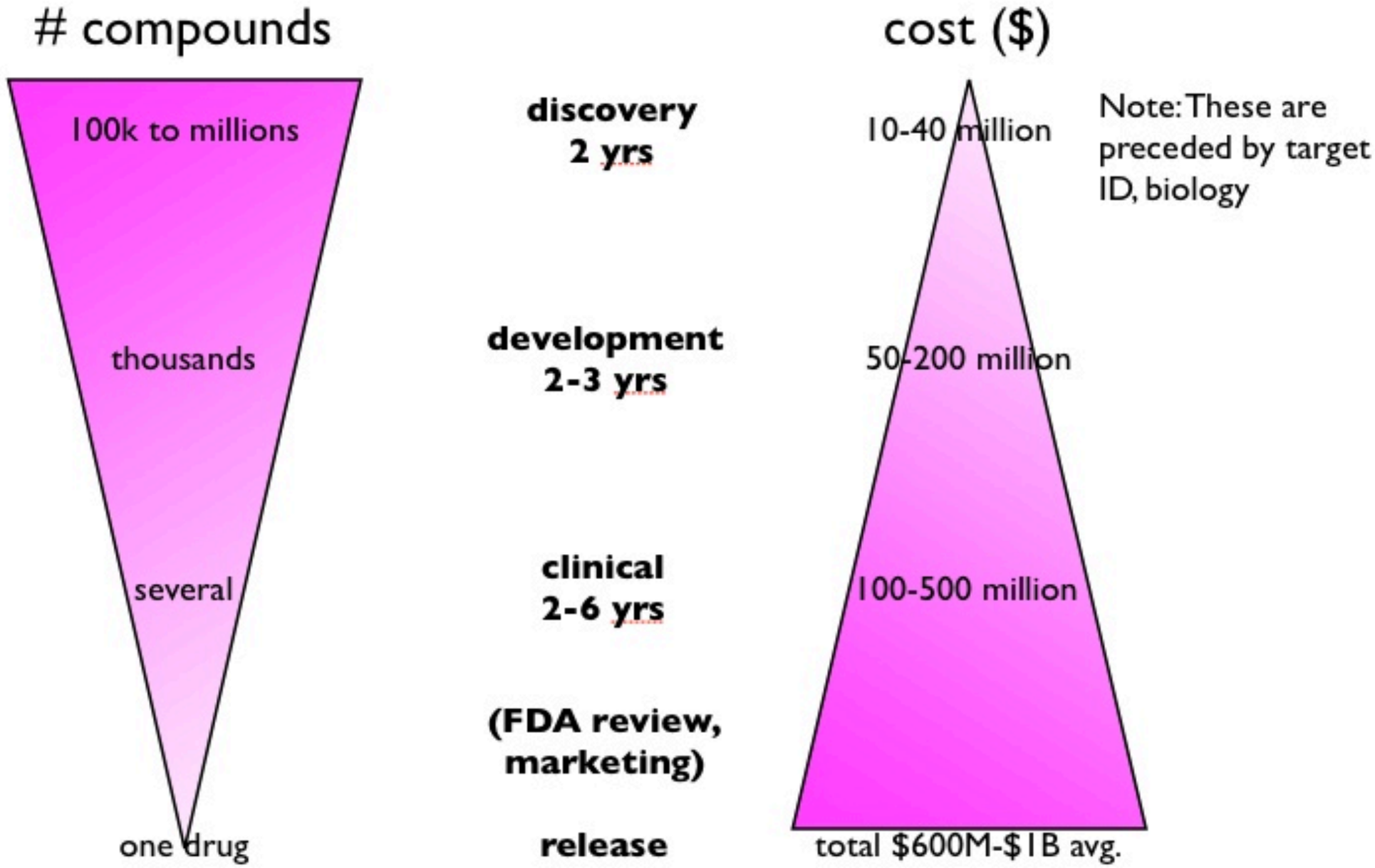
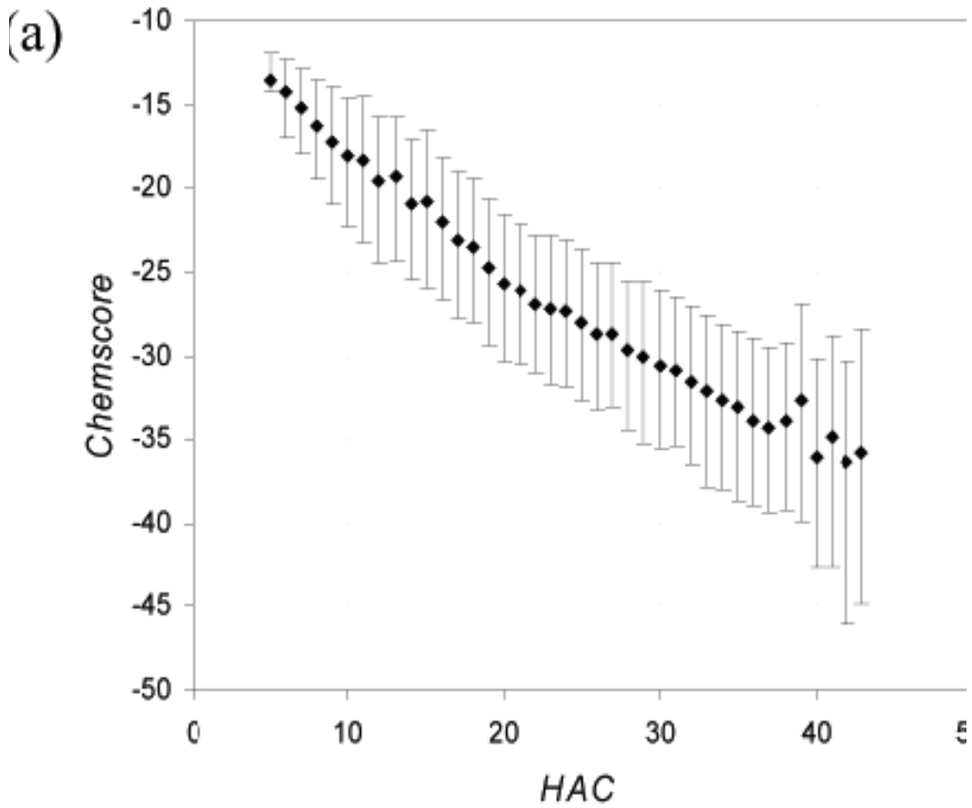
Drug discovery is a funneling process

