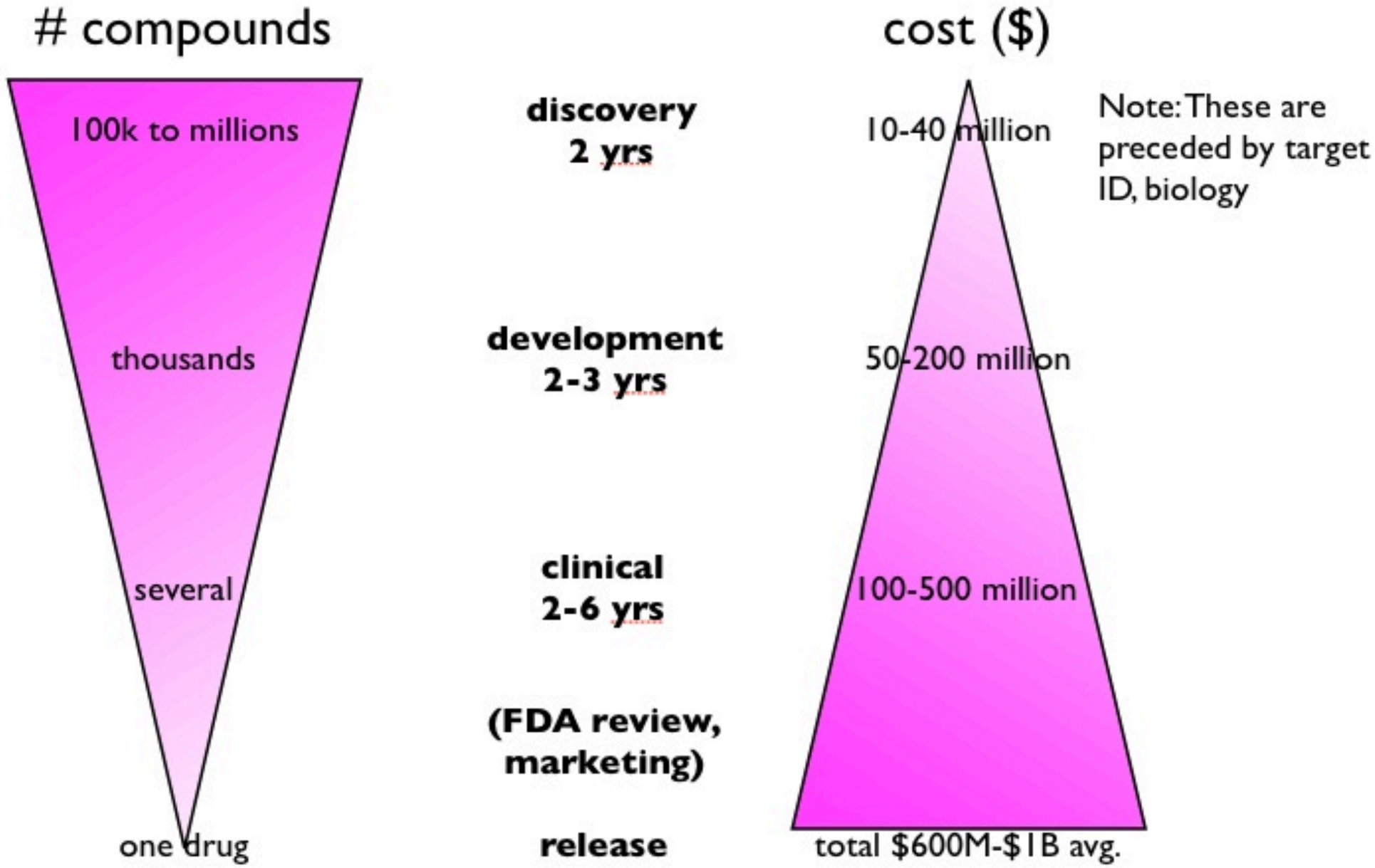
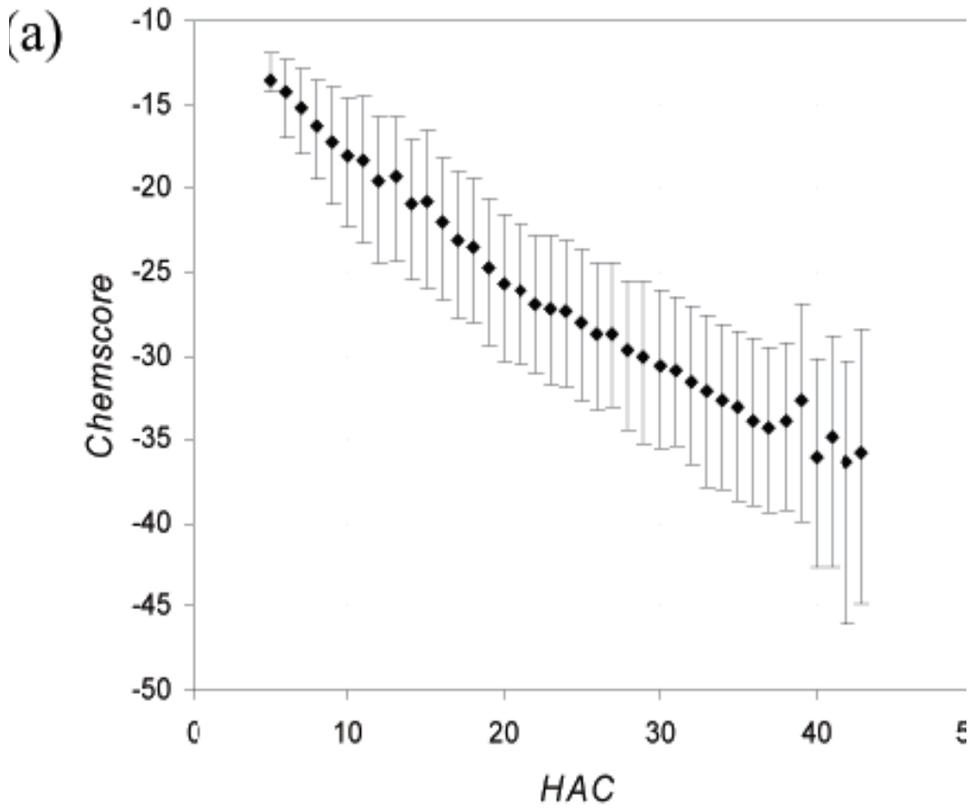
Drug discovery is a funneling process

