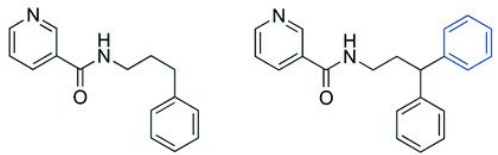
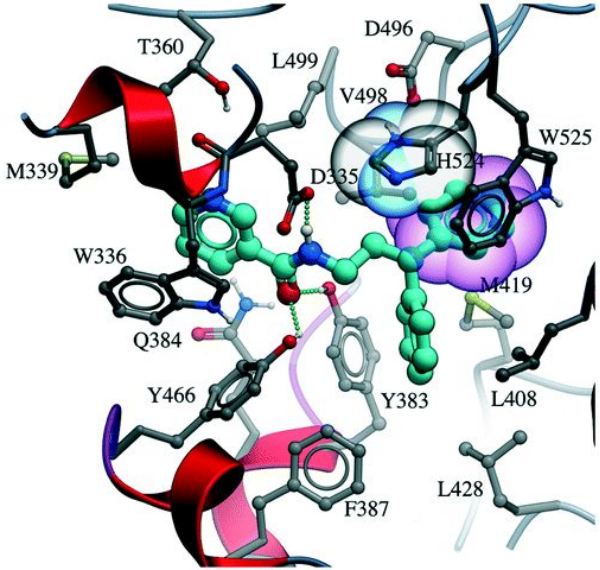
AMIDD 2024 Lecture 8: Lead identification and optimization



5
IC_{50} = 700 nM

6
IC_{50} = 7 nM

(a)

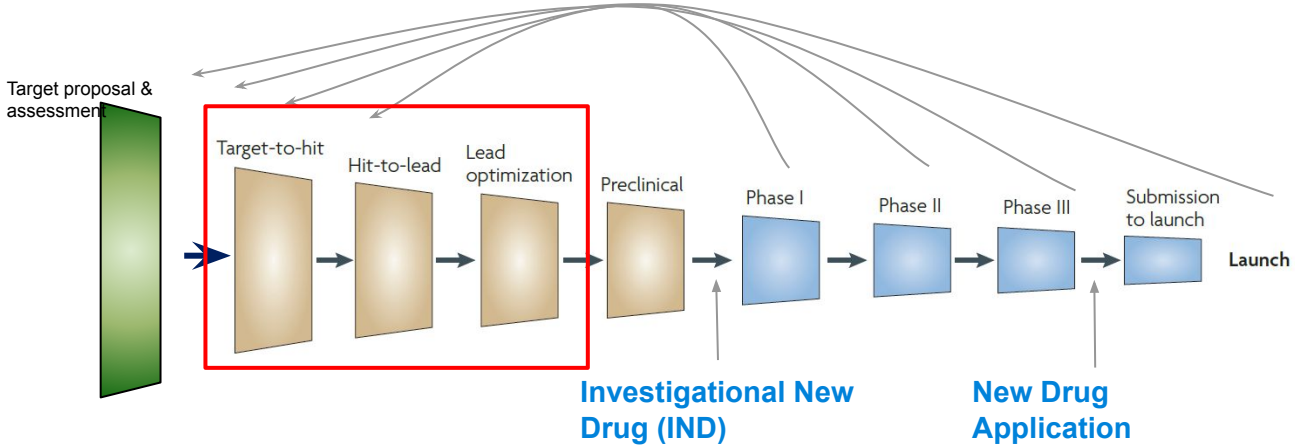


(b)

Freitas, R. F. de & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med. Chem. Commun. 8, 1970–1981 (2017)

Dr. Jitao David Zhang, Computational Biologist
[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche
[2] Department of Mathematics and Informatics, University of Basel

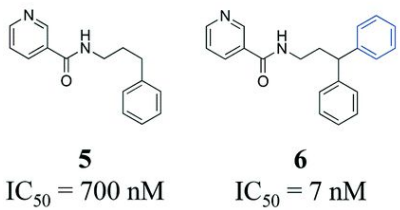
Recap: the linear view of drug discovery



Acronym	TI/TA/TV	Screening	LI	LO
Description	Target identification/ target assessment/ target validation		Lead identification	Lead optimization

Adapted from Paul *et al.* "How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge." Nature Reviews Drug Discovery, 2010

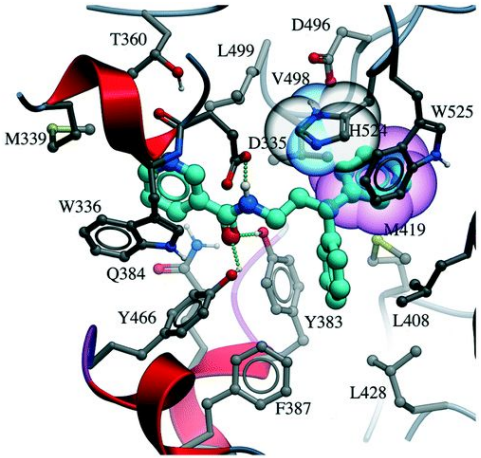
One goal of target-based drug discovery: to make a molecule that binds specifically and strongly to the target protein



5
IC_{50} = 700 nM

6
IC_{50} = 7 nM

(a)



(b)

a) Chemical structure of two inhibitors of human soluble epoxide hydrolase (sEH); b) X-ray cocrystal structure of human sEH and 6 (cyan carbons, PDB: 3I1Y). The phenyl ring (transparent CPK magenta) is positioned to allow a π-stacking interaction with H524 (shown as transparent CPK). Hydrogen bonds are displayed in dotted green lines.

Freitas, R. F. de & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med. Chem. Commun. 8, 1970–1981 (2017)

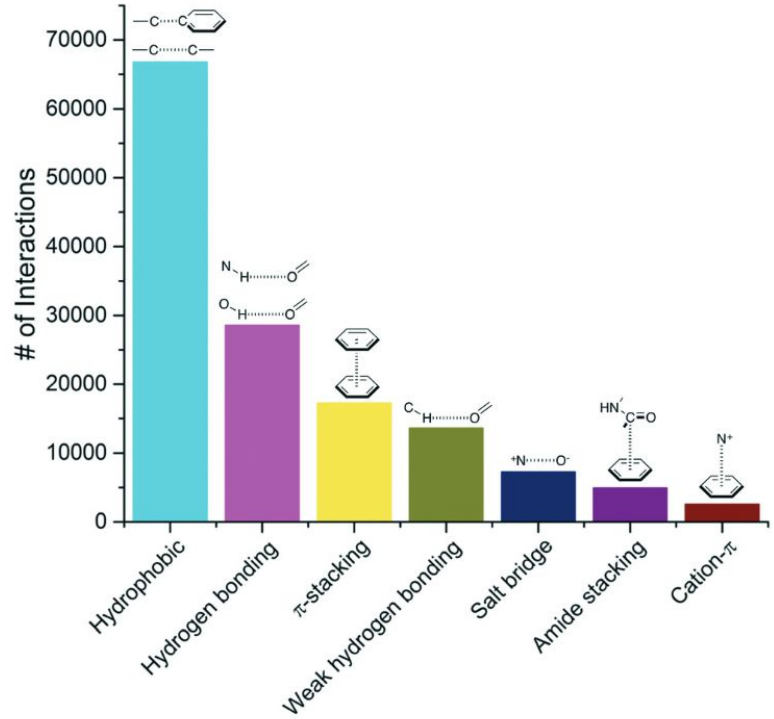


Fig. 1 Frequency distribution of the most common non-covalent interactions observed in protein–ligands extracted from the PDB.