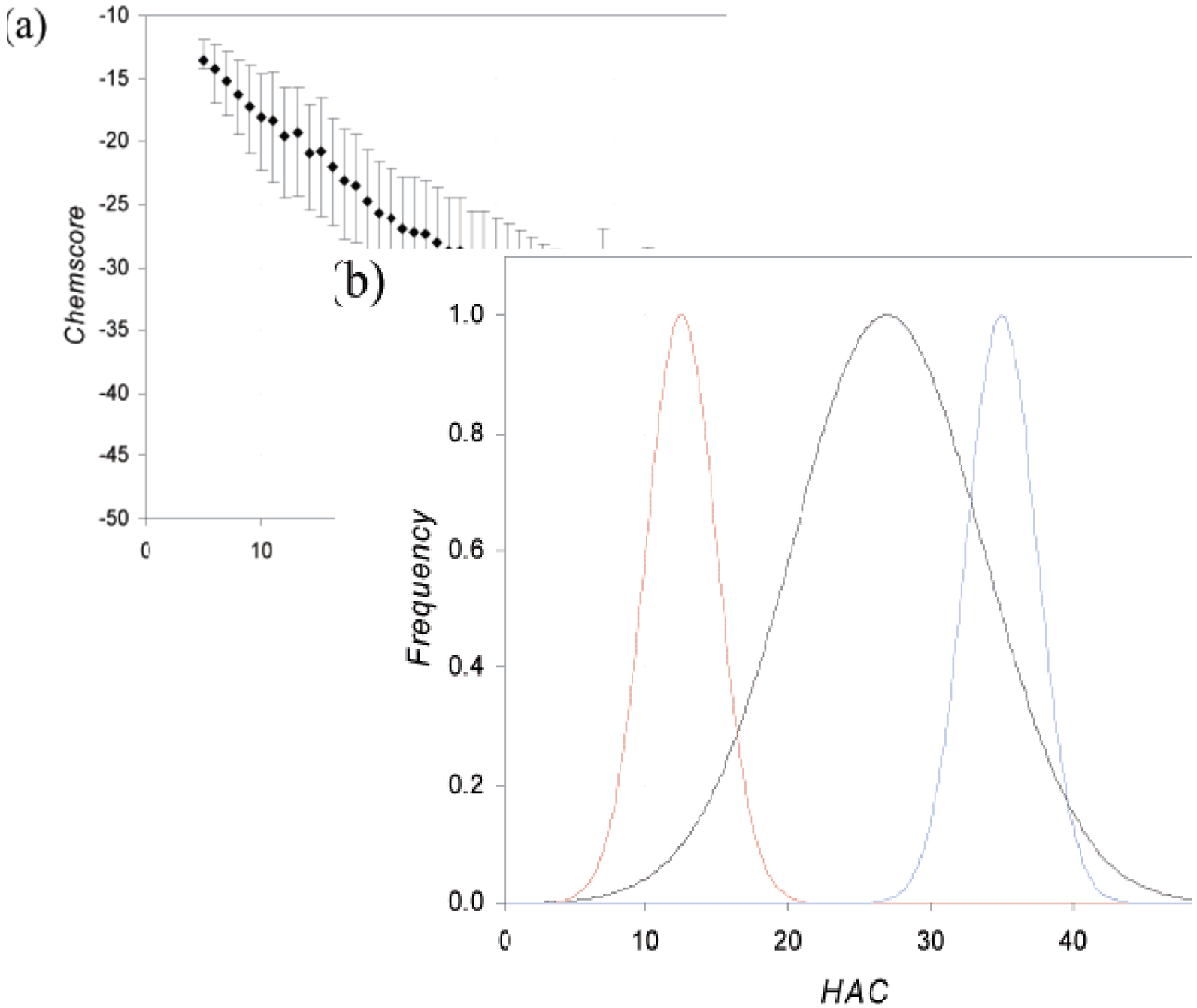


Watch out for accidental cheating



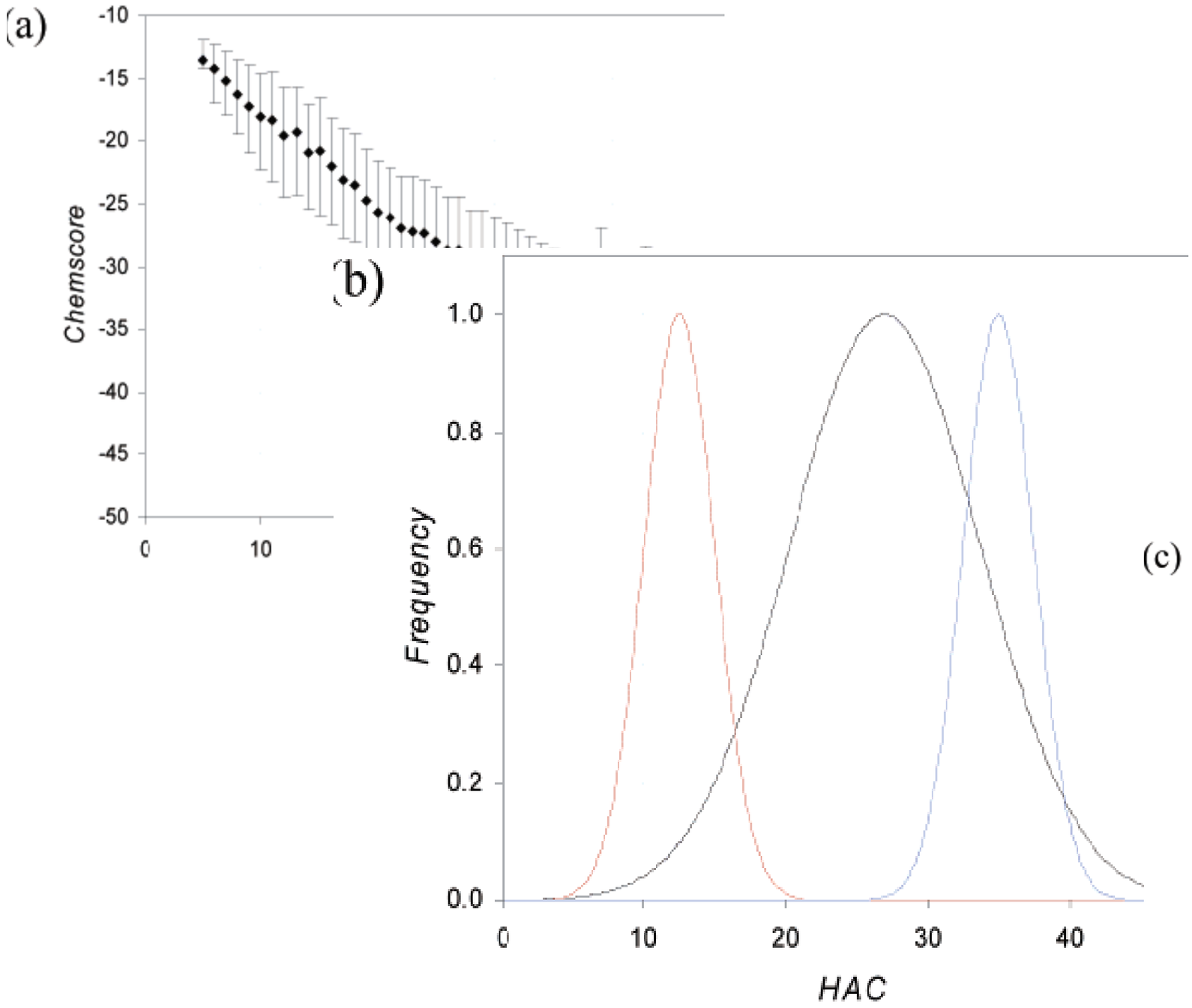
http://johnirwin.docking.org/talks/acs07-1.pdf