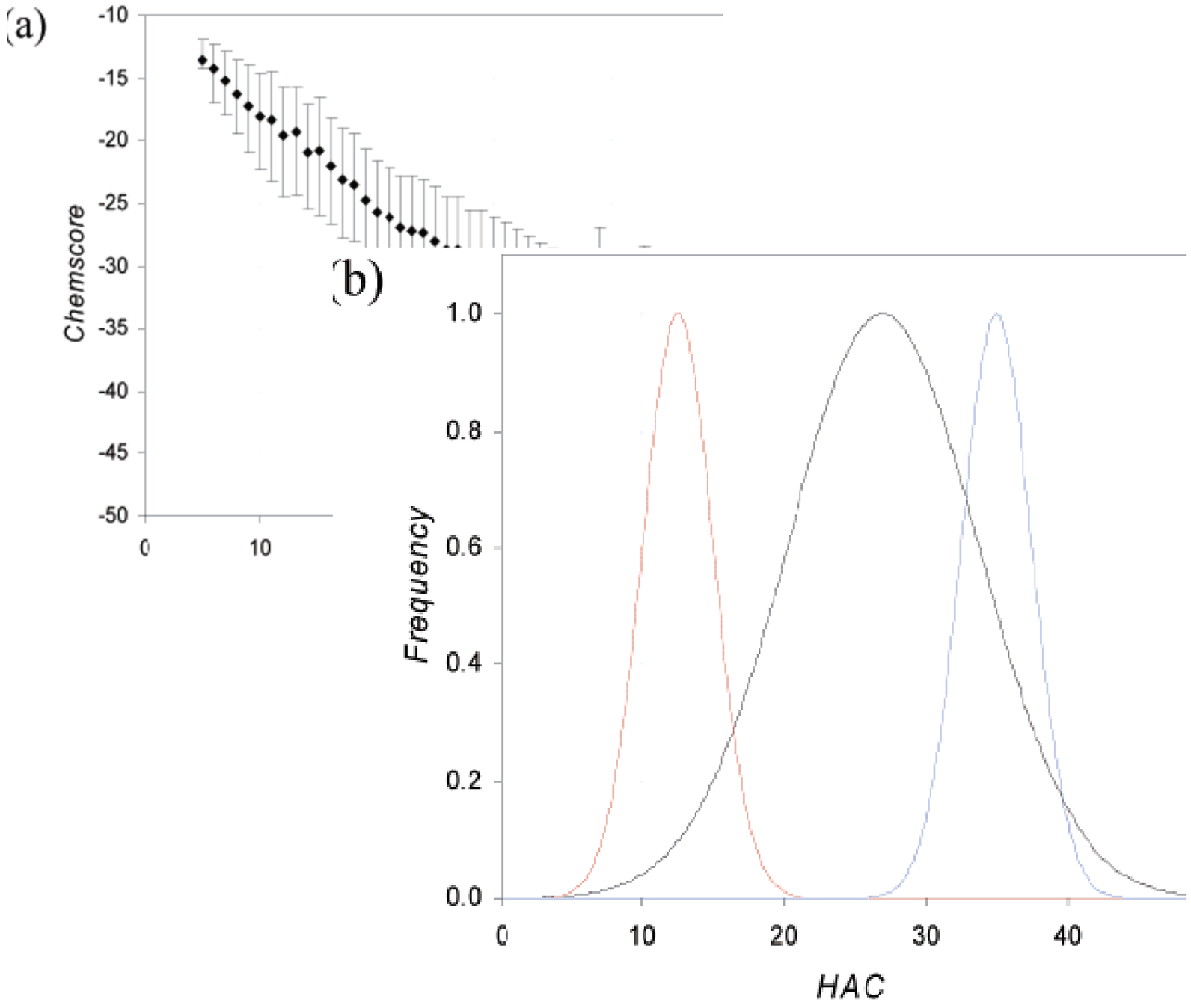


Watch out for accidental cheating



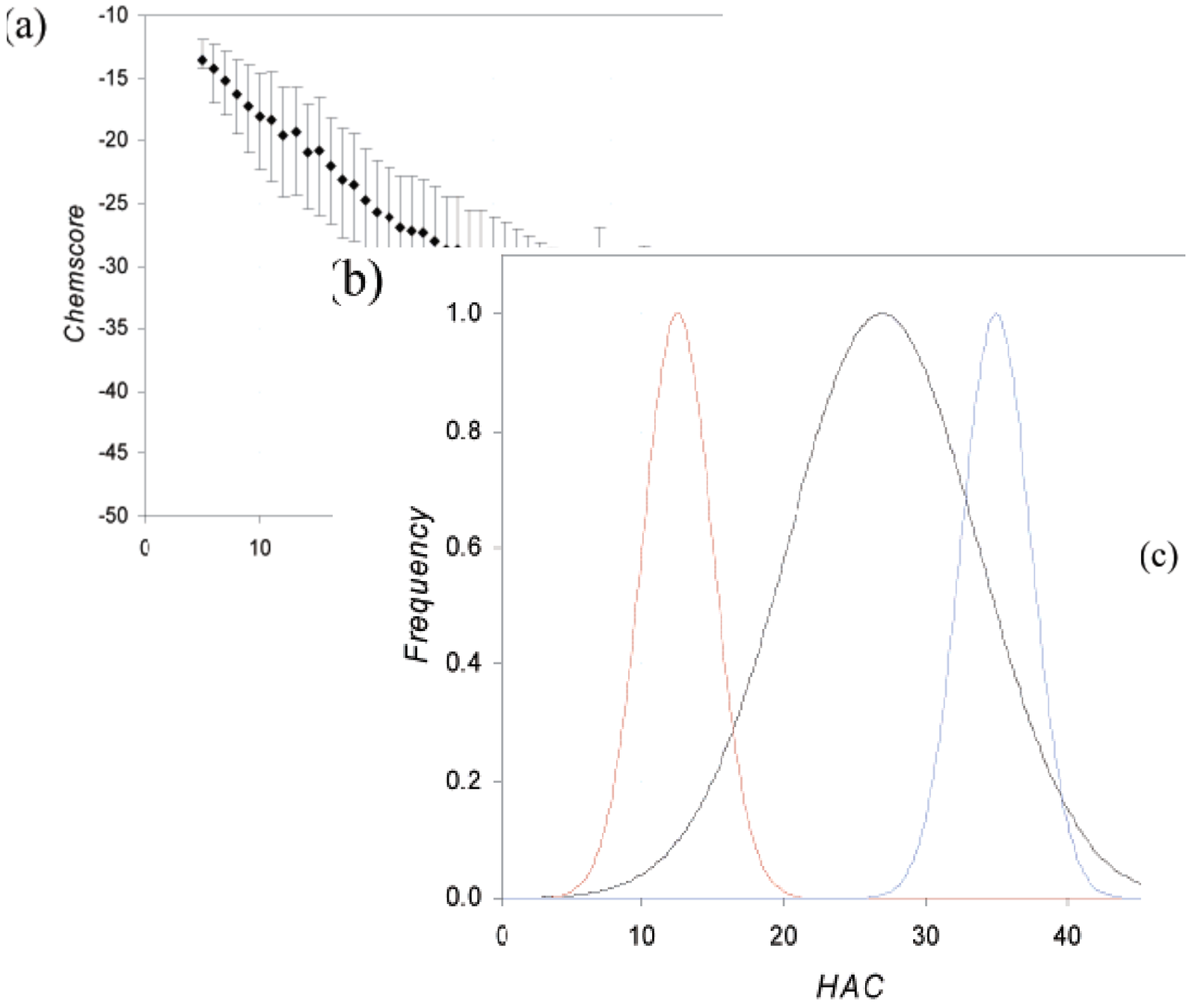
http://johnirwin.docking.org/talks/acs07-1.pdf

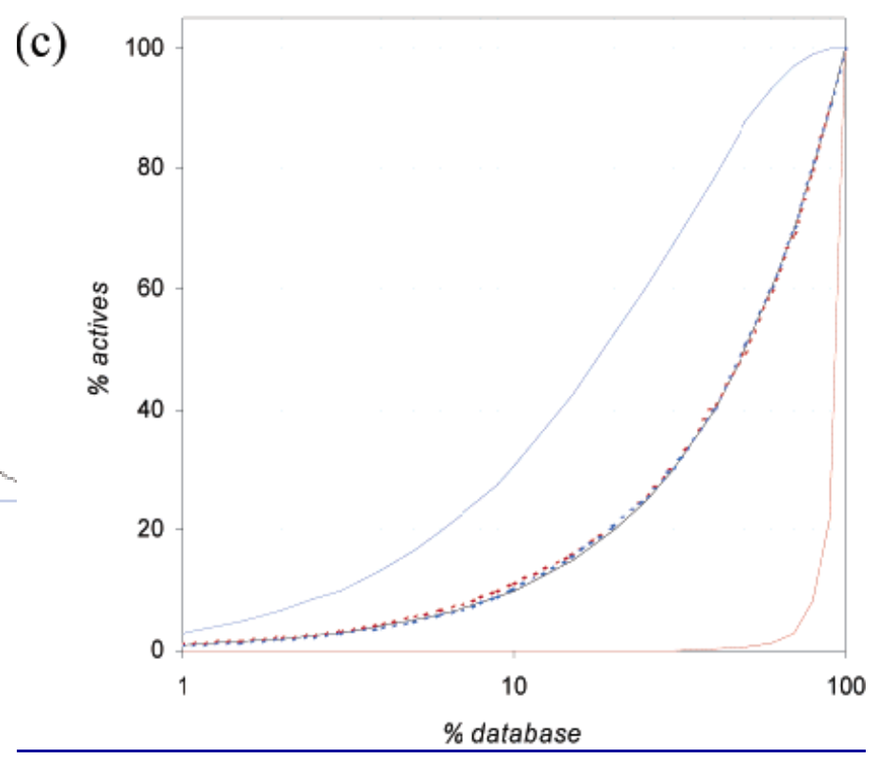
Watch out for accidental cheating



http://johnirwin.docking.org/talks/acs07-1.pdf

Watch out for accidental cheating





http://johnirwin.docking.org/talks/acs07-1.pdf

ZINC is not commercial

Name	Availability	Estimated Size (purchasable)
not-for-sale	All substances that cannot be bought as far as we know - tell us if we are mistaken	2100644 (0)
agent	All substances that are available in stock, via procurement agents	3010821 (3010821)
in-stock	Compounds purchased direct from manufacturer, already made, sitting on a shelf, ready to ship to you.	12084317 (12084317)
boutique	Boutique substances are generally much more expensive than $100/sample, often made to order, yet still cheaper than making it yourself.	12949291 (12949291)
now	Immediate delivery, includes in-stock and agent	21455156 (21455156)
bb	Available in preparative quantities, typically at least 250 mg	42908449 (42908449)
on-demand	All substances that are for sale	95008076 (95008076)
wait-ok	Compounds you can get in 8-10 weeks at modest prices, includes in-stock, agent and on-demand	116463232 (116463232)
for-sale	All substances that are for sale, includes in-stock, on-demand, and boutique	389000000 (389000000)

ZINC includes various subsets which are fairly typical for a large library

	Lead-Like	Fragment-Like	Drug-Like	All	Shards
Standard Size Updated	**Lead-Like** 6,687,370 2013-03-11	**Fragment-Like** 1,389,525 2013-10-25	**Drug-Like** 15,798,630 2013-02-08	**All Purchasable** 22,724,825 2013-12-18	**Shards** 85,247 2013-10-20
Clean Size Updated	**Clean Leads** 5,735,035 2013-11-05	**Clean Fragments** 148,310 2013-11-05	**Clean Drug-Like** 13,195,609 2013-11-05	**All Clean** 16,403,865 2013-12-18	**Clean Shards** 60,021 2013-11-05
In Stock Size Updated	**Leads Now** 2,419,472 2013-11-01	**Frags Now** 527,585 2013-10-25	**Drugs Now** 7,397,957 2013-11-11	**All Now** 9,046,036 2013-04-04	**Shards Now** 63,861 2013-10-20
Boutique Size Updated	**Boutique Leads** 5,114,169 2012-12-24	**Boutique Frags** 2,755,555 2013-11-08	**Boutique Drugs** 10,292,210 2012-11-27	**All Boutique** 12,217,845 2012-11-27	**Boutique Shards** 80,698 2013-11-08

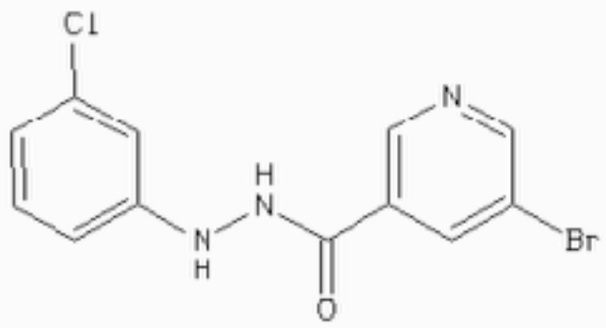
ZINC also sorts available compounds based on important biological characteristics

LogP (up to)	Molecular Weight (up to, Daltons) 200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	15,797	69,842	210,532	152,529	154,671	132,721	105,186	39,485	27,456	5,248	6,725	920,192
0	116,185	695,905	2,031,591	1,443,658	1,434,388	1,204,003	927,872	371,656	252,203	32,927	3,423	8,513,811
1	360,619	2,633,294	9,349,483	7,389,851	6,949,968	6,691,010	5,092,509	2,155,735	1,500,495	185,414	6,804	42,515,182
2	508,373	5,204,373	21,593,330	17,823,695	20,096,138	20,748,184	16,218,929	7,310,138	5,123,912	747,988	19,009	115,394,069
2.5	191,473	2,608,710	13,517,643	14,354,389	15,419,065	16,776,890	13,789,173	6,632,804	4,871,673	792,798	21,749	88,976,367
3	104,249	1,979,155	12,449,795	14,661,968	17,102,322	19,167,869	16,603,985	8,669,477	6,594,633	1,174,958	37,236	98,545,647
3.5	40,988	1,157,521	9,213,359	12,159,514	15,653,974	18,404,615	17,055,959	9,991,491	7,989,194	1,561,405	61,962	93,289,982
4	8,095	387,084	4,856,044	7,620,522	11,219,384	14,181,322	14,459,614	9,738,632	8,070,346	1,812,081	94,453	72,447,577
4.5	774	49,829	1,522,372	3,248,990	6,067,205	8,279,972	9,419,219	7,617,060	6,644,297	1,787,576	132,701	44,769,995
5	95	3,906	224,652	787,992	1,975,102	3,357,924	4,461,087	4,381,427	4,178,626	1,431,929	161,241	20,963,981
>5	28	803	21,635	106,437	369,687	952,950	1,667,709	2,197,171	2,528,737	1,531,266	809,078	10,205,701
Totals, by Weight	1,346,676	14,990,422	74,990,436	79,749,545	96,462,104	109,897,460	99,801,242	59,105,076	47,781,572	11,063,590	1,354,381	**597M** Substances **1.9K** Tranches