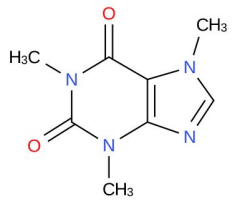
ChEMBL is an information source of small molecules

Nomenclature

caffeine
1,3,7-trimethylxanthine
methyltheobromine

Bioactivity

Affinity to human proteins and drug targets



Chemical data

Formula: $C_8H_{10}N_4O_2$
Charge: 0
Mass: 194.19

Database Xrefs

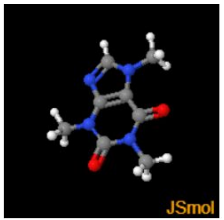
PubChem: CID2519
BindingDB: 1849

Chemical Informatics

***InChI**=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3*

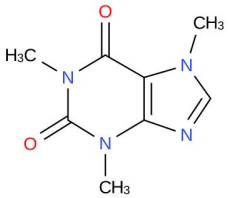
***SMILES**: CN1C(=O)N(C)c2ncn(C)c2C1=O*

Visualisation



A subset of available information from EBI ChEBI/ChEMBL, inspired by EBI's roadshow *Small Molecules in Bioinformatics*

Representation of small molecules



Molfile: View Raw | Download | Editor | Copy

Canonical SMILES: CN1C(=O)N(C)c2ncn(C)c2C1=O

Standard InChI: InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Standard InChI Key: RYYVLZVUVIJVGH-UHFFFAOYSA-N

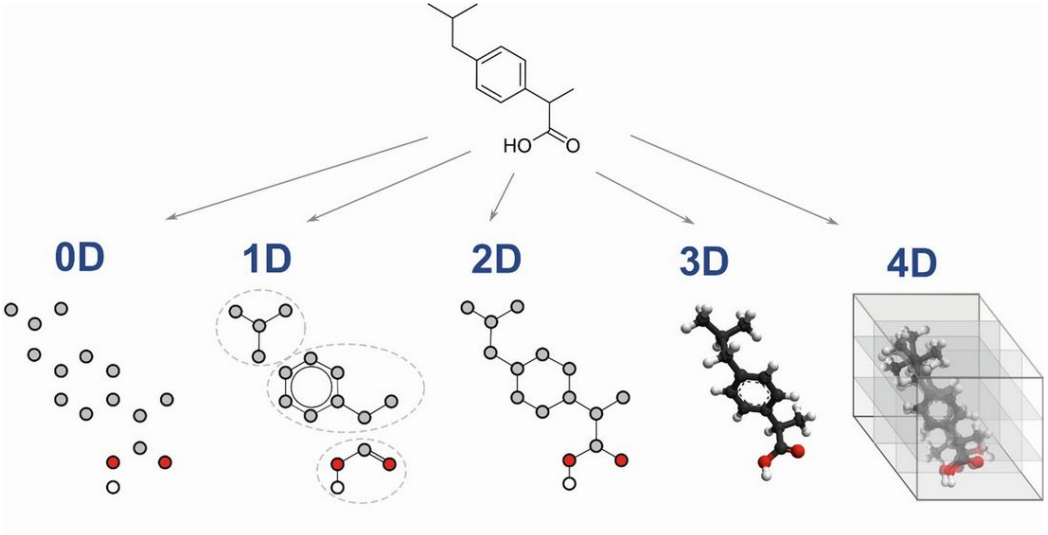
- Simplified Molecular-Input Line-Entry System (SMILES)
- IUPAC International Chemical Identifier (InChI)
- InChiKey: a 27-character, hash version of InChI
- Molfile: a type of chemical table files

```
CHEMBL113

SciTegic12231509382D

14 15 0 0 0 0 999 V2000
-1.1875 -9.6542 0.0000 C 0 0
-1.1875 -8.9625 0.0000 C 0 0
-1.8125 -10.0292 0.0000 N 0 0
-2.4167 -8.9625 0.0000 N 0 0
-2.4167 -9.6542 0.0000 C 0 0
-1.8125 -8.6000 0.0000 C 0 0
-0.5000 -9.8917 0.0000 N 0 0
-0.5000 -8.7625 0.0000 N 0 0
-0.1125 -9.3042 0.0000 C 0 0
-3.0250 -10.0375 0.0000 O 0 0
-1.8125 -7.8917 0.0000 O 0 0
-1.8125 -10.7417 0.0000 C 0 0
-3.0250 -8.6000 0.0000 C 0 0
-0.2917 -8.0750 0.0000 C 0 0
2 1 2 0
3 1 1 0
4 5 1 0
5 3 1 0
6 2 1 0
7 1 1 0
8 2 1 0
9 7 2 0
10 5 2 0
11 6 2 0
12 3 1 0
13 4 1 0
```

Molecular descriptors: numeric values that describe chemical molecules

In contrast to symbolic representations, molecular descriptors enable **quantification of molecular properties**.

Molecular descriptors allows mathematical operations and statistical analysis that associate biophysical or biochemical properties with molecule structures.



-Atom count
-Molecular weight
-Sum of atomic properties

-Fragment counts, *e.g.* # of -OH
-Fingerprints

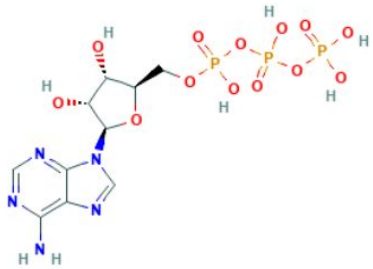
-Topological descriptors, *e.g.* the Wiener Index based on graph theory
- ECFP

-Geometrical
-Atomic coordinates
-Energy grid

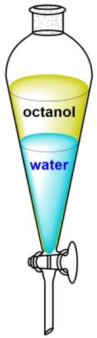
-Combination of atomic coordinates and sampling of possible conformations

Selected commonly used molecular descriptors

Molecular Weight (MW). for example, adenosine triphosphate (ATP), the *energy molecule*, has a MW of 507.



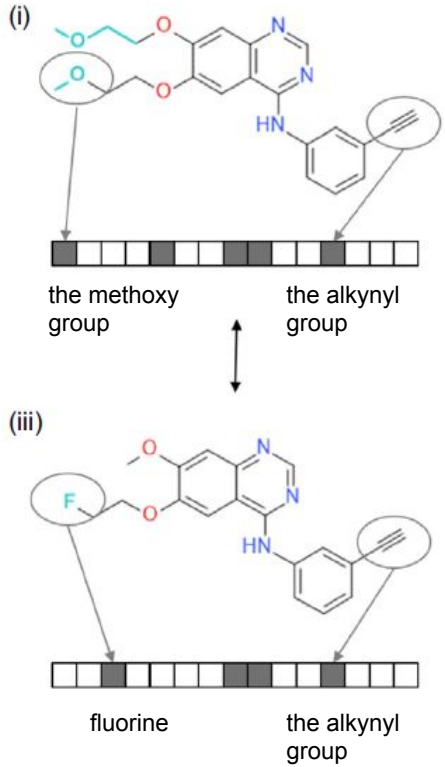
logP (partition coefficient) quantifies the hydrophilicity and hydrophobicity of a molecule. The calculated version (cLogP) exists as well.
While logP is independent of pH, logD measures the pH-dependent lipophilicity of ionizable molecules.



$$\log P_{\mathrm{oct/wat}} = \log\left(\frac{[\mathrm{solute}]^{\text{un-ionized}}_{\text{octanol}}}{[\mathrm{solute}]^{\text{un-ionized}}_{\text{water}}}\right)$$

Molecular fingerprints: a set of techniques to represent molecules in a bit array.

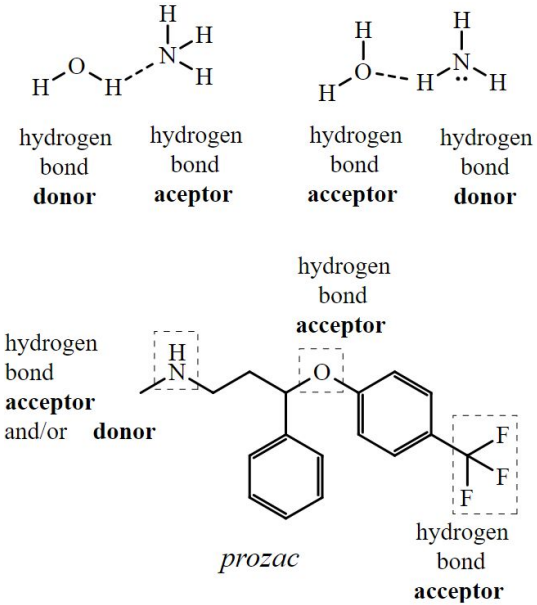
The Tanimoto coefficient, similar to Jaccard Index, makes it easy to compare molecules pairwise.