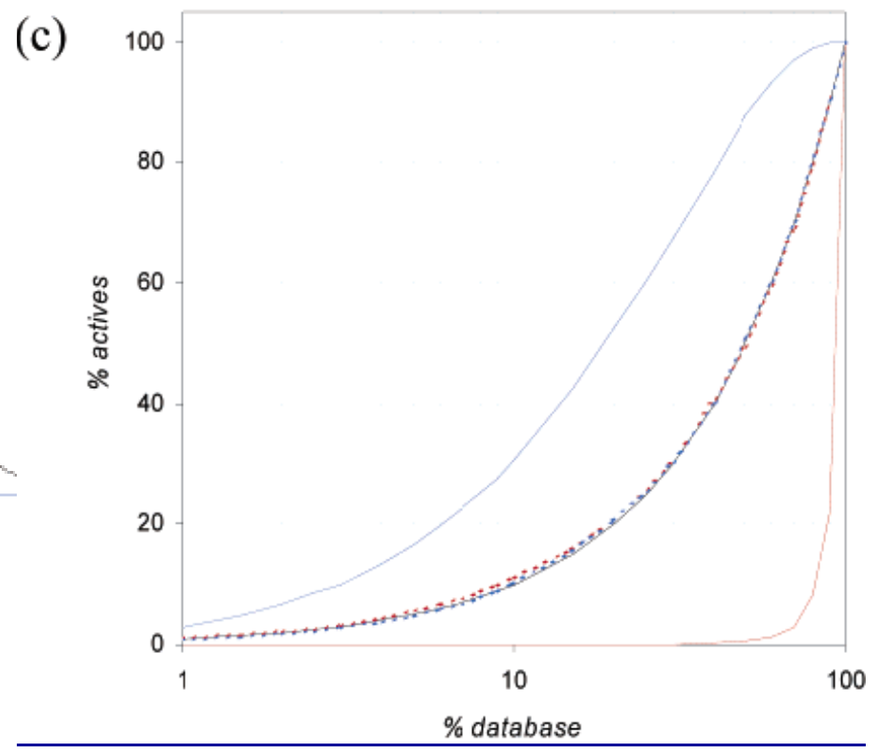
Watch out for accidental cheating



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

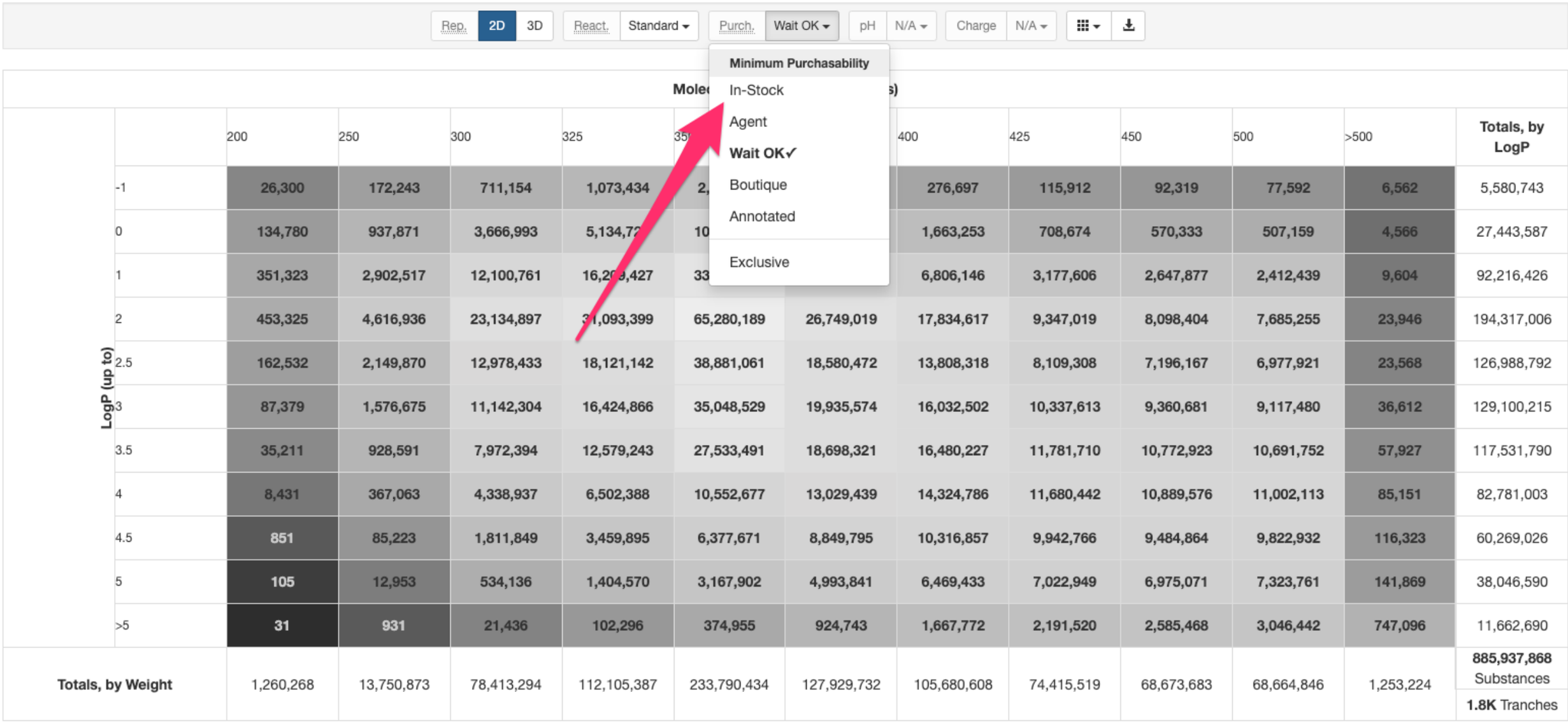
Rep. 2D 3D | React. Standard | Purch. Wait OK | pH N/A | Charge N/A

Molecular Weight (up to, Daltons)

LogP (up to)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	26,300	172,243	711,154	1,073,434	2,241,963	786,567	276,697	115,912	92,319	77,592	6,562	5,580,743
0	134,780	937,871	3,666,993	5,134,727	10,617,504	3,497,727	1,663,253	708,674	570,333	507,159	4,566	27,443,587
1	351,323	2,902,517	12,100,761	16,209,427	33,714,492	11,884,234	6,806,146	3,177,606	2,647,877	2,412,439	9,604	92,216,426
2	453,325	4,616,936	23,134,897	31,093,399	65,280,189	26,749,019	17,834,617	9,347,019	8,098,404	7,685,255	23,946	194,317,006
2.5	162,532	2,149,870	12,978,433	18,121,142	38,881,061	18,580,472	13,808,318	8,109,308	7,196,167	6,977,921	23,568	126,988,792
3	87,379	1,576,675	11,142,304	16,424,866	35,048,529	19,935,574	16,032,502	10,337,613	9,360,681	9,117,480	36,612	129,100,215
3.5	35,211	928,591	7,972,394	12,579,243	27,533,491	18,698,321	16,480,227	11,781,710	10,772,923	10,691,752	57,927	117,531,790
4	8,431	367,063	4,338,937	6,502,388	10,552,677	13,029,439	14,324,786	11,680,442	10,889,576	11,002,113	85,151	82,781,003
4.5	851	85,223	1,811,849	3,459,895	6,377,671	8,849,795	10,316,857	9,942,766	9,484,864	9,822,932	116,323	60,269,026
5	105	12,953	534,136	1,404,570	3,167,902	4,993,841	6,469,433	7,022,949	6,975,071	7,323,761	141,869	38,046,590
>5	31	931	21,436	102,296	374,955	924,743	1,667,772	2,191,520	2,585,468	3,046,442	747,096	11,662,690
Totals, by Weight	1,260,268	13,750,873	78,413,294	112,105,387	233,790,434	127,929,732	105,680,608	74,415,519	68,673,683	68,664,846	1,253,224	**885,937,868** Substances **1.8K** Tranches

http://zinc20.docking.org/tranches/home/

You can easily subset based on availability or other categories



http://zinc20.docking.org/tranches/home/

(The “in stock” subset is more modest)

105,222

Rep. 2D 3D | React. Standard | Purch. In-Stock | pH N/A | Charge N/A

Molecular Weight (up to, Daltons)