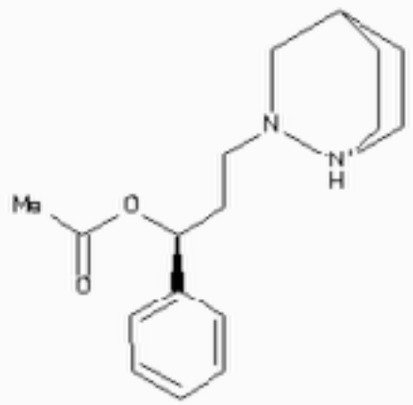
http://zinc15.docking.org/tranches/home/

Let's look at samples from the "in stock" set: Leads now

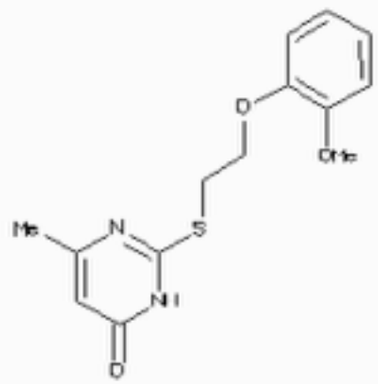
1.
20031600



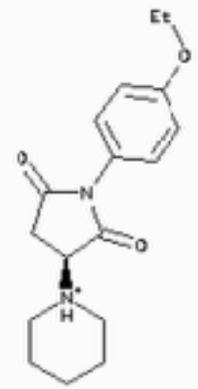
2.
19166762



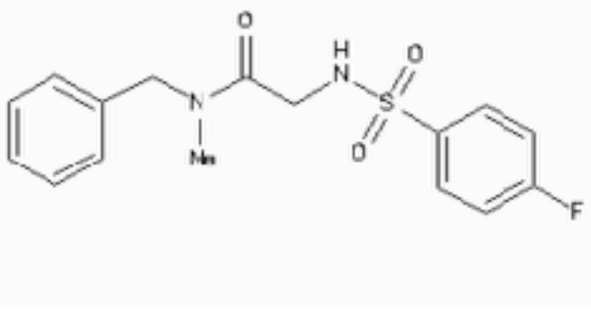
3.
3901268



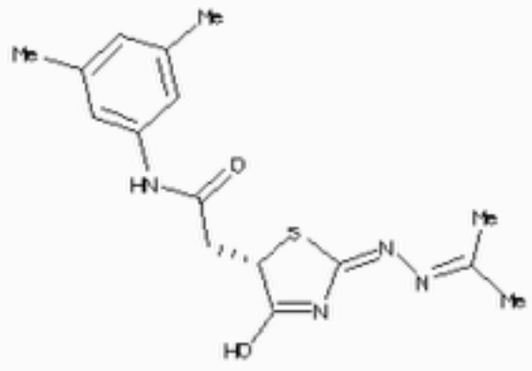
4.
19799526



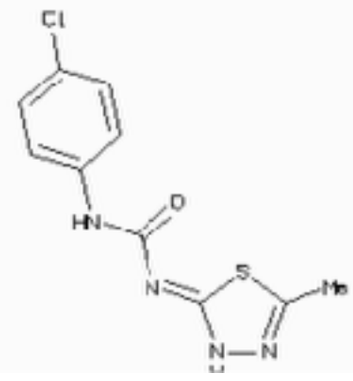
5.
982952



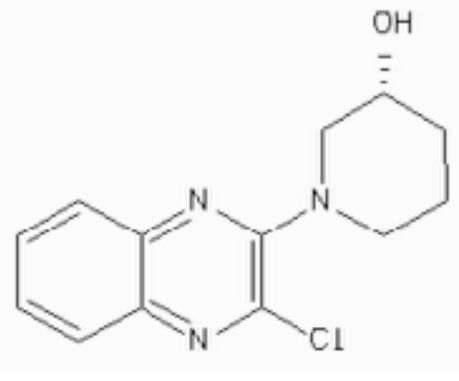
6.
8575396



7.
1240782

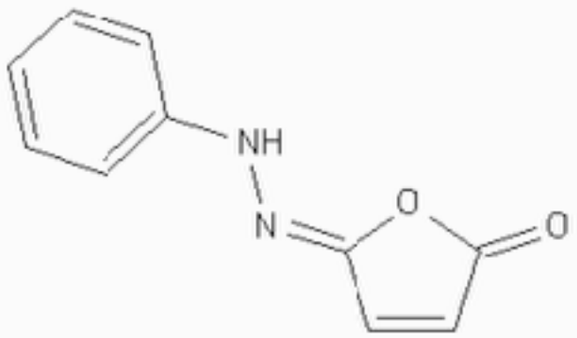


8.
984053

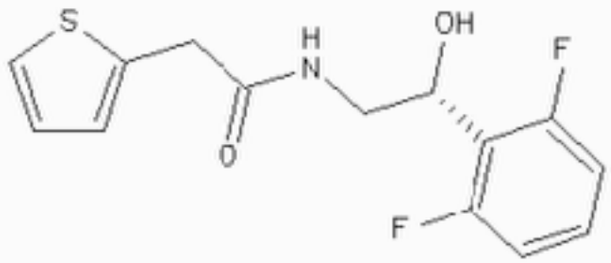


Let's look at samples from the "in stock" set: Fragments now

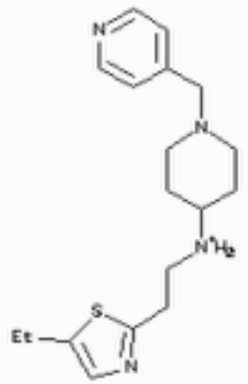
1.
95917816



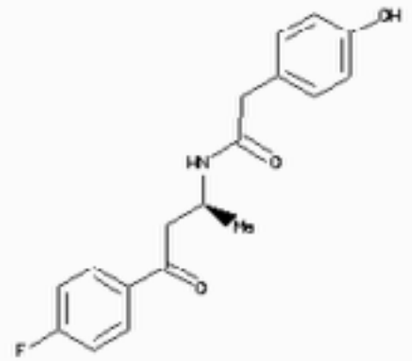
2.
95966654



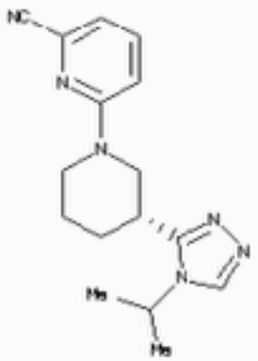
3.
95968696



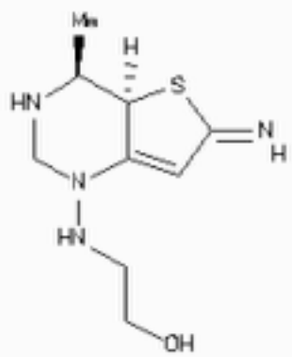
4.
95980970



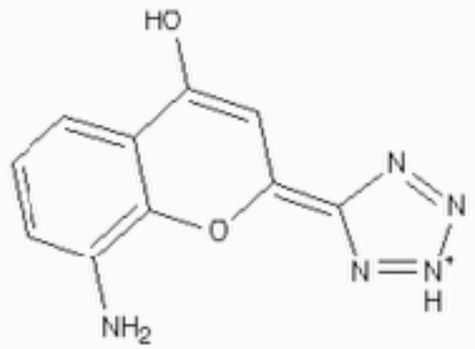
5.
95975039



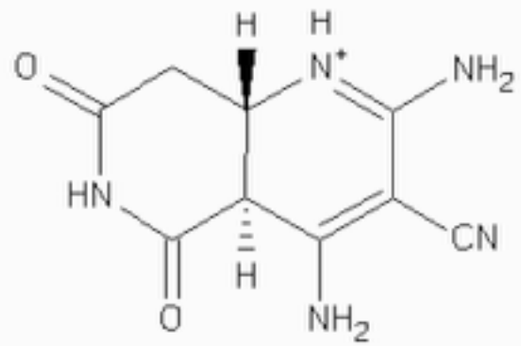
6.
95923987



7.
95929104

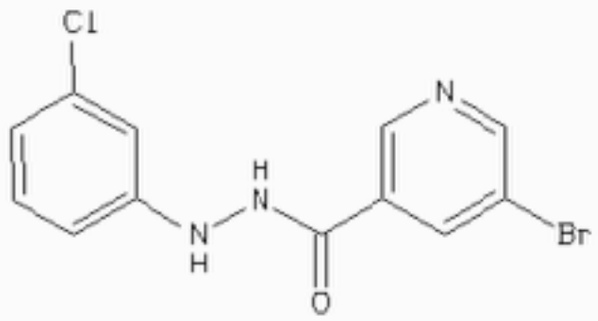


8.
95922398

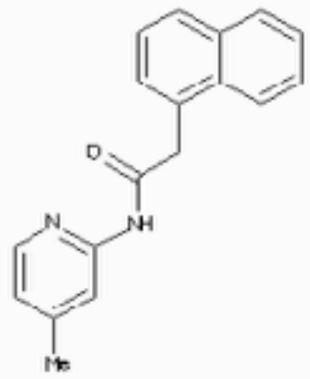


Let's look at samples from the "in stock" set: "Drugs" now

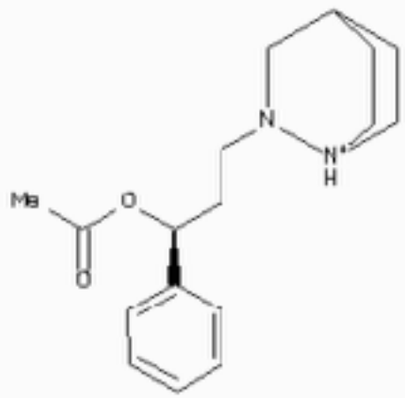