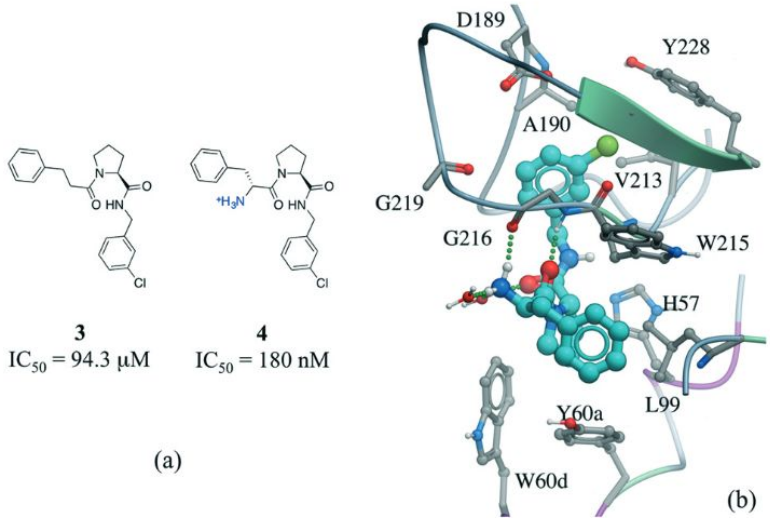


Number of hydrogen bond acceptors and donors are important descriptors, too

A **hydrogen bond**: an electrostatic force of attraction between a hydrogen (H) atom which is covalently bonded to a more electronegative "**donor**" atom or group (Dn), and another electronegative atom bearing a lone pair of electrons—the hydrogen-bond **acceptor**.

Hydrogen bonds (H-bonds) both influence the structure of the molecule and its binding to the target.

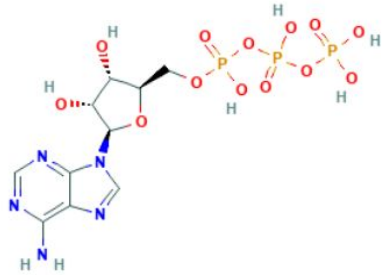




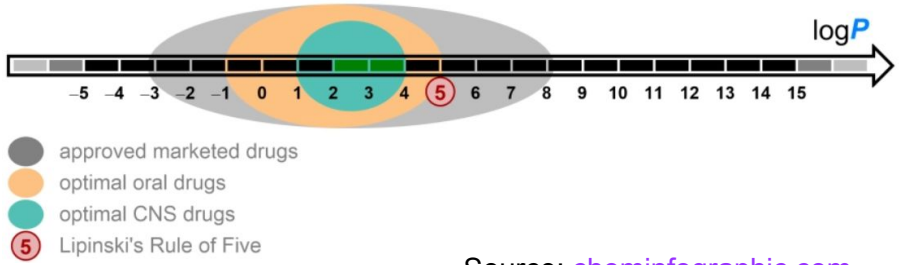
Effect of adding a hydrogen bond in a thrombin inhibitor: a) chemical structure of a pair of thrombin inhibitors; b) crystal structure of molecule 4 (cyan carbons) in complex with thrombin (PDB: 2ZC9). Hydrogen bonds are displayed in dotted green lines.

Lipinski's Rule of Five of small-molecule drugs

- **HBD<=5:** No more than **5 hydrogen-bond donors,** *e.g.* the total number of nitrogen–hydrogen and oxygen–hydrogen bonds.
- **HBA<=10**: No more than **10 hydrogen-bond acceptors**, *e.g.* all nitrogen or oxygen atoms
- **MW<500**: A **molecular weight** less than **500 Daltons, or** 500 g/mol.
- **logP<=5**: An **octanol-water partition coefficient (log P)** that does not exceed **5.** (10-based)

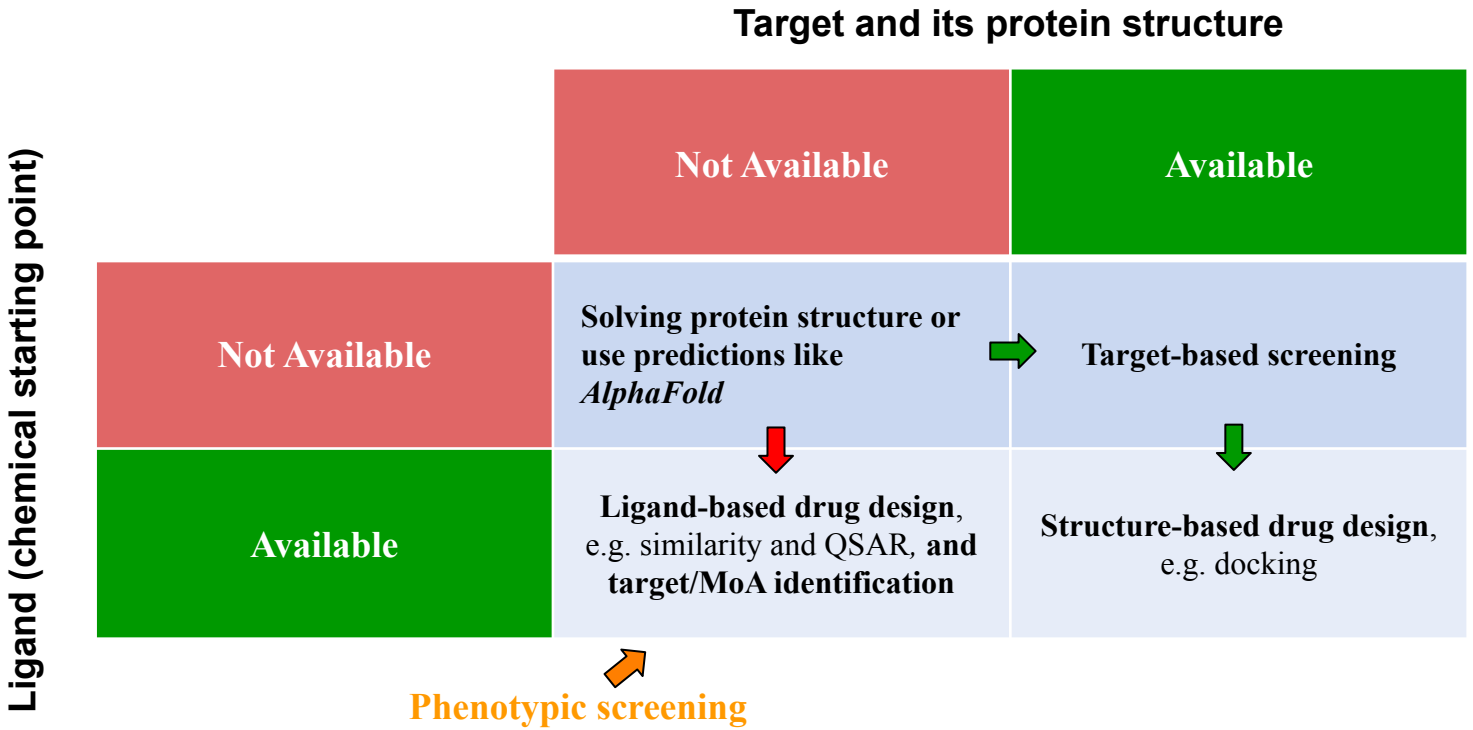


ATP (MW=507)



Source: cheminfographic.com

Ligand-based and structure-based drug design

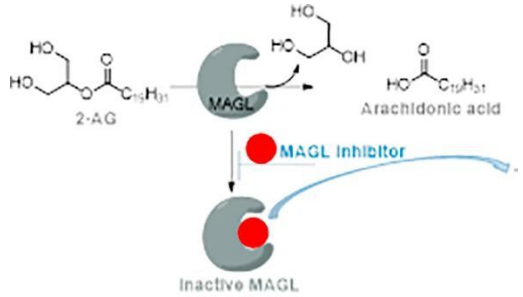


QSAR= quantitative structure activity relationship; MoA= mechanism of action, or mode of action