LogP (up to)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	7,564	5,665	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	45,094
0	28,699	20,270	35,000	21,579	22,273	14,052	8,745	4,545	3,069	3,022	4,276	165,530
1	77,543	81,302	161,967	112,333	120,921	71,383	42,380	20,170	12,238	11,221	8,510	719,968
2	105,222	181,327	399,448	318,958	389,588	230,270	149,732	76,243	50,651	42,725	19,259	1,963,423
2.5	38,264	101,944	265,621	236,384	321,651	204,539	145,627	83,368	58,497	49,925	18,830	1,524,650
3	21,956	82,779	247,196	239,988	359,194	251,845	195,747	121,779	90,466	84,258	30,366	1,725,574
3.5	9,431	51,618	182,319	196,565	310,312	253,674	218,734	155,053	122,800	122,489	51,041	1,674,036
4	2,887	24,661	109,239	126,391	209,523	215,034	209,288	171,762	144,621	158,241	78,804	1,450,451
4.5	592	8,939	53,627	71,560	122,687	151,306	169,498	159,349	144,746	174,790	111,949	1,169,043
5	103	2,586	20,624	33,421	62,058	89,134	110,752	118,808	122,285	170,826	139,628	870,225
>5	31	881	9,330	17,880	39,341	72,345	110,903	143,431	172,050	370,718	743,374	1,680,284
Totals, by Weight	292,292	561,972	1,491,949	1,379,077	1,961,694	1,556,430	1,363,497	1,056,055	922,622	1,190,340	1,212,350	**12,988,278** Substances
												960 Tranches

http://zinc20.docking.org/tranches/home/

There are also various predefined subsets

Rep. 2D 3D | React. Standard | Purch. In-Stock | pH N/A | Charge N/A

Predefined Subsets
All
None
Shards
Fragments
Flagments
Goldilocks
Lead-Like
Lugs
Drug-Like
Big-n-Greasy

Molecular Weight (up to, Daltons)

LogP (up to)	200	250	300	325	350	375	425	500	>500	Totals, by LogP
-1	7,564	5,665	7,578	4,018	4,146	,848	2,091	2,125	6,313	45,094
0	28,699	20,270	35,000	21,579	22,27	14,052	8,745	3,022	4,276	165,530
1	77,543	81,302	161,967	112,333	120,921	71,383	42,380	11,221	8,510	719,968
2	105,222	181,327	399,448	,958	389,588	230,270	149,732	42,725	19,259	1,963,423
2.5	38,264	101,944	265,621	236,384	321,651	204,539	145,627	49,925	18,830	1,524,650
3	21,956	82,779	247,196	239,988	359,194	251,845	195,747	84,258	30,366	1,725,574
3.5	9,431	51,618	182,319	196,565	310,312	253,674	218,734	122,489	51,041	1,674,036
4	2,887	24,661	109,239	126,391	209,523	215,034	209,288	158,241	78,804	1,450,451
4.5	592	8,939	53,627	71,560	122,687	151,306	169,498	174,790	111,949	1,169,043
5	103	2,586	20,624	33,421	62,058	89,134	110,752	170,826	139,628	870,225
>5	31	881	9,330	17,880	39,341	72,345	110,903	370,718	743,374	1,680,284
Totals, by Weight	292,292	561,972	1,491,949	1,379,077	1,961,694	1,556,430	1,363,497	1,190,340	1,212,350	**12,988,278** Substances
										960 Tranches

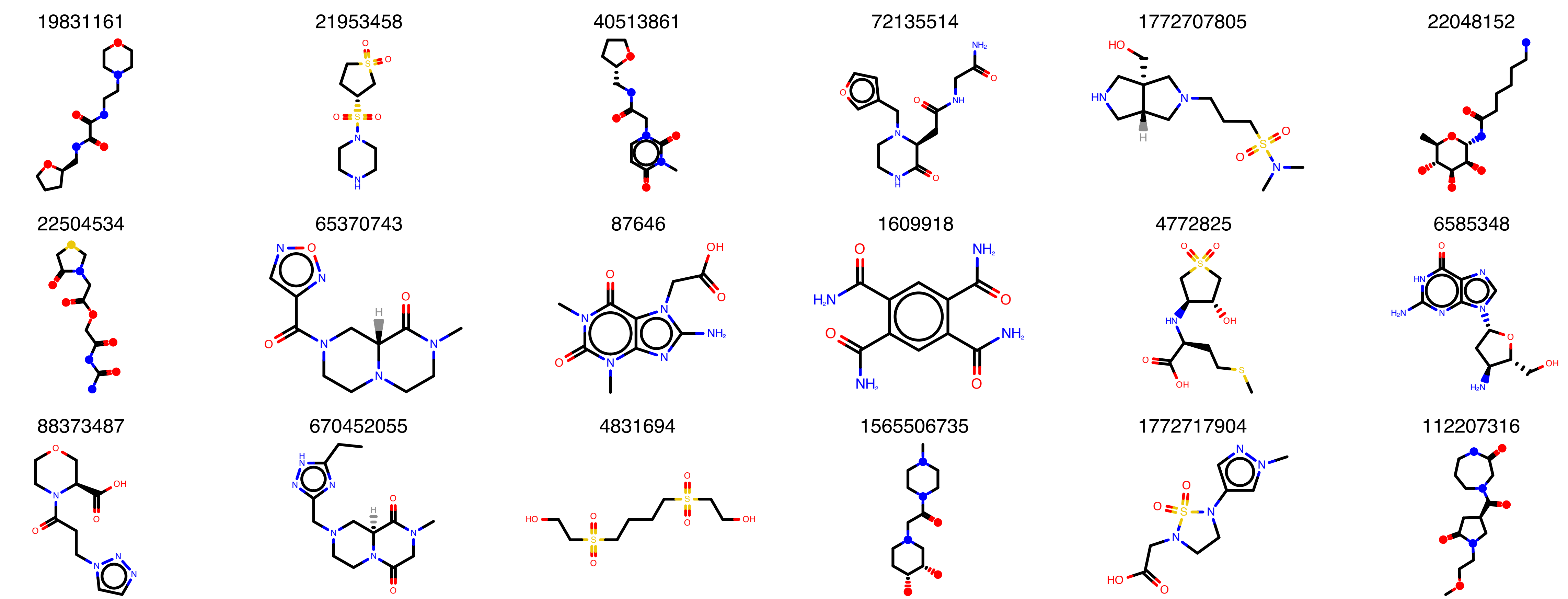
http://zinc20.docking.org/tranches/home/