1.
20031600



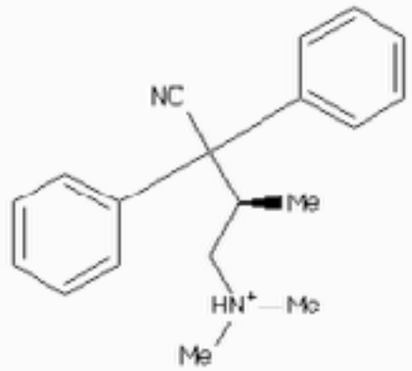
2.
9365179



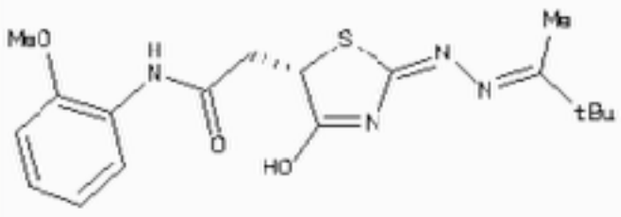
3.
19166762



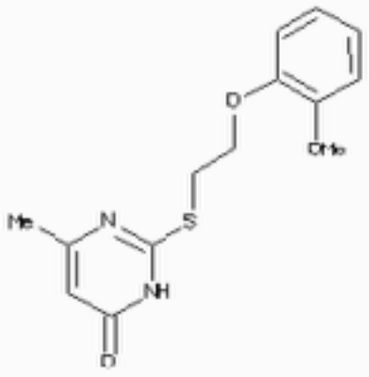
4.
1700294



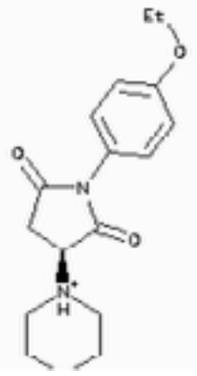
5.
12378847



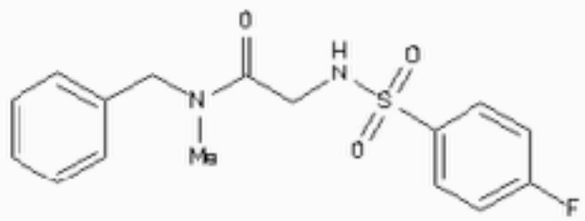
6.
3901268



7.
19799526

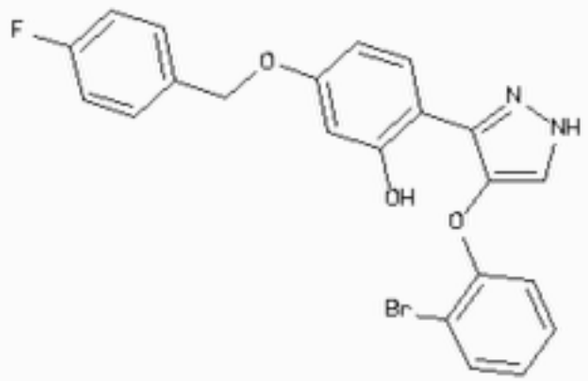


8.
982962

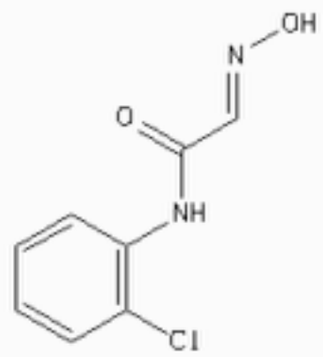


Let's look at samples from the "in stock" set: Shards now

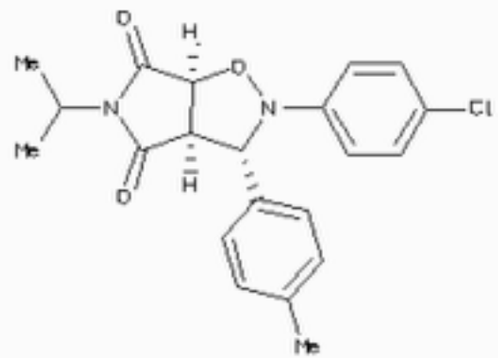
17.
95908809



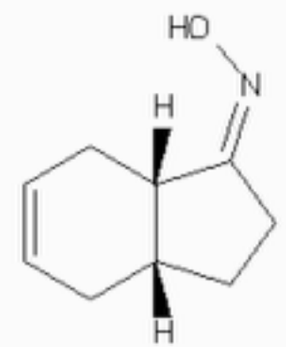
18.
95920159



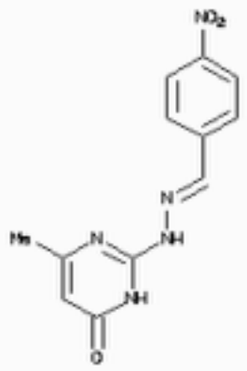
19.
95909119



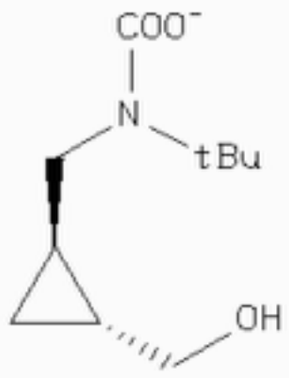
20.
95922258



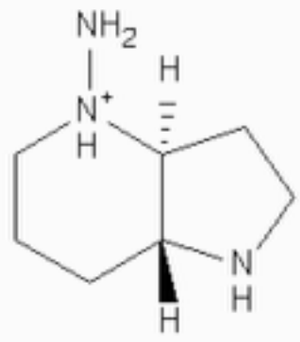
21.
8742072



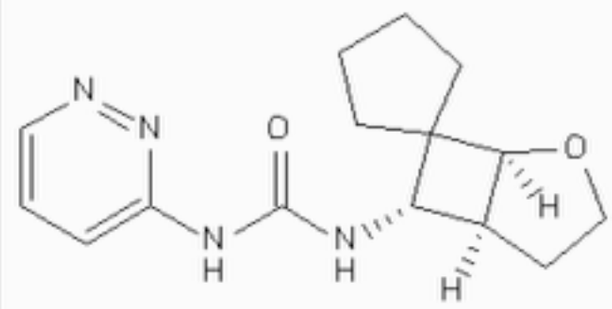
22.
95923509



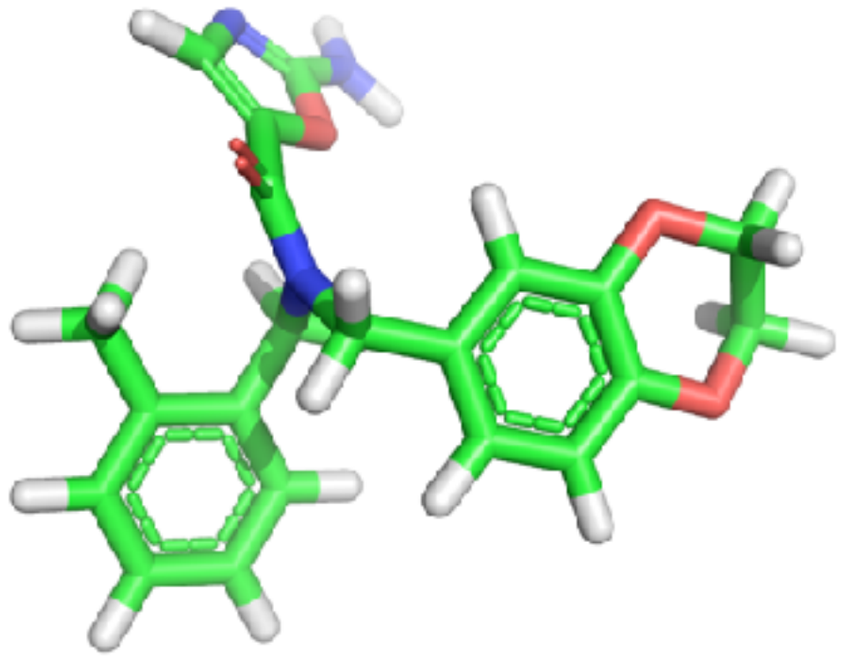
23.
95922348



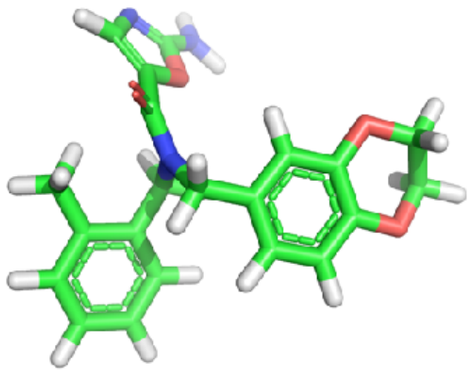
24.
95974365



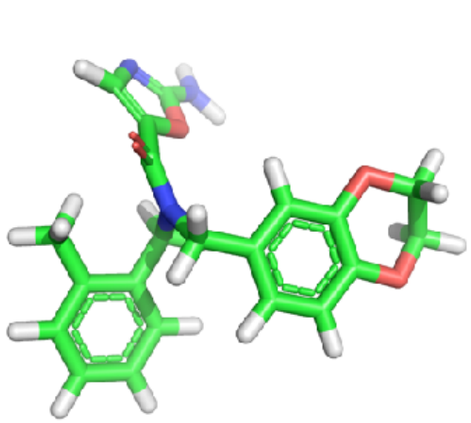
LINGO searches are extremely fast and make some chemical sense

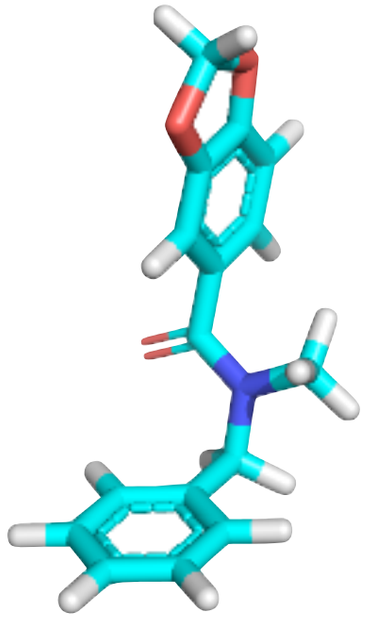


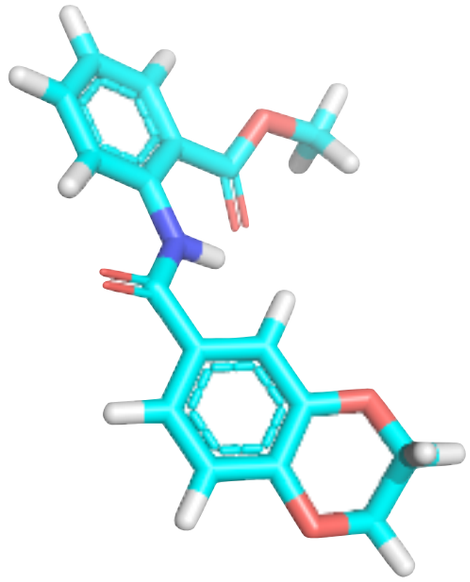
LINGO searches are extremely fast and make some chemical sense