Things you need to start a target-based drug discovery program

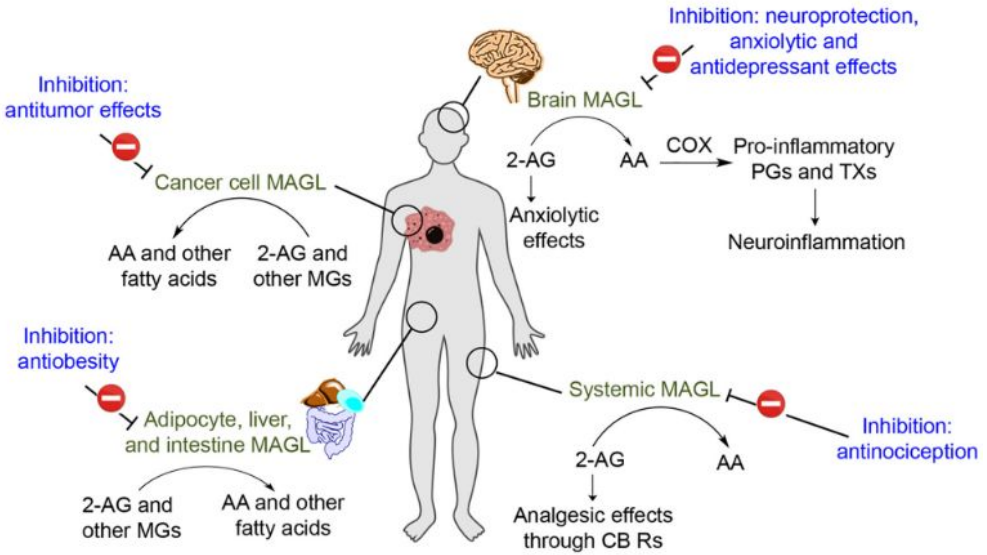
1. **Indication and patient profile**
2. **Protein target of interest**, *e.g.* evolutionary conservation, isoforms, expression profiles, structure, effect of modulation, *etc.*
3. **Modality**, shall we use small molecule, large molecule, antisense oligonucleotides or gene therapy? Shall we activate the target or inhibitor it? ...
4. **Tool compounds and competitor information,** if available
5. **Target (ideal) profile of your drug**
6. **Material for the campaign**: enough amount of proteins, compounds, etc.

Protein MAGL (MGLL) is a key protein linking metabolism and inflammation that may be therapeutically exploited



The MAGL protein, known as monoglyceride lipase, is encoded by the MGLL gene in human. MAGL is a key enzyme in the hydrolysis of the endocannabinoid 2-arachidonoylglycerol (2-AG). It converts 2-AG into arachidonic acid (AA) and glycerol.

MAGL has been proposed as a potential target for many diseases. However, there has been no approved drugs targeting MAGL.



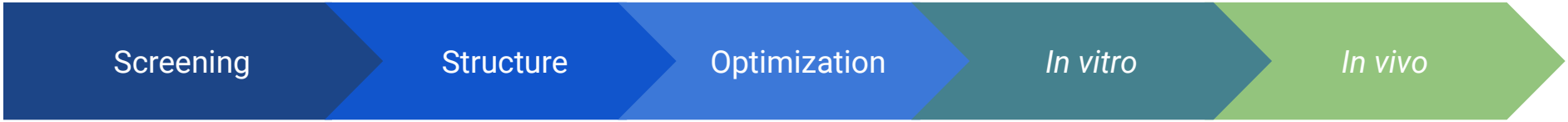
Figures above: Gil-Ordóñez *et al.*, Biochemical Pharmacology, 2018

A drug candidate, BIA 10-2474, which targets a closely related protein, fatty acid amide hydrolase (FAAH), caused serious adverse events in trial participants, leading to death of one man, in Rennes, Frances, 2016. The event has complicated the development of other drugs targeting the endocannabinoid system.

Things you need to start a target-based drug discovery program

1. **Indication and patient profile:** Multiple sclerosis (among others)
2. **Protein target of interest**: MAGL
3. **Modality**: small-molecule inhibitor
4. **Tool compounds and competitor information**: patents, publications, etc.
5. **Target (ideal) profile of your drug**: reversible, selective, brain penetrating
6. **Material for the campaign**: ready

Key steps in lead identification (LI) and lead optimization (LO)



Screening	Structure	Optimization	*In vitro*	*In vivo*
Identify molecules (*hits* and *leads*) that bind to and modulate the target protein	Determine the mode of binding between the lead molecule and the target	Optimize the potency, selectivity, and physico-chemical properties of the molecule	Evaluate pharma-cology, ADME, and safety profiles of the molecule *in vitro*	Evaluate pharma-cology, ADME/PK, and safety in animals (or alternative systems)

Example: Identification of a novel, reversible MAGL inhibitor with high potency and selectivity, and excellent absorption, distribution, metabolism, and excretion (ADME) properties

Screening → Structure → Optimization → *in vitro* → *in vivo*

Journal of Medicinal Chemistry

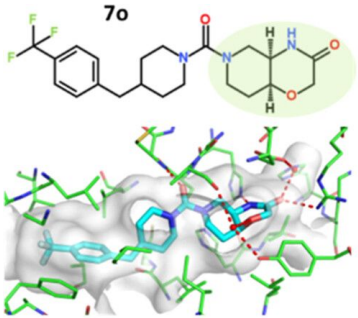
pubs.acs.org/jmc Article

Structure-Guided Discovery of *cis*-Hexahydro-pyrido-oxazinones as Reversible, Drug-like Monoacylglycerol Lipase Inhibitors

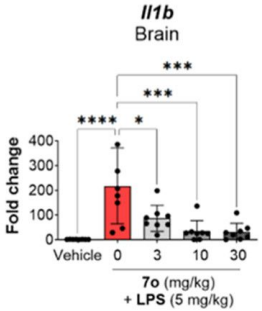
Bernd Kuhn,* Martin Ritter, Benoit Hornsperger, Charles Bell, Buelent Kocer, Didier Rombach, Marius D. R. Lutz, Luca Gobbi, Martin Kuratli, Christian Bartelmus, Markus Bürkler, Raffael Koller, Paolo Tosatti, Iris Ruf, Melanie Guerard, Anto Pavlovic, Juliane Stephanus, Fionn O'Hara, Dennis Wetzl, Wiebke Saal, Martine Stihle, Doris Roth, Melanie Hug, Sylwia Huber, Dominik Heer, Carsten Kroll, Andreas Topp, Manfred Schneider, Jürg Gertsch, Sandra Glasmacher, Mario van der Stelt, Andrea Martella, Matthias Beat Wittwer, Ludovic Collin, Jörg Benz, Hans Richter, and Uwe Grether*

Cite This: *J. Med. Chem.* 2024, 67, 18448–18464 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



