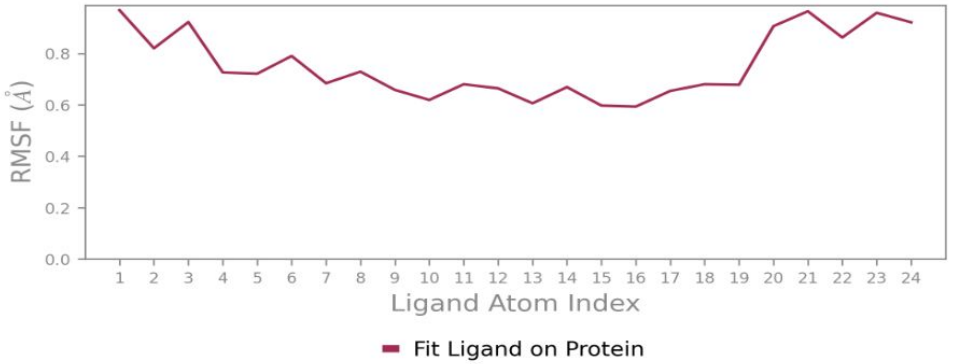




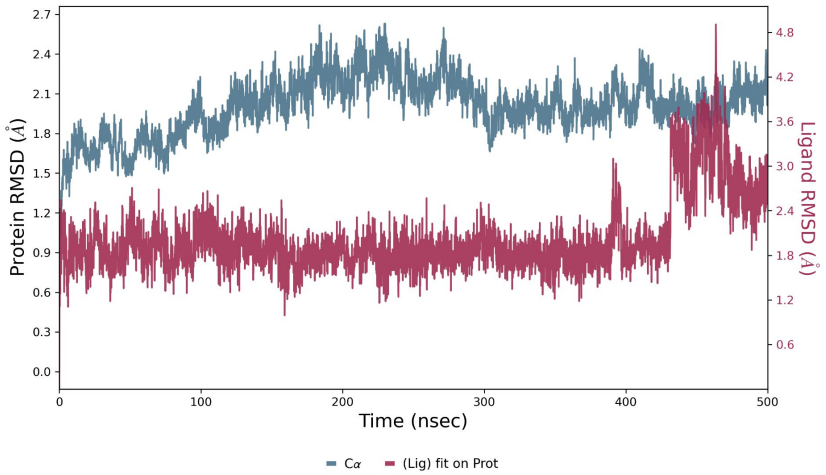



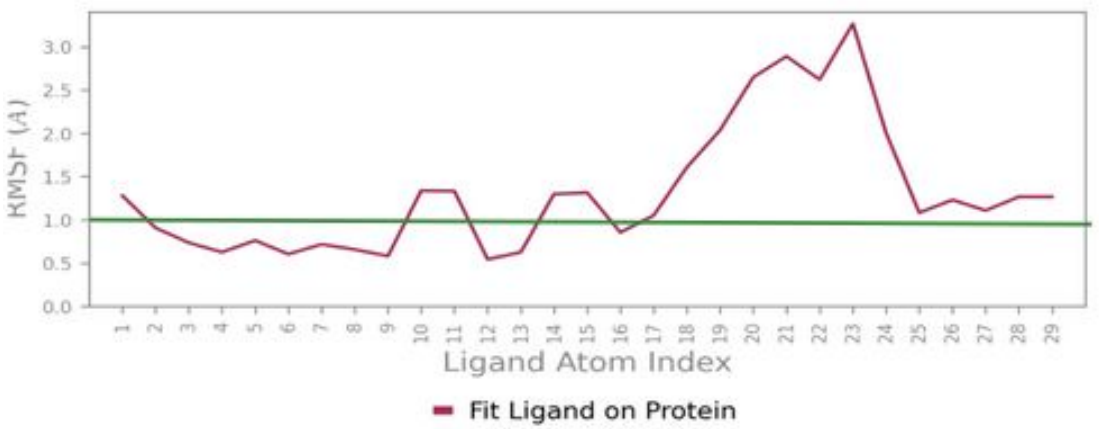

H-bonds Hydrophobic Ionic Water bridges

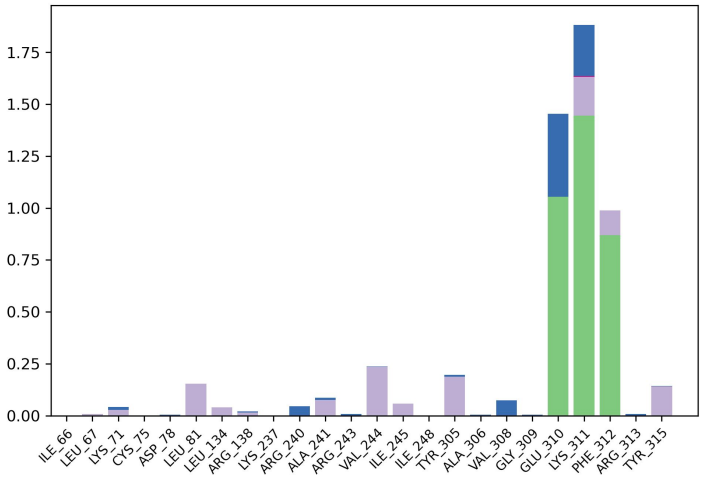


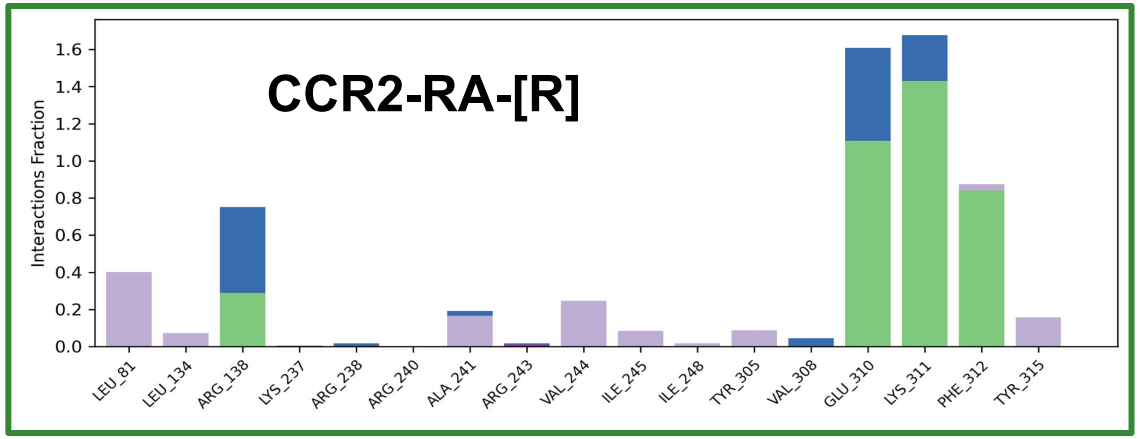



# Candidate v4_007_a

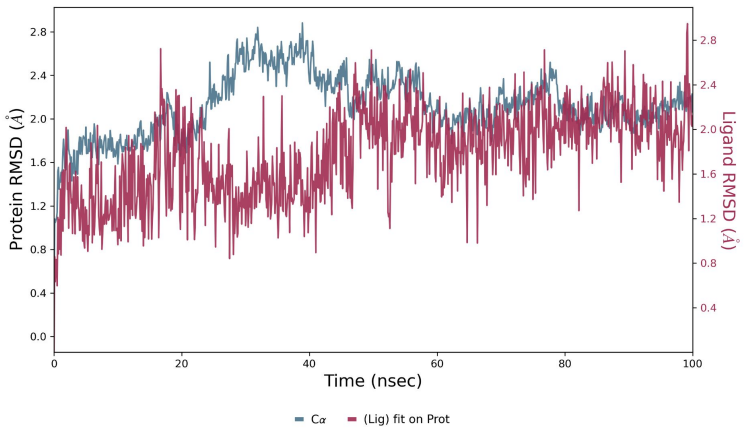


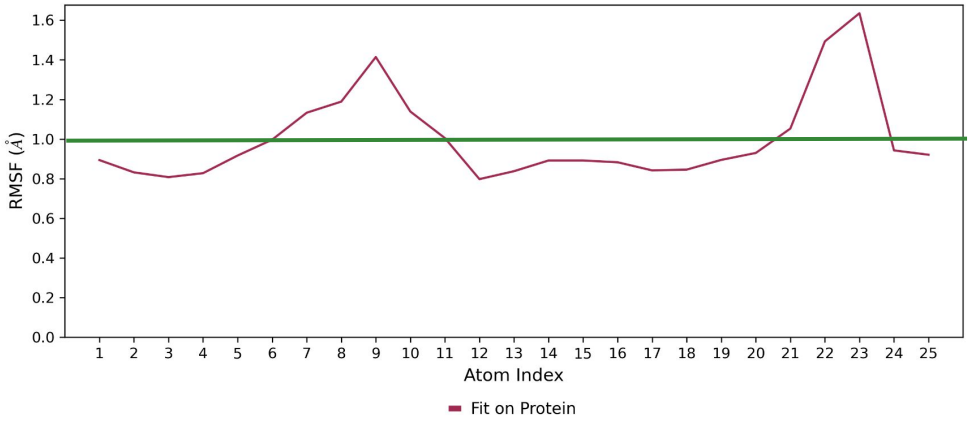



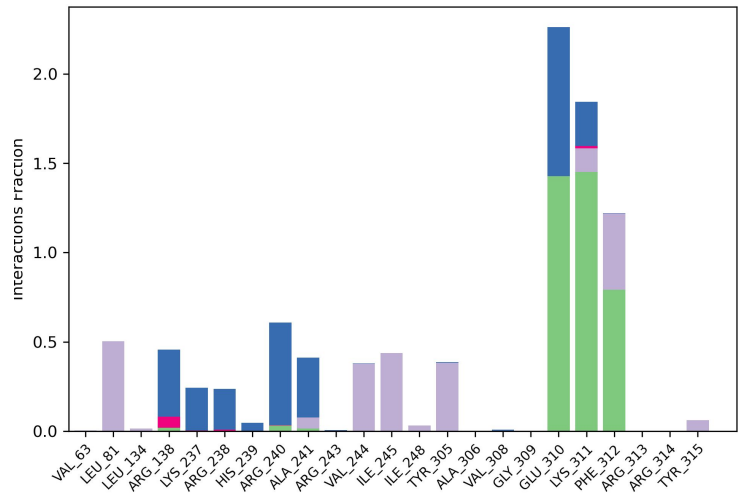



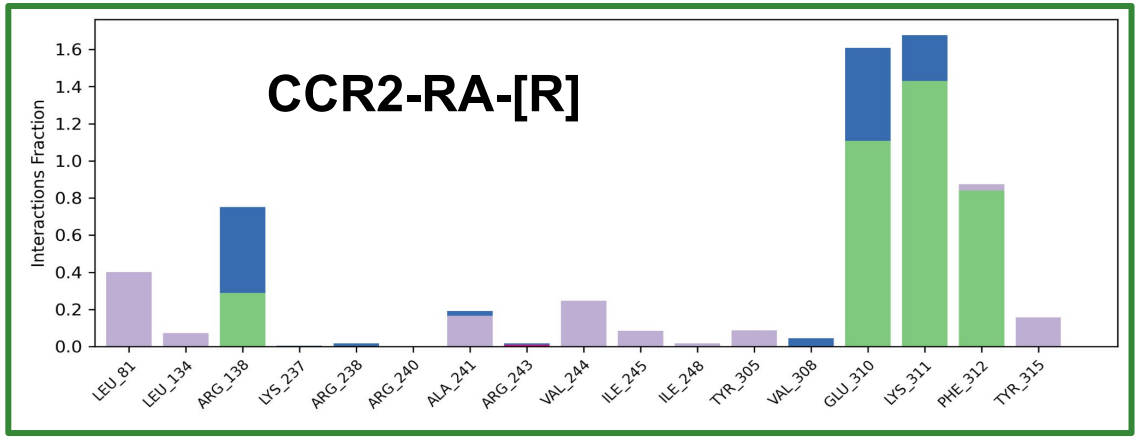