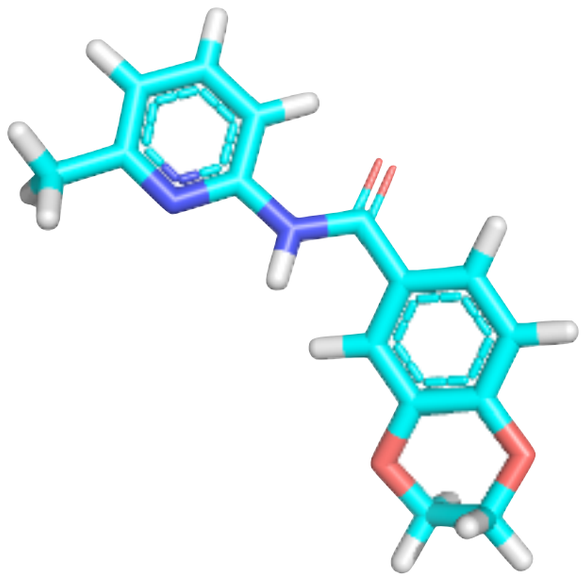


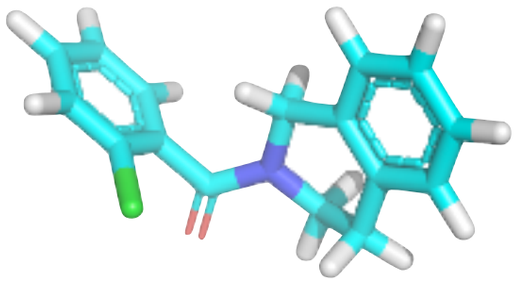
LINGO searches are extremely fast and make some chemical sense



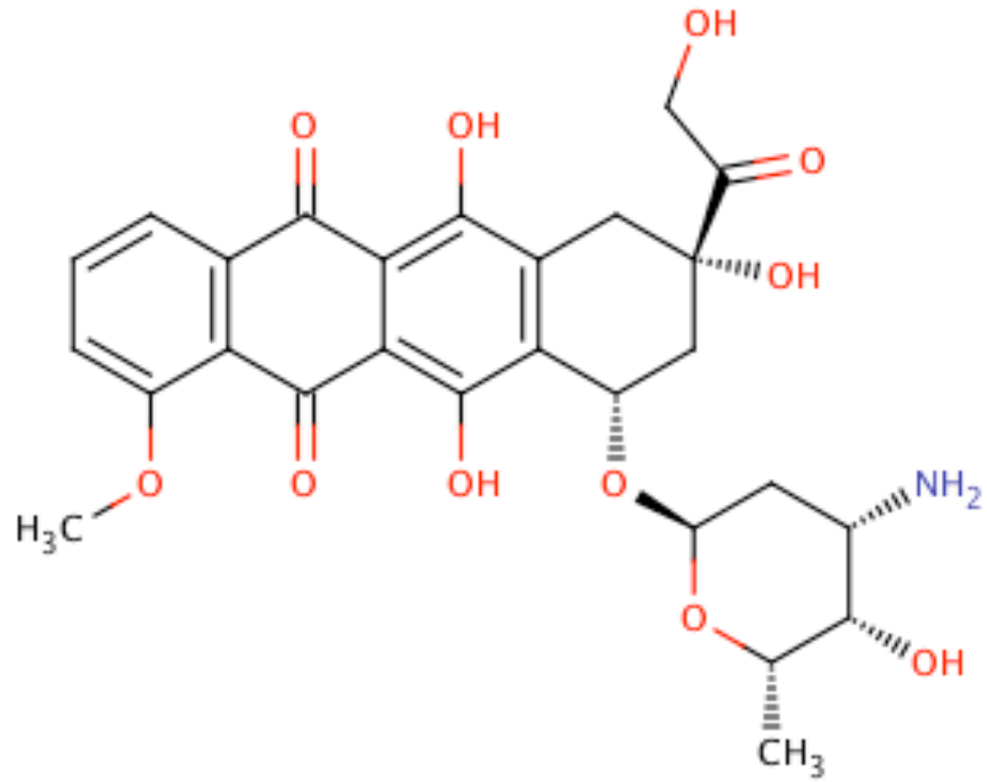








LINGO searches work based on SMILES strings



CC1C(C(CC(O1)OC2CC(Cc3c2c(c4c(c3O)C(=O)c5ccc
c(c5C4=O)OC)O)(C(=O)CO)O)N)O

1D representations include name, SMILES strings

- "1D": Conveys identity of compound in a way that can in principle be converted to 2D structure
- Most obvious 1D representation: Chemical name (i.e. IUPAC name)
- Numbering schemes
    - Chemical Abstracts Service (CAS) numbers
    - PubChem numbers
    - Like social security numbers
- Smiles strings

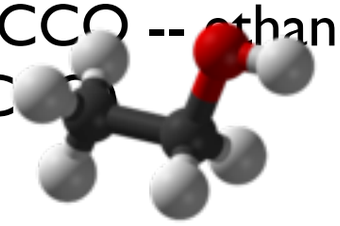
SMILES strings are a powerful, informative way to represent molecules

- "Simplified molecular-input line-entry specification"
- Simpler than reading/writing IUPAC names
- Human readable
- Element names and types (hydrogens implied)
  - C, B, N, O, P, S, F, Cl, Br, I. Other elements must be in brackets. Lowercase for aromatic
  - Hydrogens are assumed (except in brackets)
  - Charges need to be indicated
  - Bonding shown by =, # (single bonds implied)
  - Ex. CCO -- ethanol; C=CO -- vinyl alcohol (ethenol) which tautomerizes to acetaldehyde (CC=O)

Images from Wikipedia

SMILES strings are a powerful, informative way to represent molecules

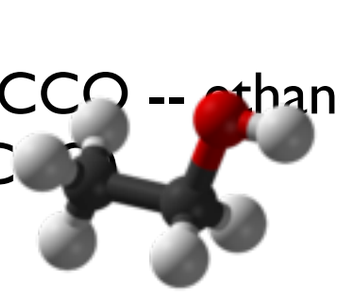
- "Simplified molecular-input line-entry specification"
- Simpler than reading/writing IUPAC names
- Human readable
- Element names and types (hydrogens implied)
  - C, B, N, O, P, S, F, Cl, Br, I. Other elements must be in brackets. Lowercase for aromatic
  - Hydrogens are assumed (except in brackets)
  - Charges need to be indicated
  - Bonding shown by =, # (single bonds implied)
  - Ex. CCO -- ethanol; C=CO -- vinyl alcohol (ethenol) which tautomerizes to acetaldehyde (CC

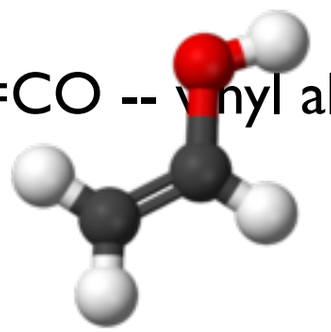


Images from Wikipedia

SMILES strings are a powerful, informative way to represent molecules

- "Simplified molecular-input line-entry specification"
- Simpler than reading/writing IUPAC names
- Human readable
- Element names and types (hydrogens implied)
  - C, B, N, O, P, S, F, Cl, Br, I. Other elements must be in brackets. Lowercase for aromatic
  - Hydrogens are assumed (except in brackets)
  - Charges need to be indicated
  - Bonding shown by =, # (single bonds implied)
  - Ex. CCO -- ethanol; C=CO -- vinyl alcohol (ethenol) which tautomerizes to acetaldehyde (CC

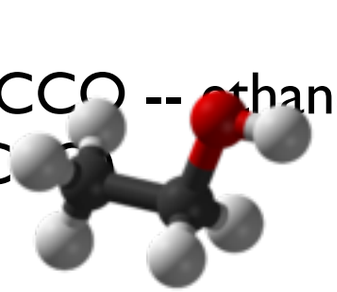


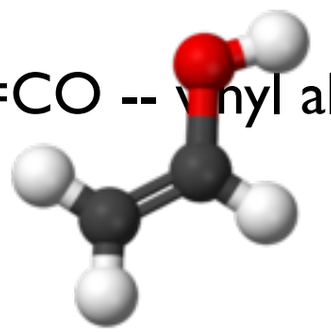


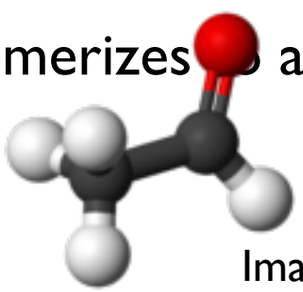
Images from Wikipedia

SMILES strings are a powerful, informative way to represent molecules

- "Simplified molecular-input line-entry specification"
- Simpler than reading/writing IUPAC names
- Human readable
- Element names and types (hydrogens implied)
  - C, B, N, O, P, S, F, Cl, Br, I. Other elements must be in brackets. Lowercase for aromatic
  - Hydrogens are assumed (except in brackets)
  - Charges need to be indicated
  - Bonding shown by =, # (single bonds implied)
  - Ex. CCO -- ethanol; C=CO -- vinyl alcohol (ethenol) which tautomerizes to acetaldehyde (CC=O)