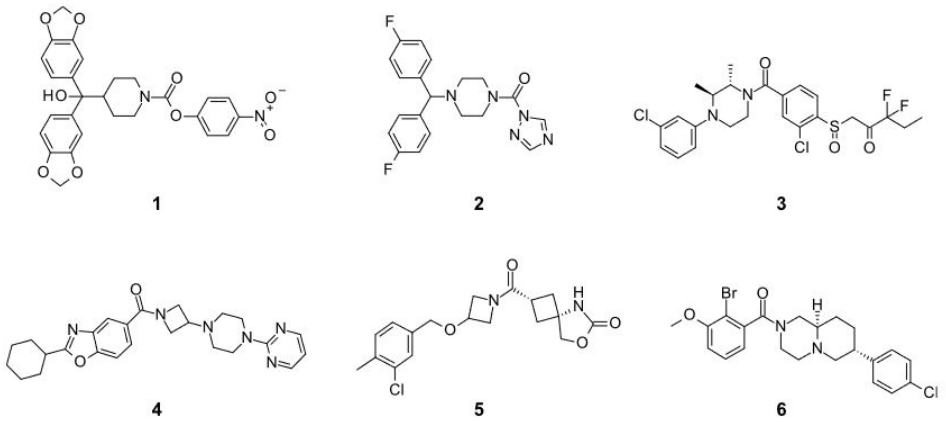
high MAGL potency
excellent selectivity

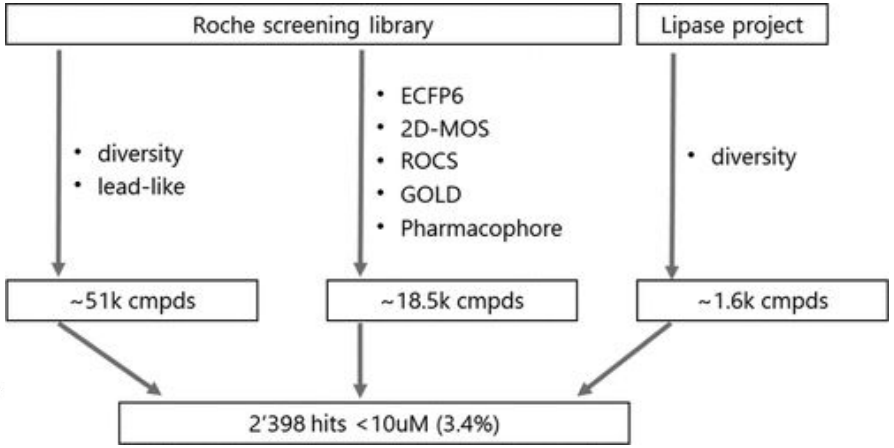


drug-like ADMET properties
in vivo efficacy

Kuhn, B. et al. Structure-Guided Discovery of *cis*-Hexahydro-pyrido-oxazinones as Reversible, Drug-like Monoacylglycerol Lipase Inhibitors. J. Med. Chem. 67, 18448–18464 (2024).

Starting point: a library of compounds that may modulate the target





Left: selected published compounds targeting MAGL protein. Right: library assembly

ECFP6: extended connectivity fingerprints, diameter of bond distance 6, using compounds extracted from existing MAGL patents as references

2D-MOS: two-dimensional graph-based maximum overlapping spheres similarity search

ROCS: three-dimensional shape-based similarity search

GOLD: Docking program

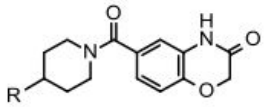
Pharmacophore: molecular features that necessary for molecular recognition between ligand and target

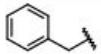
Screening hit 7a and early optimizations

The hit 7a (benzoxazinone) was selected because of the high *ligand efficiency* (LE).

Ligand efficiency is defined as the binding energy (ΔG) divided by the number of non-hydrogen atoms. The binding energy is a function of IC_{50}: higher affinity means larger absolute value of the binding energy.

The co-structure of 7a and human MAGL protein was soon solved with X-ray crystallography (see right). The structure, when compared with the co-structures of other compounds, helps to guide medicinal chemists' work to optimize the molecule.



Compound	R	MAGL inh. (IC_{50}, μM)[a]	MAGL $pIC_{50} \pm SD$[b]	LE[c]	logD[d]	Solubility[f] [μg/mL]
7a		0.075	7.13 ± 0.08	0.38	3.2	1.1

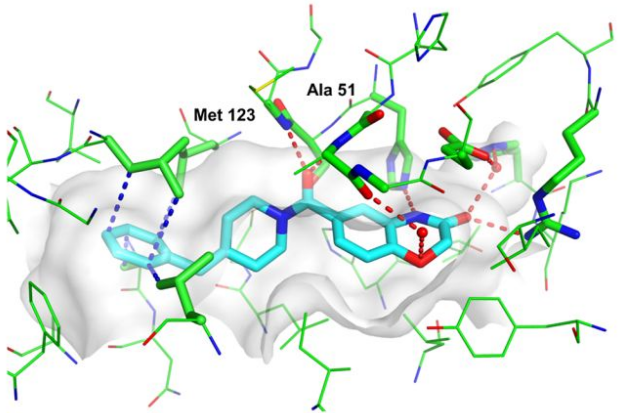
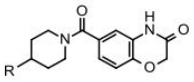


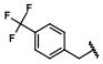
Figure 3. Co-crystal structure of human MAGL (green) with focused screening hit **7a** (cyan, pdb code: 9F8A). Residues of the oxyanion hole (Ala51 and Met123) which are on top of the catalytic Ser122 are labeled. Hydrogen bonds between the ligand and the protein or water are shown as red, dashed lines. Dispersion interactions with short distances (<4.0 Å) in the hydrophobic pocket on the left are highlighted as blue lines. Water molecules are removed except for the ones that form hydrogen bonds with the ligand.

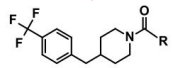
Further optimizations guided by structures, experiments, and machine learning predictions

Medicinal chemists, computer-aided drug design (CADD) experts, structural biologists and other colleagues worked together to build hypotheses about which changes may improve potency and ADME properties.

They synthesized new compounds, solved more structures, comparing structures and activities, thereby gaining insights in structure-activity relationship (SAR).



Compound	R	MAGL inh. (IC_{50}, μM)[a]	MAGL $pIC_{50} \pm SD$[b]	LE[c]	logD[d]	Solubility[f] [μg/mL]
7a		0.075	7.13 ± 0.08	0.38	3.2	1.1
7b		0.062	7.24 ± 0.17	0.37	3.3	0.8
7c		0.032	7.54 ± 0.24	0.34	3.8	4.1
7d		0.051	7.31 ± 0.12	0.37	3.6	36
7e		0.046	7.35 ± 0.09	0.34	3.9	15
7f		0.0005	9.36 ± 0.15	0.40	3.8	<0.1

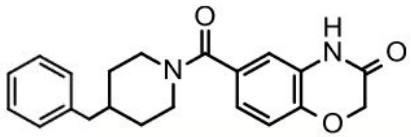


Compound	R		MAGL inh. (IC_{50}, μM)[a]	MAGL $pIC_{50} \pm SD$[b]	LE[c]	logD[d]	Solubility[f] [μg/mL]
7c			0.032	7.54 ± 0.24	0.34	3.8	4.1
7k			2.5	5.62 ± 0.11	0.26	3.7[e]	6.9
7l			0.965	6.04 ± 0.14	0.29	3.8[e]	2.9
7m		and enantiomer	0.370	6.46 ± 0.17	0.30	3.2	78
7n		and enantiomer	0.065	7.22 ± 0.17	0.33	3.2	225
7o			0.032	7.55 ± 0.22	0.34	3.3	231
7p			2.2	5.65 ± 0.04	0.26	3.3	114
7q		or enantiomer	0.569	6.30 ± 0.24	0.30	3.2	377

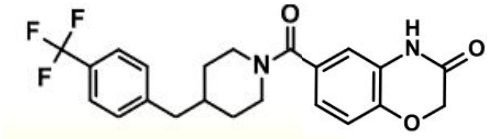
Some parameters, for instance logD values marked by superscripts *e*, are predicted by in-house machine-learning algorithms.

From hit to lead, and from lead to *in vivo* candidate HHPO 7o

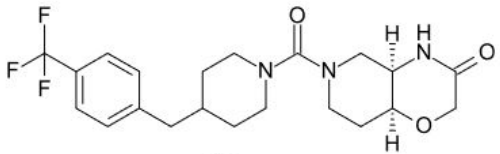
Hit:



Lead



In vivo candidate HHPO 7o:



7o

Table 3. Physicochemical Properties of HHPO 7o

MW [g/mol]	425.45
PSA [Å^2]	55
solubility[a] [μg/mL]	231
log D[b]	3.3
LIMBA log D_{brain}[c]	1.8
PAMPA P_{eff} [cm/s*10 × 10^{-6}][c]	12.8
chemical stability in aqueous buffer	stable at pH 1, 4, 6.5, 8, and 10 for 2 h at 37 °C

[a]Aqueous solubility at pH 6.5 in 0.05 M phosphate buffer. [b]pH 7.4. [c]See the Experimental Section. HHPO=cis-hexahydro-pyrido-oxazinone

- MW: Molecular Weight
- PSA: Polar Surface Area (drugs with PSA<90 Å² penetrates blood-brain barrier more easily).
- LIMBA: lipid membrane binding assay. Low value→lower risk of non-specific binding in the brain.
- PAMPA: Parallel artificial membrane permeability assay. High value→penetrates cell membrane.