# Candidate v4_100_d5_i



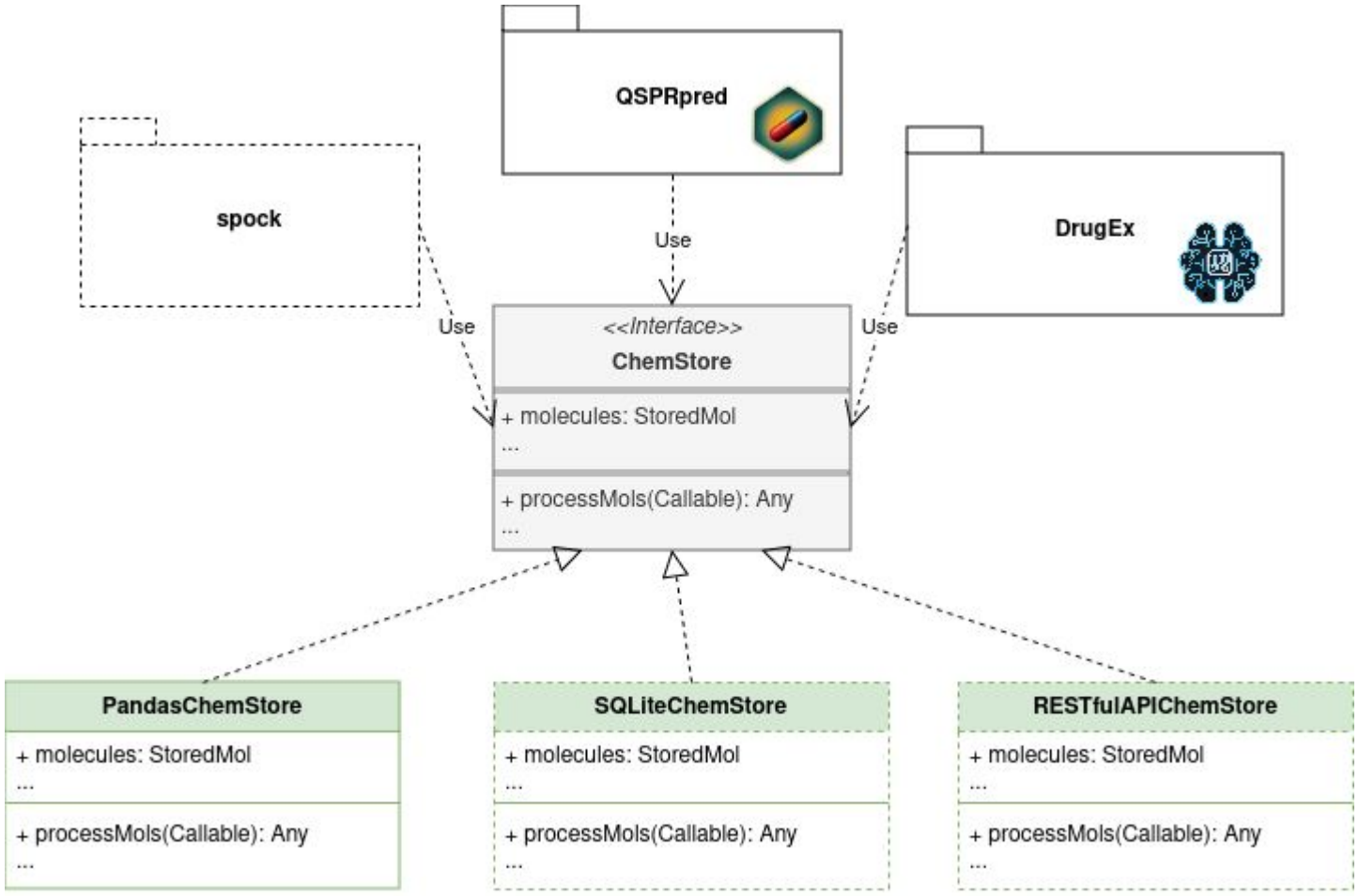




Hydrophobic Ionic Water bridges

# Software Development Perspective – Current Work

https://github.com/CDDLeiden/QSPRpred/tree/dev




- **ChemStore**
  - efficient processing
    - multi-CPU
    - Dask
    - ...
  - molecule representation hierarchy
    - standardization
    - unique identification
    - conformers
    - tautomers
    - ...
  - multiple implementations
    - Pandas
    - SQL Databases
    - ....

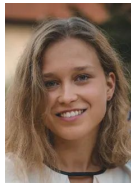
Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

# Conclusions

- Large potential of AI and new methods in de novo drug design
  - -> **validation through real projects is important**
- Good software development practices can help a lot:
  - Documentation and **Tutorials**
  - **GUI** (Streamlit, DataWarrior…)
  - Continuous Integration and Testing
  - Standardization of approaches and extensible API development (even just within one research group)



**Enhancing Evaluation: Scaffold Metrics in Molecular Generator Assessment**

Valeriia Fil [1], Svozil Daniel [1,2]

[1] Department of Informatics and Chemistry & CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Faculty of Chemical Technology, University of Chemistry and Technology, Technická 5, 16628, Prague, Czech Republic.
[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic.