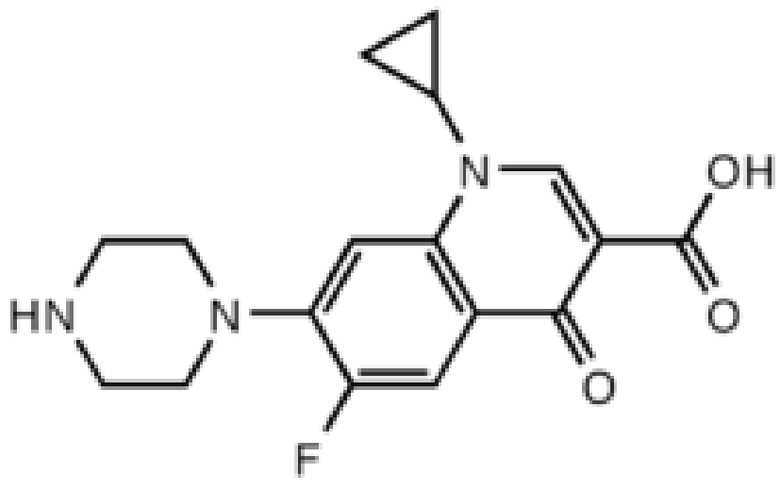




Images from Wikipedia

SMILES strings are a powerful, informative way to represent molecules

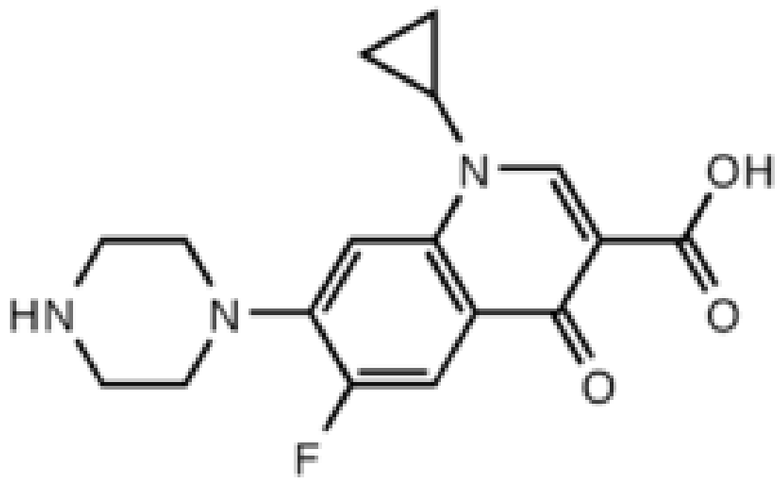
- Branching is indicated using parentheses
- Loop closure by numerically labeling atoms
  - Cyclohexane C1CCCCC1; dioxane O1CCCOCC1
- Generation is done by breaking cycles and writing as branches off a main backbone
- Ex:

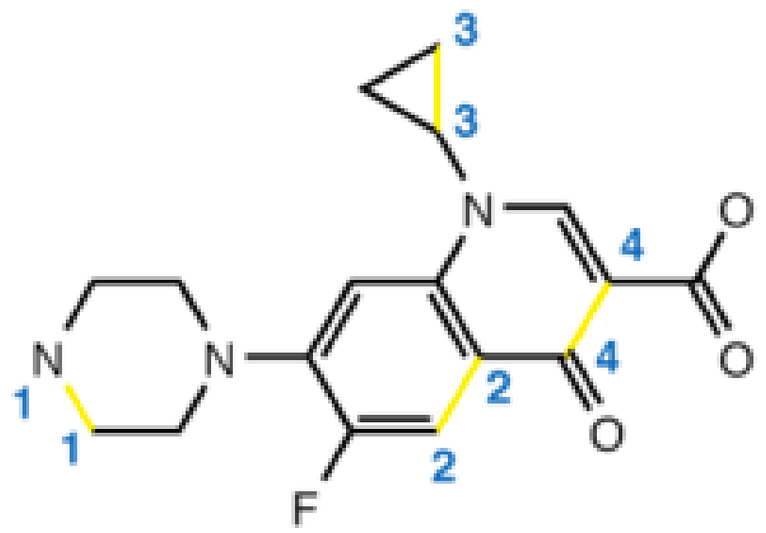


Some material, images from Wikipedia

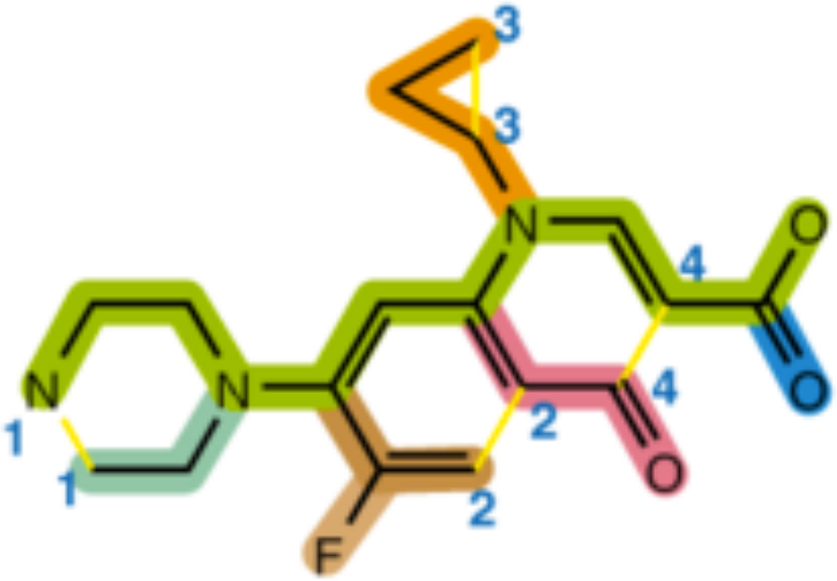
Example: Ciprofloxacin, an antibiotic

Break rings and identify backbone and places to close rings:





Write SMILES string



N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Some material, images from Wikipedia

Example: Ciprofloxacin, an antibiotic

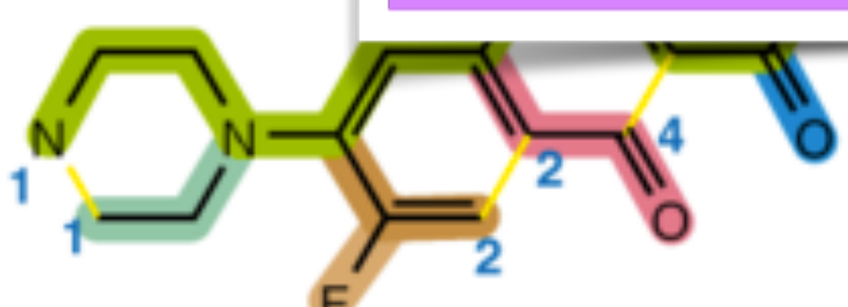
Break rings and identify backbone and places to close rings:

Much easier than writing the IUPAC name! And, easier to turn SMILES string into structure than doing so from the IUPAC name:

1-cyclopropyl-6-fluoro-4-oxo-7-piperazin-1-yl-quinoline-3-carboxylic acid

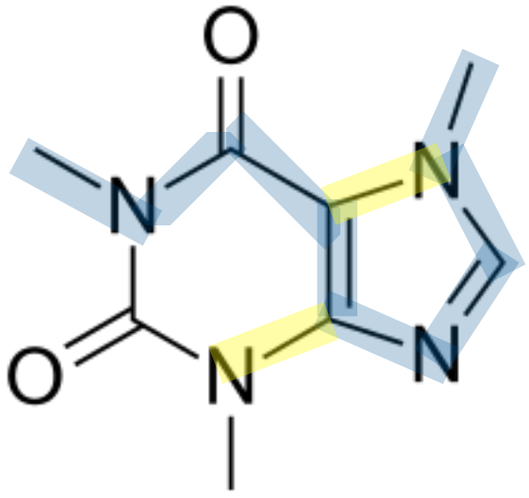
Write SM

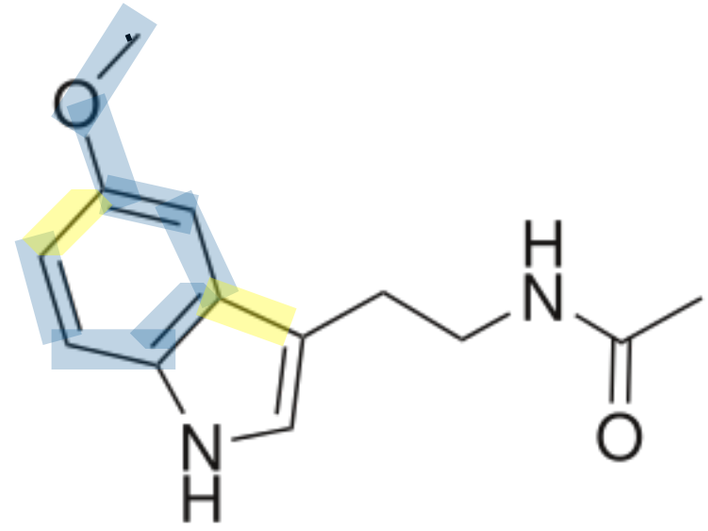
3CC3)C=C4C(=O)O



Some material, images from Wikipedia

Let’s try a couple examples together

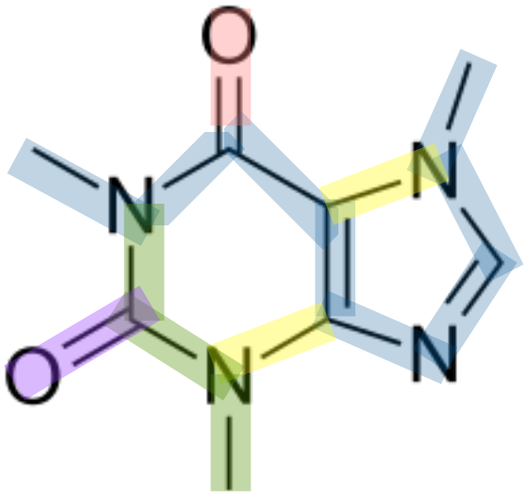


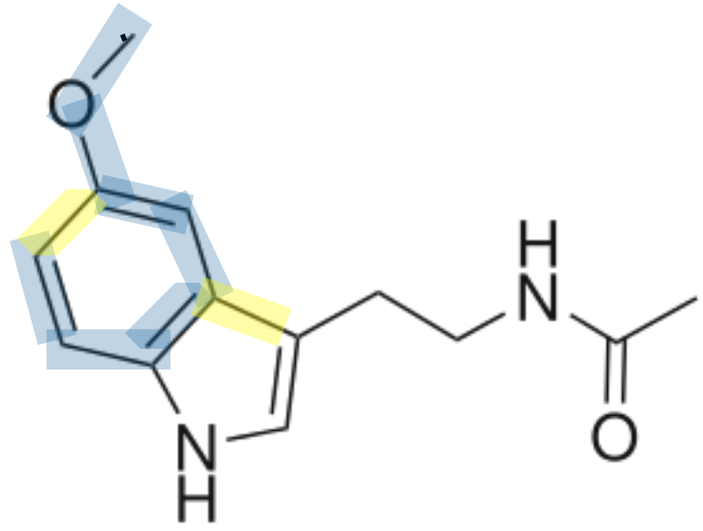


(remember: hydrogens implied)