In vitro pharmacology, ADME, and safety profile

Table 4. In Vitro Pharmacology Data of HHPO 7o

human/mouse/rat/cyno MAGL inhibition IC_{50} [nM][a]	32/91/71/16
human/rat MAGL SPR K_d [nM][b]	45/64
human/rat MAGL SPR	$2.5 \times 10^5/2.6 \times 10^5$
k_{on} [1/M/s]/k_{off} [1/s]	$7.5 \times 10^{-3}/16.8 \times 10^{-3}$
off-target panel screen against 50 representative proteins at 10 μM[c]	inhibition <50% except for serotonin transporter (57%)
hydrolase selectivity in gel-based ABPP assay at 10 μM[d]	no off-targets

Table 5. ADME and Rat PK Profile of MAGL Inhibitor 7o as well as Mouse PK Profile of the Racemic Mixture 7n

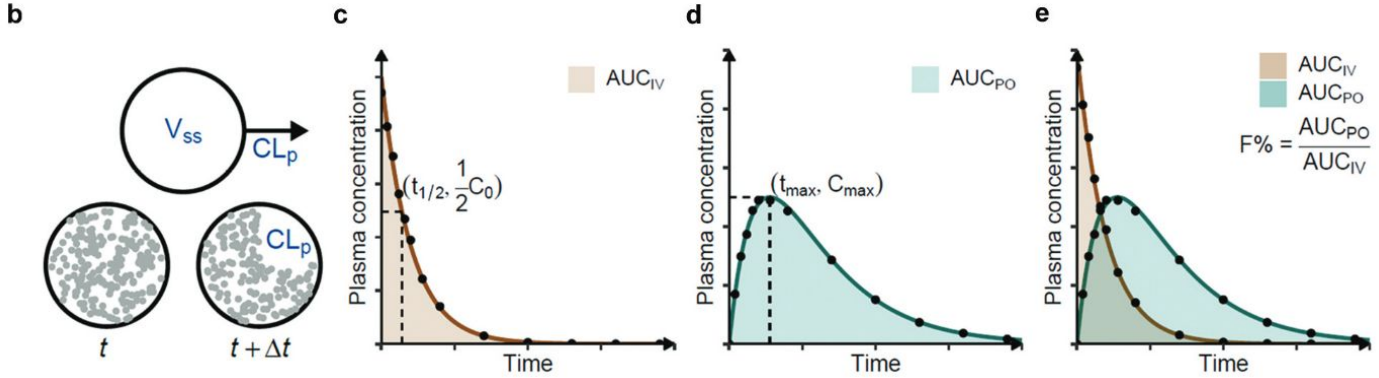
human/mouse/rat clearance in microsomes [μL/min/mg protein][a]	$<10^b/<10^b/11$
human/mouse/rat clearance in hepatocytes [μL/min/10×10^6 cells][c]	1.9/8.9/4.0
human/mouse/rat plasma protein binding [%][d]	4.3/4.7/3.2
human/mouse P-glycoprotein efflux ratio[d]	2.4/4.0
$P_{app,\ AB\ inhibitor}$ [nm/s]	176.3/243.9

Table 6. In Vitro Safety Profile of MAGL Inhibitor 7o

CYP inhibition@10 μM (3A4/2C9/2D6) [%][a]	−195/3/−12
GSH (human liver microsomes) adducts[b]	none detected
hERG IC_{50} [μM][c]	7.7
Ames/MNT/phototoxicity	all negative

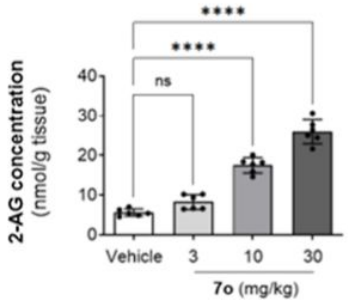
Offline activity: using mandatory reading material and any resource (including Wiki and ChatGPT) to understand and explain following concepts: (1) SPR, (2) K_d, (3) k_{on} and k_{off}, (4) microsomal clearance, (5) plasma protein binding, (6) GSH adducts, (7) hERG assay, (8) Ames test, (9) micronucleus test (MNT), and (10) phototoxicity.

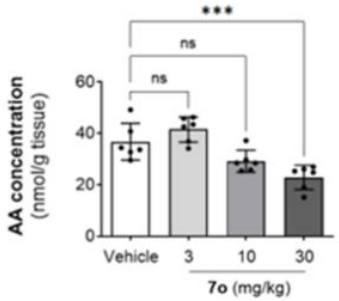
In vivo PK profiles

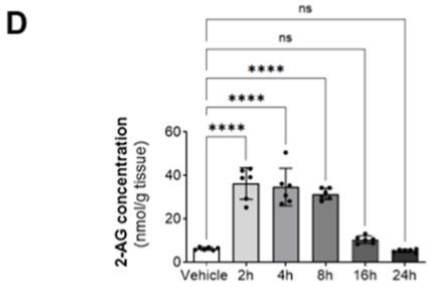


mouse PK profile[e]		rat PK profile[f]	
CL [mL/min/kg] (i.v.)	23	CL [mL/min/kg] (i.v.)	4.6
V_{ss} [L/kg] (i.v.)	3.0	V_{ss} [L/kg] (i.v.)	1.3
$T_{1/2}$ [h]	1.5 (i.v.)/2.9 (i.p.)	$T_{1/2}$ [h]	5.1 (i.v.)/2.7 (i.p.)
C_{max} [ng/mL]	896 (p.o.)/12,200 (i.p.)	C_{max} [ng/mL]	1290 (p.o.)/12,200 (i.p.)
T_{max} [h]	2.7 (p.o.)/1.5 (i.p.)	T_{max} [h]	6.0 (p.o.)/0.25 (i.p.)
AUC_{last} [h*ng/mL]	4560 (p.o.)/62,100 (i.p.)	AUC_{last} [h*ng/mL]	19,100 (p.o.)/40,200 (i.p.)
F [%]	127 (p.o.)/368 (i.p.)	F [%]	115 (p.o.)/111 (i.p.)
brain/plasma ratio (average)	1.7 (i.p.)	brain/plasma ratio (average)	1.1 (i.p.)
CSF/plasma ratio (average); $K_{p,uu}$	0.0164 (i.p.); 0.35	CSF/plasma ratio (average); $K_{p,uu}$	0.017 (i.p.); 0.53

Compound 7a shows reasonable target engagement





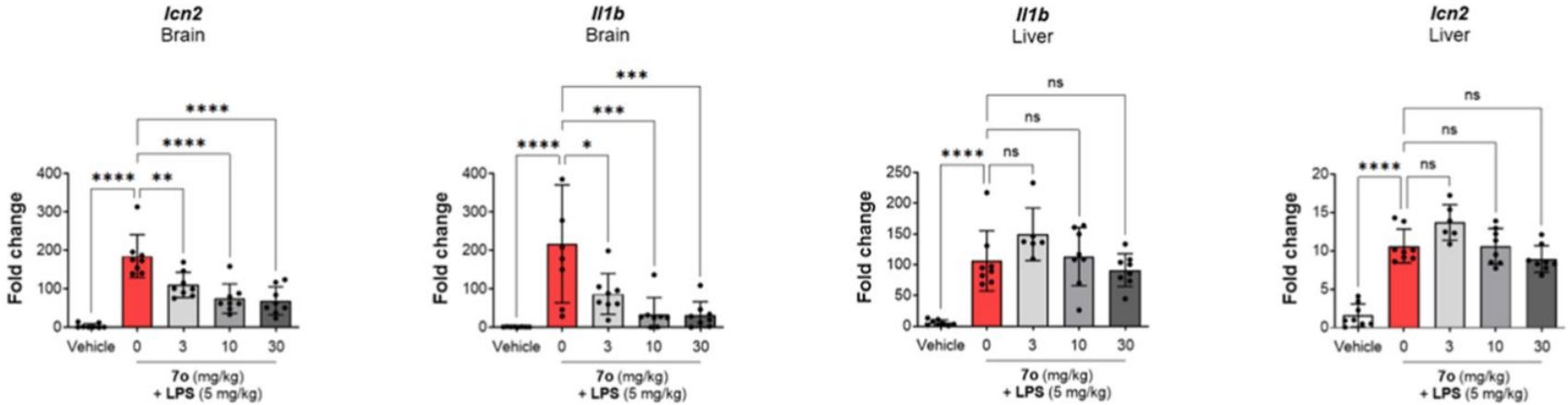


Dose (mg/kg)	Measured average plasma conc. (nM)	Calculated CSF conc. (nM)	Calculated % target occupancy
3	1901	31.2	49
10	7496	123	79
30	27765	455	93