**Introduction**

The exploration of chemical space is crucial for creating new virtual compounds, which are vital for the next generation of drug-like molecules. To achieve this goal, researchers have developed various molecular generation tools. However, evaluating their effectiveness presents a complex challenge. Current evaluation methods primarily focus on technical aspects, verifying that a generated molecule is valid and unique. However, they often neglect the potential biological activity, the ultimate goal in drug discovery. **This work proposes a novel approach to benchmarking chemical structure generators. We focus on scaffolds, the core structures of a molecule, to identify biologically active patterns that the generator can discover, even if they aren't part of the training data.**

**Metrics**

**1. True positive recall all (TUPOR)** - measures recall rate of unique scaffolds.

$$TUPOR = \frac{NAS}{UAS}$$

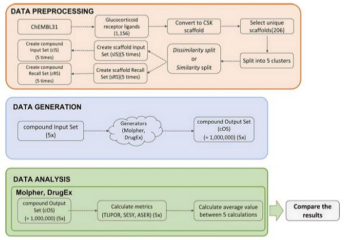
**2. Set Scaffold yield (SESY)** - assesses diversity of output set.

$$SESY = \frac{NS}{SS}$$

**3. Absolute set scaffold recall (ASER)** - evaluates generator's effectiveness.

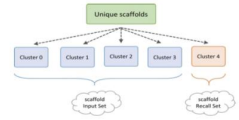
$$ASER = \frac{tRS}{SS}$$

**Project**

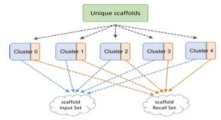



**Split approach**

Using two different approaches for split data for Input Sets(IS) and Recall Sets(RS) to test the bottom and top limits of our metrics. This helps evaluate how generators perform with varying distances between the IS and RS.
**Dissimilarity Split:** sRS was assembled from Cyclic Skeleton (CSK) scaffolds that differed from the CSK scaffolds in the sIS.
**Similarity Split:** sRS was assembled from scaffolds similar to those in sIS.




*Dissimilarity split*




*Similarity split*

**Results**

| Name | Splitting_type | Scaffold | Set_size | TUPOR | SESY | ASER |
|---|---|---|---|---|---|---|
| Molpher_mean | dis | csk | 1 035 435 | 0,5136 | 0,1242 | 0,0046 |
| DrugEx_mean | dis | csk | 999 912 | 0,5555 | 0,3221 | 0,0127 |
| Molpher_mean | sim | csk | 1 047 941 | 0,6786 | 0,1233 | 0,01139 |
| DrugEx_mean | sim | csk | 999 886 | 0,6391 | 0,3388 | 0,01801 |

| Name | Splitting_type | Scaffold | Set_size | TUPOR | SESY | ASER |
|---|---|---|---|---|---|---|
| Molpher_mean | dis | murcko | 1 035 441 | 0,1661 | 0,2585 | 0,0016 |
| DrugEx_mean | dis | murcko | 1 000 000 | 0,1628 | 0,7088 | 0,0012 |
| Molpher_mean | sim | murcko | 1 047 959 | 0,2953 | 0,2582 | 0,0034 |
| DrugEx_mean | sim | murcko | 1 000 000 | 0,2717 | 0,7184 | 0,0036 |

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE
CZ OPENSCREEN

# Acknowledgments

## LACDR Colleagues

- **Supervisors & Consultants:**
  - **Gerard van Westen**
  - **Willem Jespers**
  - Anthe Janssen
  - Laura Heitman
- **Postdocs**
  - Sohvi Luukkonen

## LACDR Students

- **PhD Students**
  - **Helle van den Maagdenberg**
  - Olivier Béquignon
  - **Alan Kai Hassen**
  - **Andrius Bernatavicius**
  - Yorick van Aalst
  - Remco van den Broek
- **Master Students**
  - Chara Spyropoulou
  - Sem Egbers

GitHub

## VŠCHT DICH

- Daniel Svozil
- Wim Dehaen
- **Valeria Fil**
- **Petr Palivec**
- **Jozef Fulop**

GAČR
GRANTOVÁ AGENTURA ČESKÉ REPUBLIKY

CZ OPENSCREEN

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

# Thank you.

Universiteit Leiden

UNIVERSITY OF CHEMISTRY AND TECHNOLOGY PRAGUE

LACDR

Bij ons leer je de wereld kennen