Attrb: creative commons

Let’s try a couple examples together



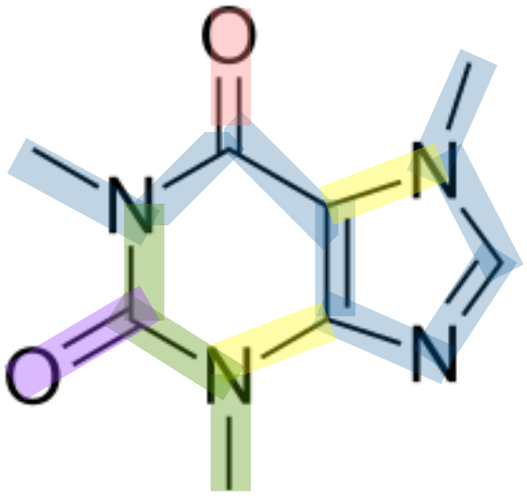


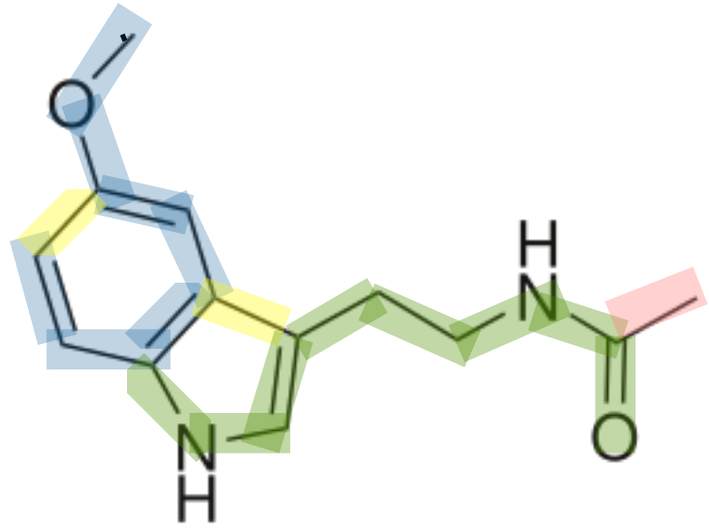
(remember: hydrogens implied)

Attrb: creative commons

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Let’s try a couple examples together





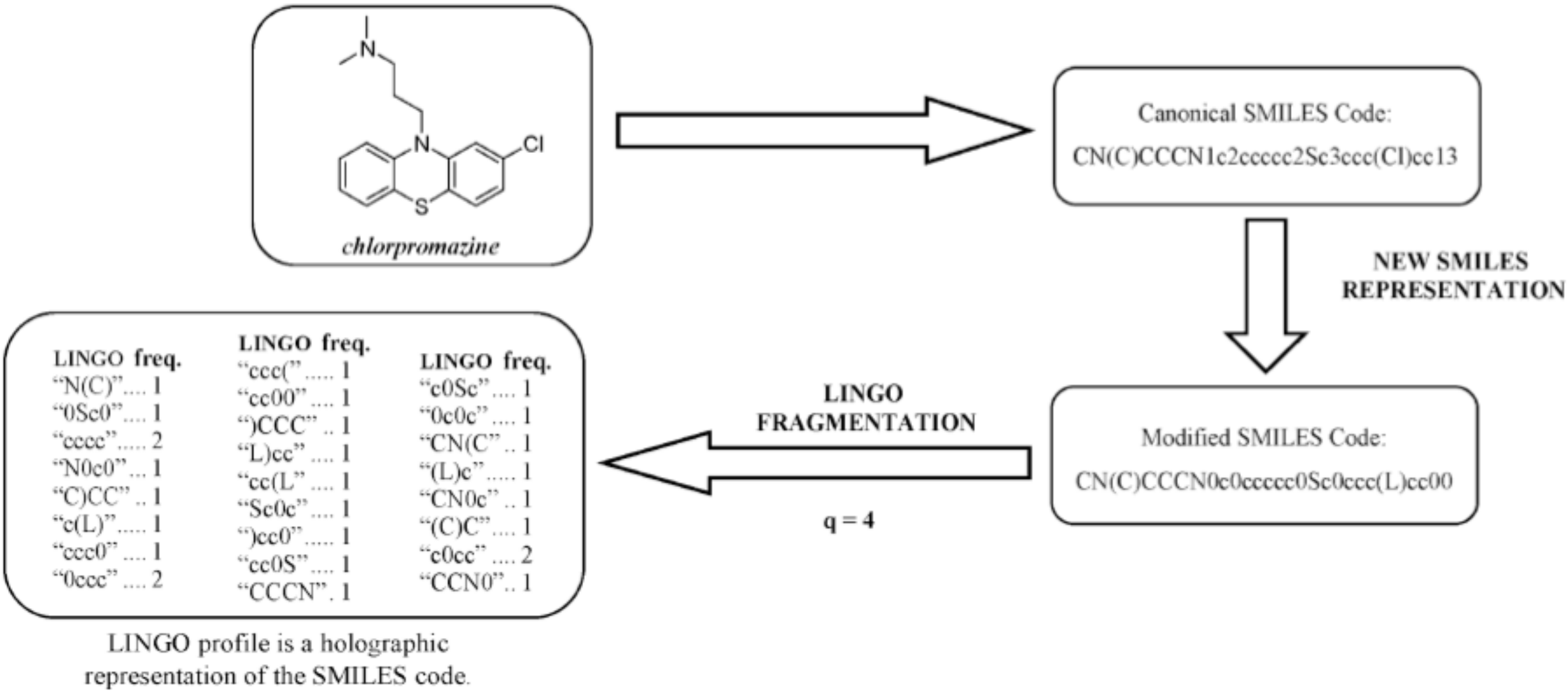
(remember: hydrogens implied)

Attrb: creative commons

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

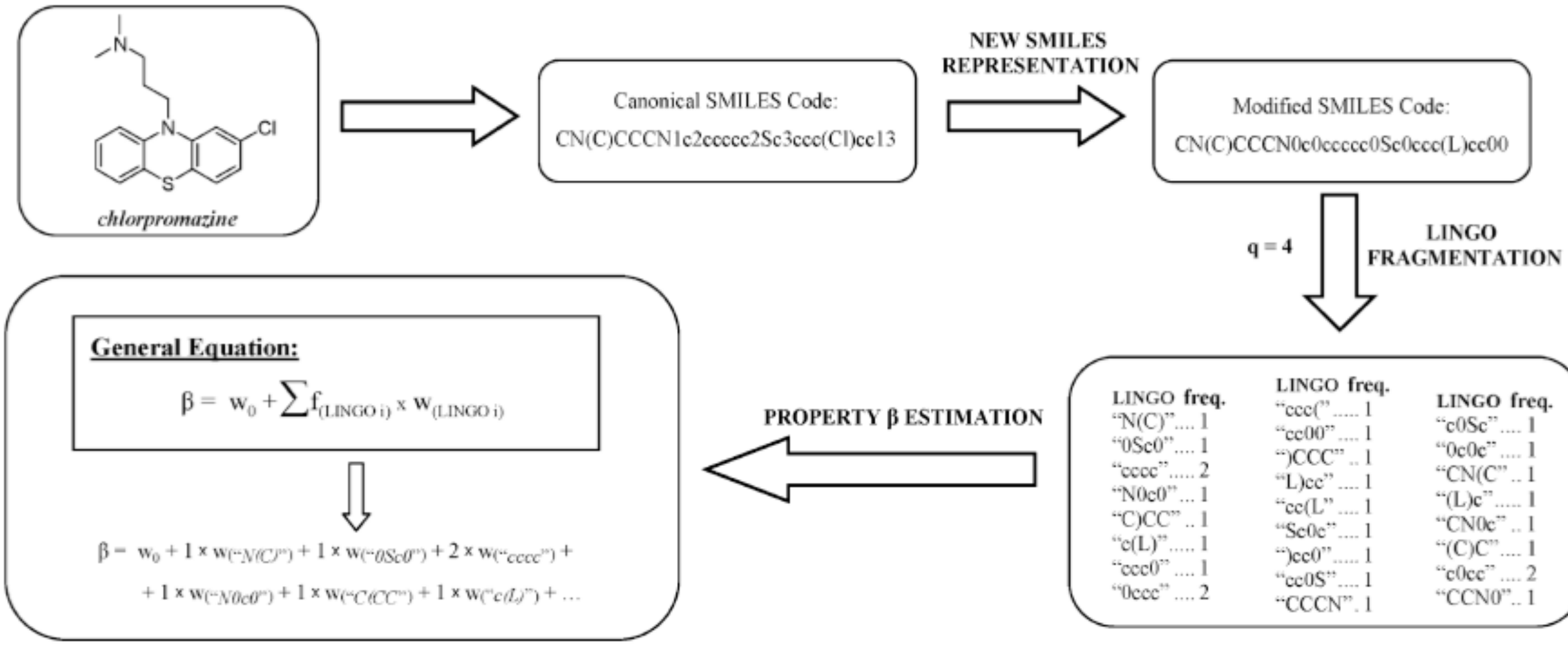
COC1=CC2=C(NC=C2CCNC(C)=O)C=C1

SMILES strings are modified to remove numbering and then fragmented; frequencies are compared

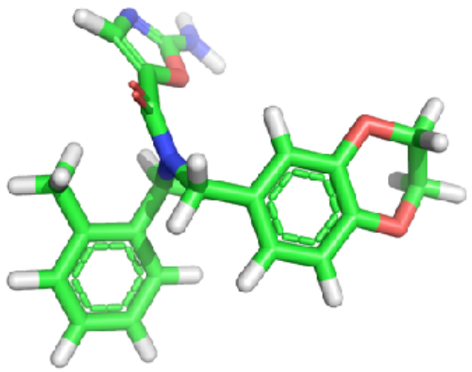


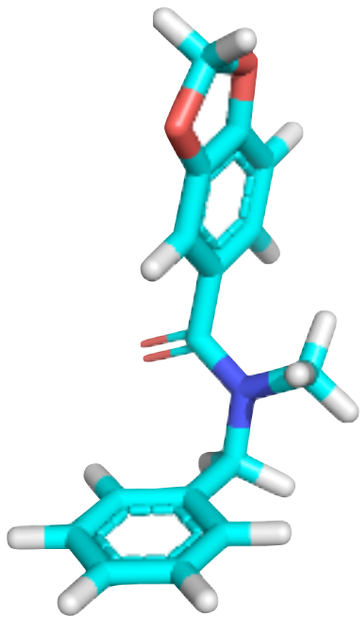
LINGO profile is a holographic representation of the SMILES code.