Time (h)	Measured average plasma conc. (nM)	Calculated CSF conc. (nM)	Calculated % target occupancy
2	42040	689	96
4	38588	633	95
8	21089	346	92
16	2955	48	60
24	100	2	5

Left: sampling after 4h of varying doses. Right: one dose (30 mg/kg), varying time points

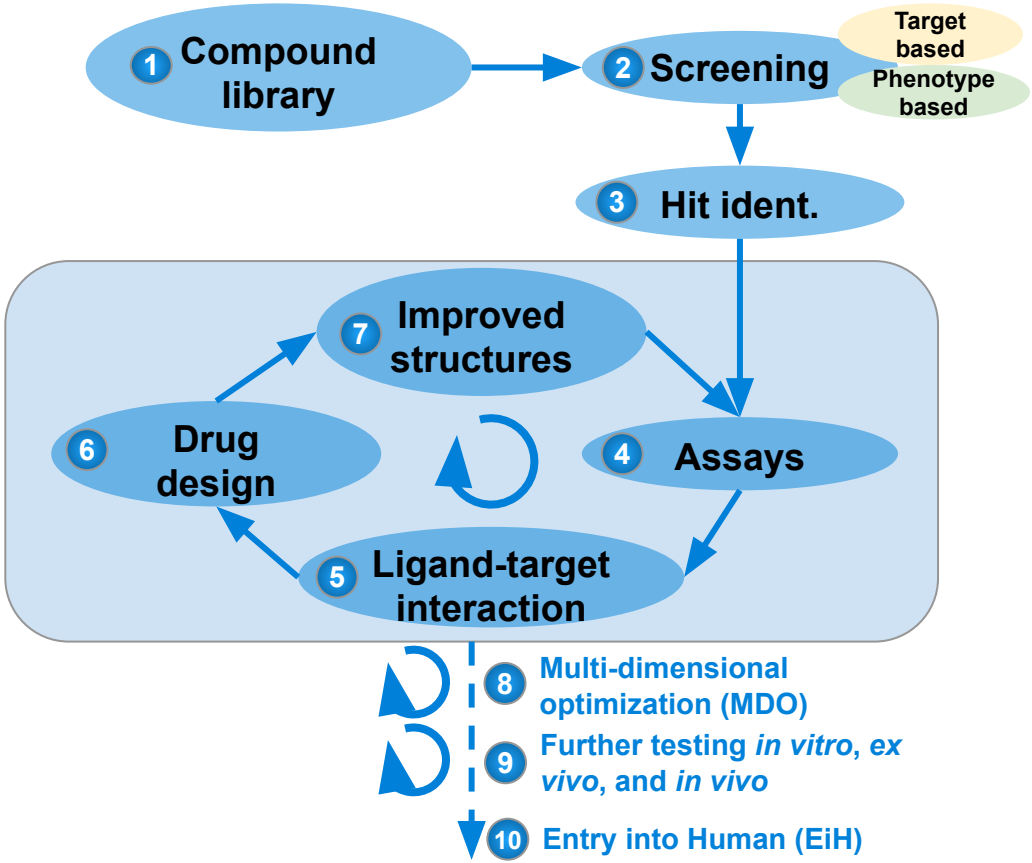
Biomarker study confirms a brain-specific anti-inflammatory effect



Efficacy of the compound 7o in the LPS (lipopolysaccharide) model of neuroinflammation. Mice were dosed with vehicle, LPS, or LPS + 7o (3, 10, 30 mg/kg intraperitoneal). mRNA analyses of the Icn2 gene (lipocalin-2) and the Il1b (interleukin 1-beta) demonstrated the anti-inflammatory properties of compound 7o specifically in the brain, not in liver.

Workflow in a typical target-based drug-discovery program

1. Compound library construction (small molecules, large molecules, RNA therapeutics, or other modalities)
2. Screening compounds with ***bioassays***, or ***assays***, which determine potency of a chemical by its effect on biological entities: proteins, cells, *etc*;
3. Hit identification and clustering;
4. More assays, complementary to the assays used in the screening, maybe of lower throughput but more biologically relevant;
5. Analysis of ligand-target interactions, for instance by getting the co-structure of both protein (primary target, and off-targets if necessary) and the hit;
6. *Drug design,* namely to modify the structure of the drug candidate;
7. Analog synthesis and testing (back to step 4);
8. Multidimensional Optimization (MDO), with the goal to optimize potency, selectivity, safety, bioavailability, *etc;*
9. Further *in vitro*, *ex vivo*, and *in vivo* testing, and preclinical development;
10. Entry into human (Phase 0 or phase 1 clinical trial).



A schematic presentation of structure-based drug discovery

Summary

1. Target-based drug discovery relies on screening, structure information, and iterative optimization to identify drug-like molecules.
2. Drug-like molecules are tested *in vitro* and *in vivo* to ensure their pharmacological, ADME, and safety profiles.
3. Chemoinformatics, computer-aided drug design, and machine learning play important roles in lead identification and optimization.

Backup slides

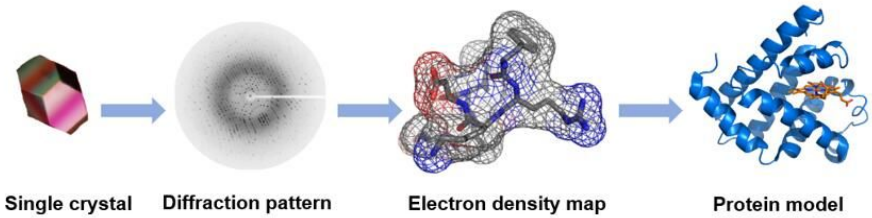
Free training opportunity offered by U.S. Food & Drug Agency: FDA Clinical Investigator Training Course 2024

The primary goal of this clinical investigator training course is to provide participants with the essential knowledge and skills to conduct clinical trials effectively, ethically, and in accordance with regulatory standards. Participants will acquire a practical understanding of:

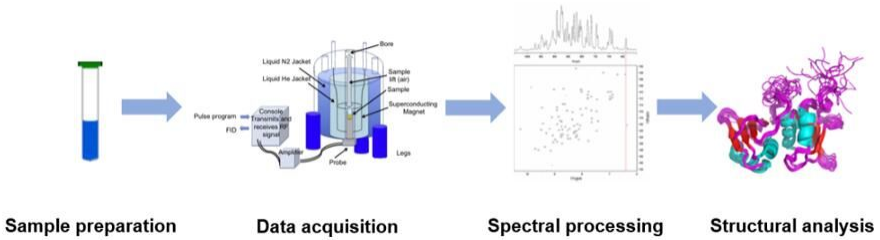
- FDA's approach to trial design
- Statistical issues in the analysis of trial data
- Safety concerns in the development of medical products
- Understanding preclinical information relevant to medical product development
- Clinical investigator responsibilities

The agenda is designed for clinical investigators, health care professionals (physicians, nurses, pharmacists, *etc.*), and individuals involved in biomedical research and the development of drugs and biological products. The course will take place on December 10-12th, 2024 via Zoom, between 11AM and ~ 4PM ET (17:30-22:00 in Switzerland). Information and registration: https://www.fda.gov/drugs/news-events-human-drugs/fda-clinical-investigator-training-course-citc-2024-12102024