Let's look at in-stock lead-like samples

		Molecular Weight (up to, Daltons)											
		200	250	300	325	350	375	400	425	450	500	>500	**Totals, by LogP**
LogP (up to)	-1	7,564	5,665	**7,578**	**4,018**	**4,146**	2,848	2,091	1,547	1,190	2,125	6,313	15,742
	0	28,699	20,270	**35,000**	**21,579**	**22,273**	14,052	8,745	4,545	3,049	3,022	4,276	78,852
	1	77,543	81,302	**161,967**	**112,333**	**120,921**	71,383	42,360	20,170	12,238	11,221	8,510	395,221
	2	105,222	181,327	**399,448**	**318,958**	**389,588**	230,270	149,732	76,243	50,651	42,725	19,259	1,107,994
	2.5	38,284	101,944	**265,621**	**236,384**	**321,651**	204,539	145,627	83,368	58,497	49,925	18,830	823,656
	3	21,956	82,779	**247,196**	**239,988**	**359,194**	251,845	195,747	121,779	90,466	84,258	30,366	846,378
	3.5	9,431	51,618	**182,319**	**196,565**	**310,312**	253,674	218,734	155,053	122,800	122,489	51,041	689,196
	4	2,887	24,661	109,239	126,391	209,523	215,034	209,288	171,762	144,621	158,241	78,804	0
	4.5	592	8,939	53,627	71,560	122,687	151,306	169,498	159,349	144,746	174,790	111,949	0
	5	103	2,586	20,624	33,421	62,058	89,134	110,752	118,808	122,285	170,826	139,628	0
	>5	31	881	9,330	17,880	39,341	72,345	110,903	143,431	172,050	370,718	743,374	0
Totals, by Weight		0	0	1,299,129	1,129,825	1,528,085	0	0	0	0	0	0	**3,957,039** Substances **168** Tranches

Let’s look at in-stock lead-like samples

LogP (up to)	Molecular Weight (up to, Daltons): 200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	7,564	5,665	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	15,742
0	28,699	20,270	35,000	21,579	22,273	14,052	8,745	4,545	3,069	3,022	4,276	78,852
1	77,543	81,302	161,967	112,333	120,921	71,383	42,380	20,170	12,238	11,221	8,510	395,221
2	105,222	181,327	399,448	318,958	389,588	230,270	149,732	76,243	50,651	42,725	19,259	1,107,994
2.5	38,264	101,944	265,621	236,384	321,651	204,539	145,627	83,368	58,497	49,925	18,830	823,656
3	[illegible]	[illegible]	247,196	239,988	359,194	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	846,378
3.5												
4												
4.5												
5												
>5												
Totals, by Weight												



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

		Molecular Weight (up to, Daltons)											
		200	250	300	325	350	375	400	425	450	500	>500	**Totals, by LogP**
LogP (up to)	-1	**7,564**	**5,665**	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	13,229
	0	**28,699**	**20,270**	35,000	21,579	22,273	14,052	8,745	4,545	3,059	3,022	4,276	48,969
	1	**77,543**	**81,302**	161,967	112,333	120,921	71,383	42,380	20,170	12,238	11,221	8,510	158,845
	2	**105,222**	**181,327**	399,448	318,958	389,588	230,270	149,732	76,243	50,651	42,725	19,259	286,549
	2.5	**38,264**	**101,944**	265,621	236,384	321,651	204,539	145,627	83,368	58,497	49,925	18,830	140,208
	3	**21,956**	**82,779**	247,196	239,988	359,194	251,845	195,747	121,779	90,466	84,258	30,366	104,735
	3.5	**9,431**	**51,618**	182,319	196,565	310,312	253,674	218,734	155,053	122,800	122,489	51,041	61,049
	4	**2,887**	**24,661**	109,239	126,391	209,523	215,034	209,288	171,762	144,621	158,241	78,804	0
	4.5	**592**	**8,939**	53,627	71,560	122,687	151,306	169,498	159,349	144,746	174,790	111,949	0
	5	**103**	**2,586**	20,624	33,421	62,058	89,134	110,752	118,808	122,285	170,826	139,628	0
	>5	**31**	**881**	9,330	17,880	39,341	72,345	110,903	143,431	172,050	370,718	743,374	0
Totals, by Weight		288,679	524,905	0	0	0	0	0	0	0	0	0	**813,584** Substances **112** Tranches

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

Molecular Weight (up to, Daltons)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	7,564	5,665	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	13,229
0	28,699	20,270	35,000	21,579	22,273	14,052	8,745	4,545	3,059	3,022	4,276	48,969
1	77,543	81,302	161,967	112,333	120,921	71,383	42,380	20,170	12,238	11,221	8,510	158,845
2												
2.5												
3												
3.5												
4												
4.5												
5												
>5												
Totals, by Weight												

LogP (up to)

200	250	300	325	350	375	400	425	450	500	>500
34370327	39401552	1575495	1576365	2263862	38415755
70992702	83119925	95708091	95644286	98210169	166514298
3861798	8727446	2504425	3257688	20258057	39778426
137815004	675072173	1689056	2570758	4096823	13514753

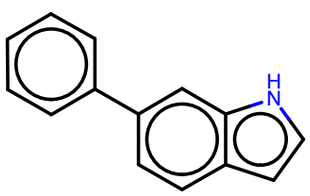
Let's look at samples from the "in stock" set: Shards now