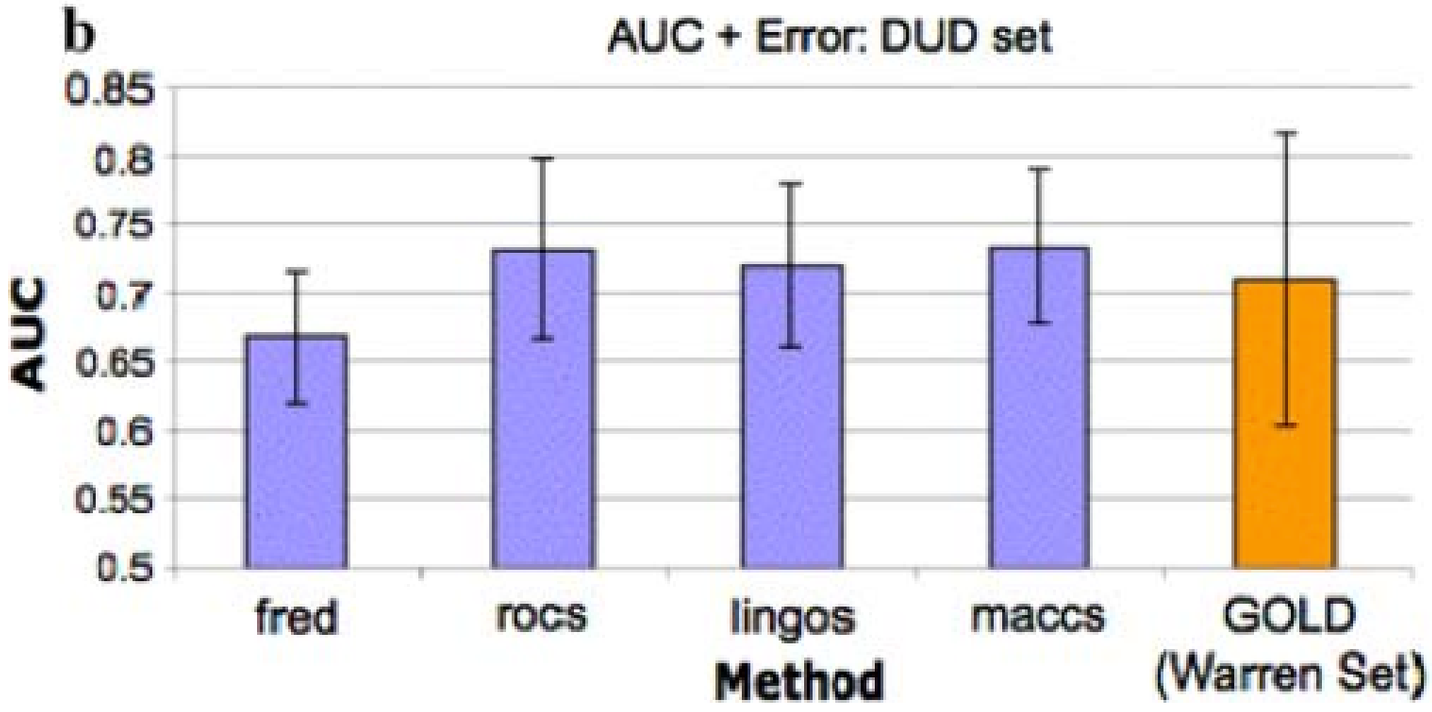
Some empirical models use LINGO to estimate properties based on functional groups



In virtual screening, LINGO searches compare reasonably favorably with shape and docking approaches

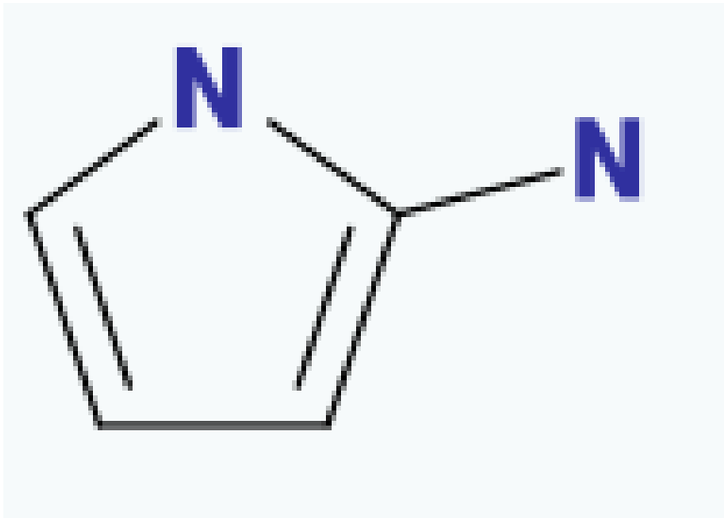


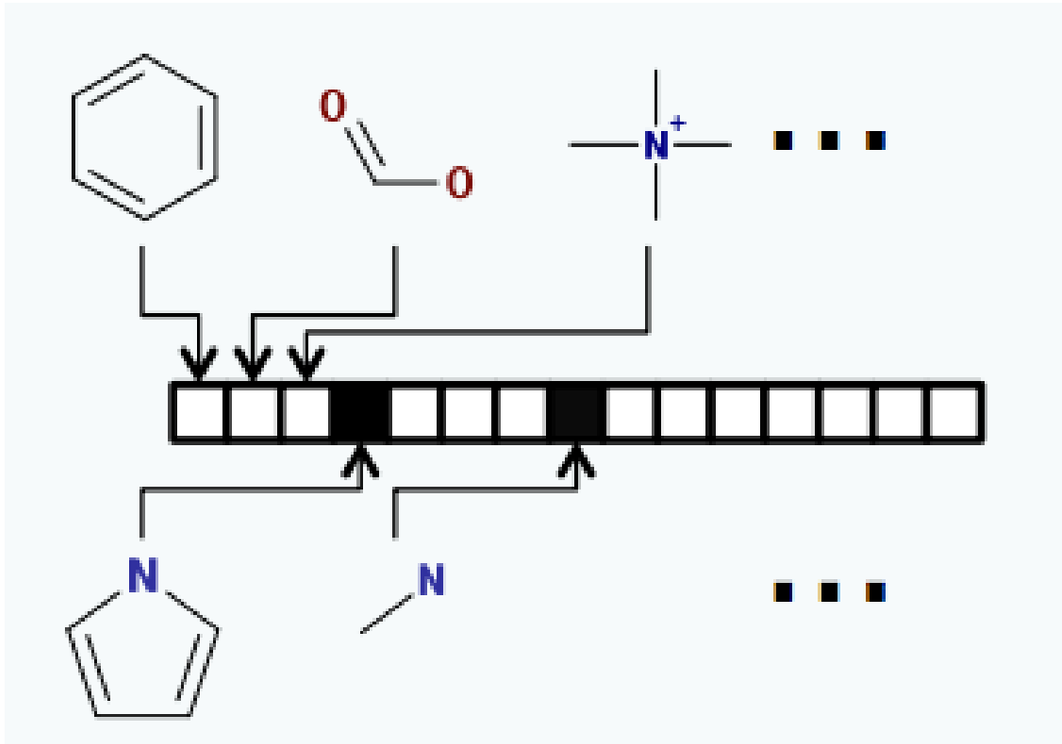




Nicholls, J. Comput. Aided Mol. Design 22:239 (2008)

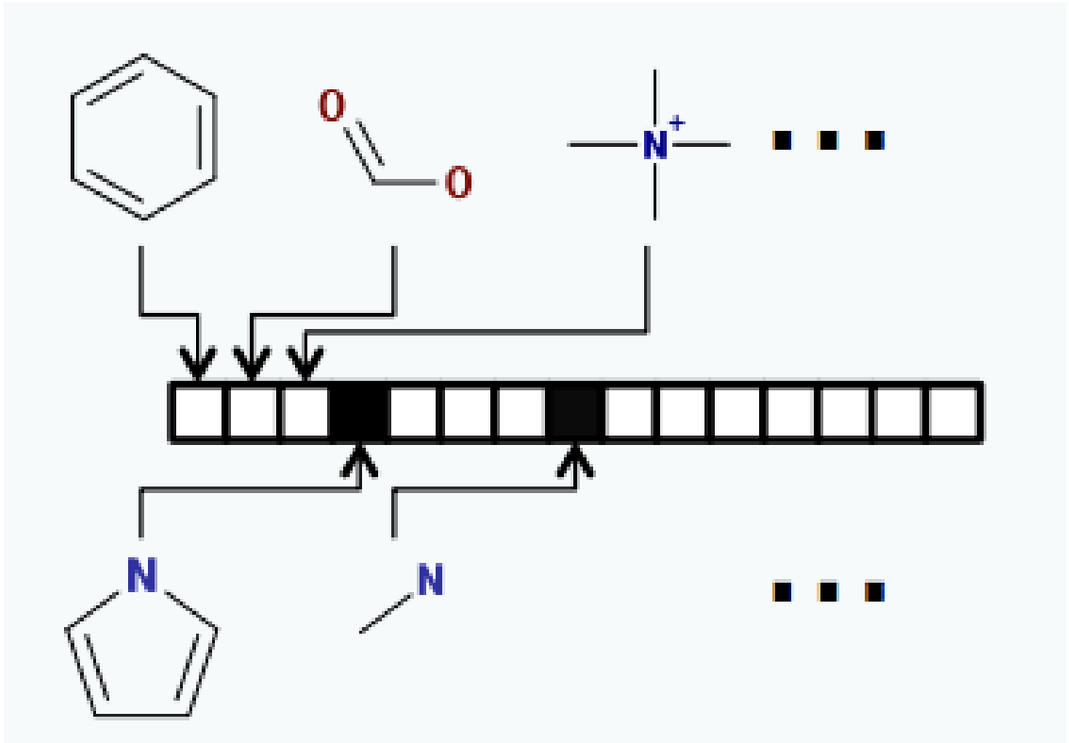
Fingerprint methods encode molecular descriptors in a fast way, often binary





http://www.eyesopen.com/2010_cup_presentations/GraphSim_presentation.pdf

What kind of information might we encode?



- Specific functional groups of interest
- Other patterns, bioisosteres
- Hydrogen bond donors, acceptors
- Aromatic rings generally
- Arrangement of functional groups
- Anything we want that can be framed in terms of a yes/no question