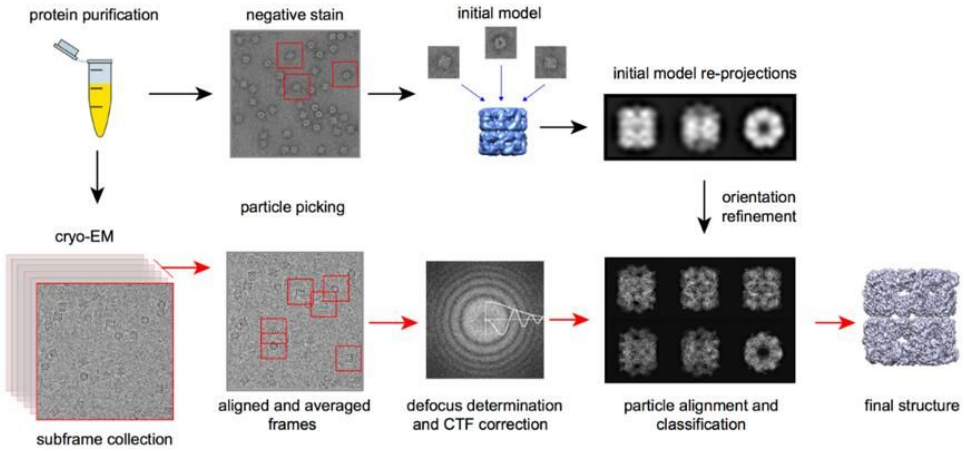
X-ray, NMR, and CryoEM are major experimental approaches to determining protein structures



X-ray crystallography



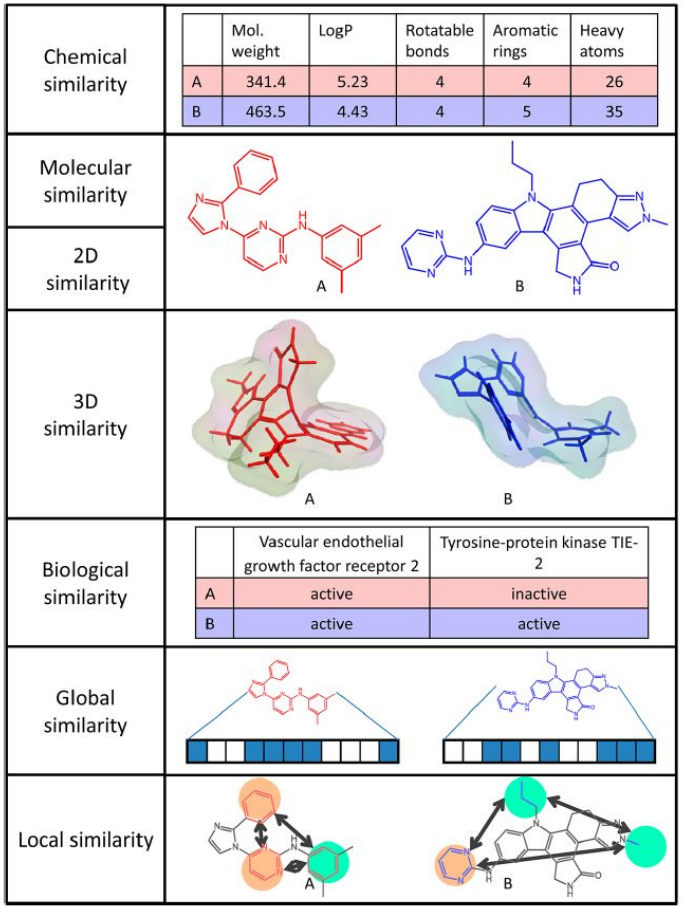
Nuclear Magnetic Resonance (NMR)



Cryo-electron microscopy (CryoEM)

Figure sources:
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

Molecular similarity and similarity measures



	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
A	341.4	5.23	4	4	26
B	463.5	4.43	4	5	35

	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
A	active	inactive
B	active	active

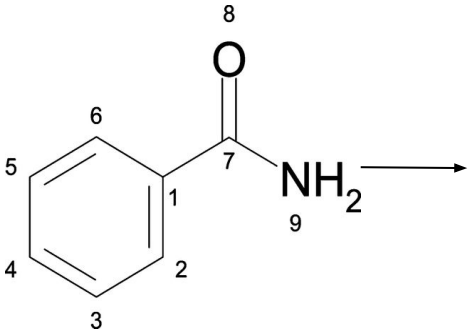
Table 2 Formulas for the various similarity and distance metrics

Distance metric	Formula for continuous variables[a]	Formula for dichotomous variables[a]
Manhattan distance	$D_{A,B}=\sum_{j=1}^{n}\lvert x_{jA}-x_{jB}\rvert$	$D_{A,B}=a+b-2c$
Euclidean distance	$D_{A,B}=\left[\sum_{j=1}^{n}(x_{jA}-x_{jB})^2\right]^{1/2}$	$D_{A,B}=[a+b-2c]^{1/2}$
Cosine coefficient	$S_{A,B}=\left[\sum_{j=1}^{n}x_{jA}x_{jB}\right]/\left[\sum_{j=1}^{n}(x_{jA})^2\sum_{j=1}^{n}(x_{jB})^2\right]^{1/2}$	$S_{A,B}=\frac{c}{[ab]^{1/2}}$
Dice coefficient	$S_{A,B}=\left[2\sum_{j=1}^{n}x_{jA}x_{jB}\right]/\left[\sum_{j=1}^{n}(x_{jA})^2+\sum_{j=1}^{n}(x_{jB})^2\right]$	$S_{A,B}=2c/[a+b]$
Tanimoto coefficient	$S_{A,B}=\frac{\left[\sum_{j=1}^{n}x_{jA}x_{jB}\right]}{\left[\sum_{j=1}^{n}(x_{jA})^2+\sum_{j=1}^{n}(x_{jB})^2-\sum_{j=1}^{n}x_{jA}x_{jB}\right]}$	$S_{A,B}=c/[a+b-c]$
Soergel distance[b]	$D_{A,B}=\left[\sum_{j=1}^{n}\lvert x_{jA}-x_{jB}\rvert\right]/\left[\sum_{j=1}^{n}max(x_{jA},x_{jB})\right]$	$D_{A,B}=1-\frac{c}{[a+b-c]}$

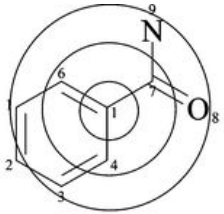
S denotes similarities, while *D* denotes distances. The two can be converted to each other by *similarity=1/(1+distance)*. x_{jA} means the j-th feature of molecule A. a is the number of *on* bits in molecule A, b is number of *on* bits in molecule B, while c is the number of bits that are *on* in both molecules.

(Left) Maggiora, Gerald, Martin Vogt, Dagmar Stumpfe, und Jürgen Bajorath. „Molecular Similarity in Medicinal Chemistry“. *Journal of Medicinal Chemistry* 57, Nr. 8 (24. April 2014): 3186–3204. (Right) Bajusz, Dávid, Anita Rácz, and Károly Héberger. 2015. “Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?” Journal of Cheminformatics 7 (1): 20.

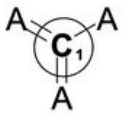
Extended-connectivity fingerprints (ECFPs) and Functional-class fingerprints (FCFPs) extract and compare (multi-)sets of subgraphs



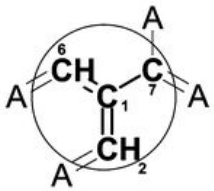
Benzamide



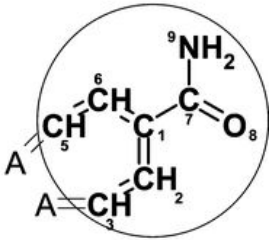
Considering atom 1 in benzoic acid amide



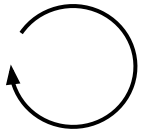
Iteration 0



Iteration 1



Iteration 2



hash(Daylight atomic invariants*)

*# non-H neighbors, bond order, atom number, atom mass, atom charge, attached Hs.

A bit array as output: *[0,0,0,0,0, …, 1, 0, 0,..., 1,0,...]*

Implemented in RDKit and other software. Publication and tutorials: (1) Rogers, David, and Mathew Hahn. "Extended-Connectivity Fingerprints." Journal of Chemical Information and Modeling (2010). (2) Tutorial by Manish Kumar and (3) Tutorial by Leo Klarner.