		Molecular Weight (up to, Daltons)											
		200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
LogP (up to)	-1	7,564	5,665	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	7,564
	0	28,699	20,270	35,000	21,579	22,273	14,052	8,746	4,545	3,059	3,022	4,276	28,699
	1	77,543	81,302	161,967	112,333	120,921	71,383	42,360	20,170	12,238	11,221	8,510	77,543
	2	105,222	181,327	399,448	318,958	389,588	230,270	149,732	76,243	50,651	42,725	19,259	105,222
	2.5	38,264	101,944	265,621	236,384	321,651	204,539	145,627	83,388	58,497	49,925	18,830	38,264
	3	21,956	82,779	247,196	239,988	359,194	251,845	195,747	121,779	90,466	84,258	30,366	21,956
	3.5	9,431	51,618	182,319	196,565	310,312	253,674	218,734	155,053	122,600	122,489	51,041	9,431
	4	2,887	24,661	109,239	126,391	209,523	215,034	209,288	171,762	144,621	158,241	78,804	2,887
	4.5	592	8,939	53,627	71,560	122,687	151,306	169,498	159,349	144,746	174,790	111,949	592
	5	103	2,586	20,624	33,421	62,058	89,134	110,752	118,808	122,285	170,826	139,628	103
	>5	31	881	9,330	17,880	39,341	72,345	110,903	143,431	172,050	370,718	743,374	31
Totals, by Weight		292,292	0	0	0	0	0	0	0	0	0	0	**292,292** Substances **80** Tranches

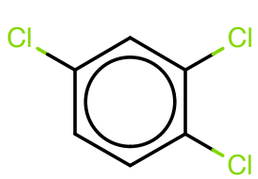
Let's look at samples from the "in stock" set: Shards now

LogP (up to)	Molecular Weight (up to, Daltons) 200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	7,564	5,665	7,578	4,018	4,146	2,848	2,091	1,547	1,199	2,125	6,313	7,564
0	28,699	20,270	35,000	21,579	22,273	14,052	8,746	4,545	3,059	3,022	4,276	28,699
1	77,543	81,302	161,967	112,333	120,921	71,383	42,360	20,170	12,238	11,221	8,510	77,543
2	105,222	181,327	399,448	318,958	389,588	230,270	149,732	76,243	50,651	42,725	19,259	105,222
2.5	38,264	101,944	265,621	236,384	321,651	204,539	145,627	83,388	58,497	49,925	18,830	38,264
3	21,956	82,779	247,196	239,988	359,194	251,845	195,747	121,779	90,466	84,258	30,366	21,956
3.5	9,431	51,618	182,319	196,565	310,312	253,674	218,734	155,053	122,800	122,489	51,041	9,431
4	2,887	24,661	109,239	126,391	209,523	215,034	209,288	171,762	144,621	158,241	78,804	2,887
4.5	592											
5	103											
>5	31											
Totals, by Weight	292,292											

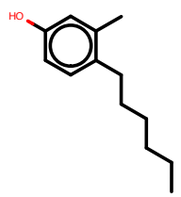
34099947



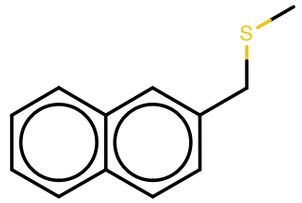
388208



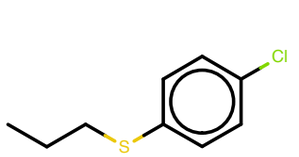
71773014



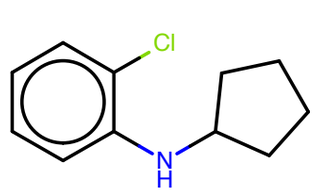
2172928



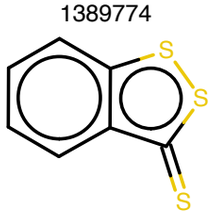
96024730



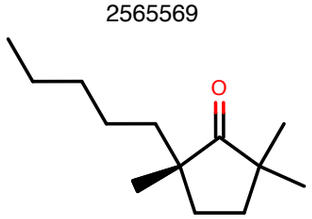
19942546



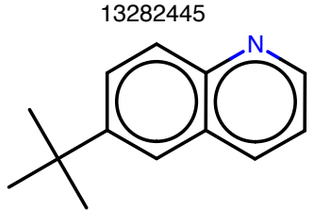
1389774



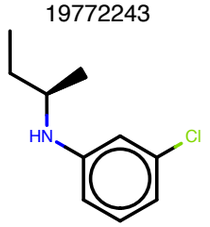
2565569



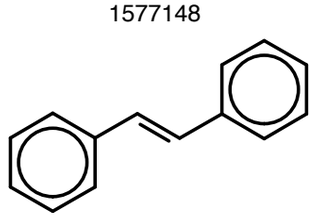
13282445



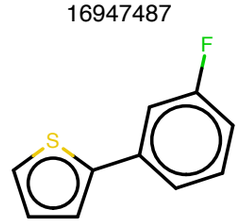
19772243



1577148



16947487



1615760

1681578

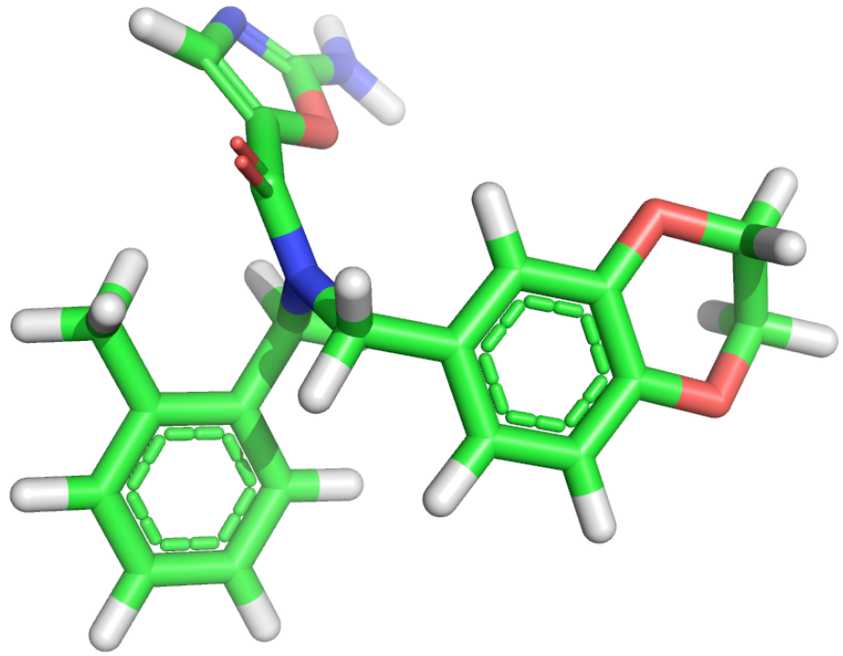
2757367

104279496

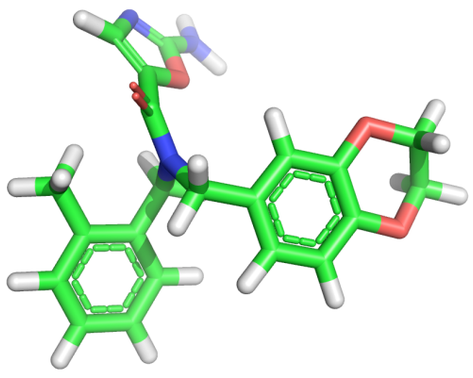
1272335

1589955

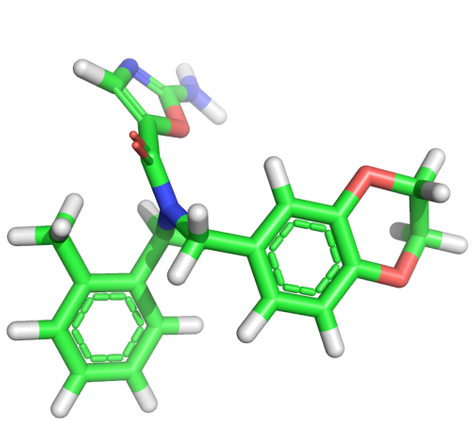
LINGO searches are extremely fast and make some chemical sense

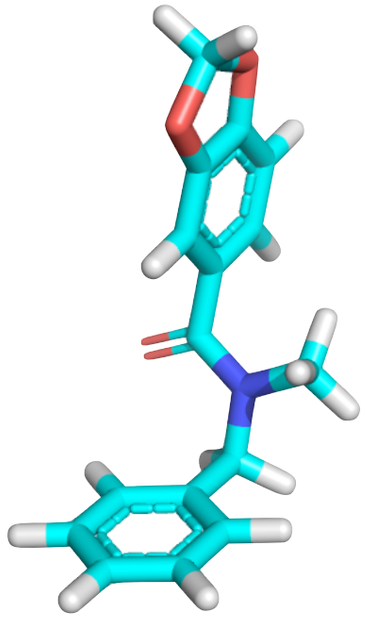


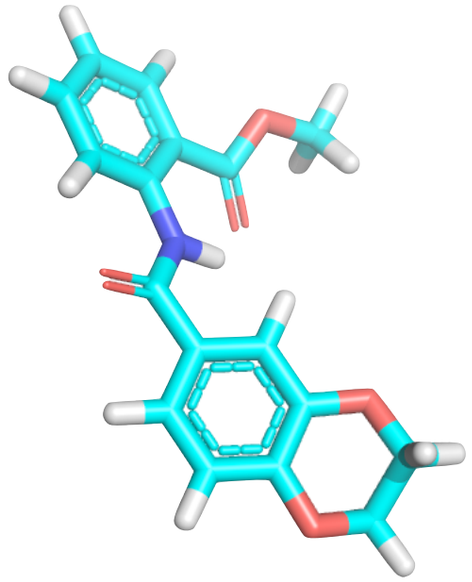
LINGO searches are extremely fast and make some chemical sense

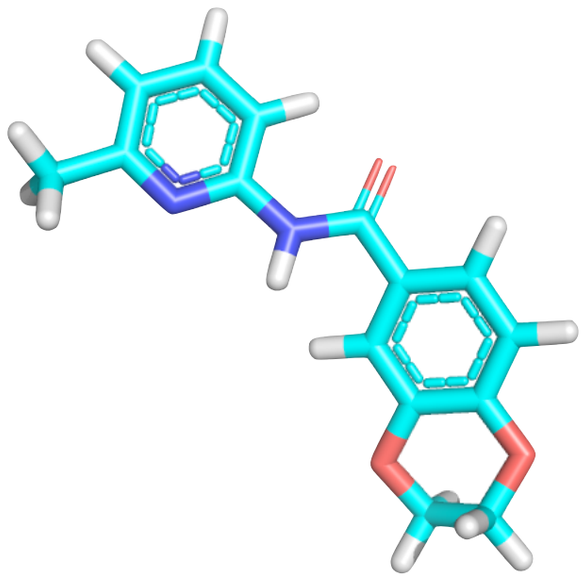


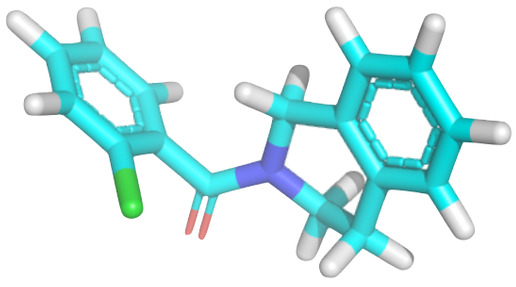
LINGO searches are extremely fast and make some chemical sense



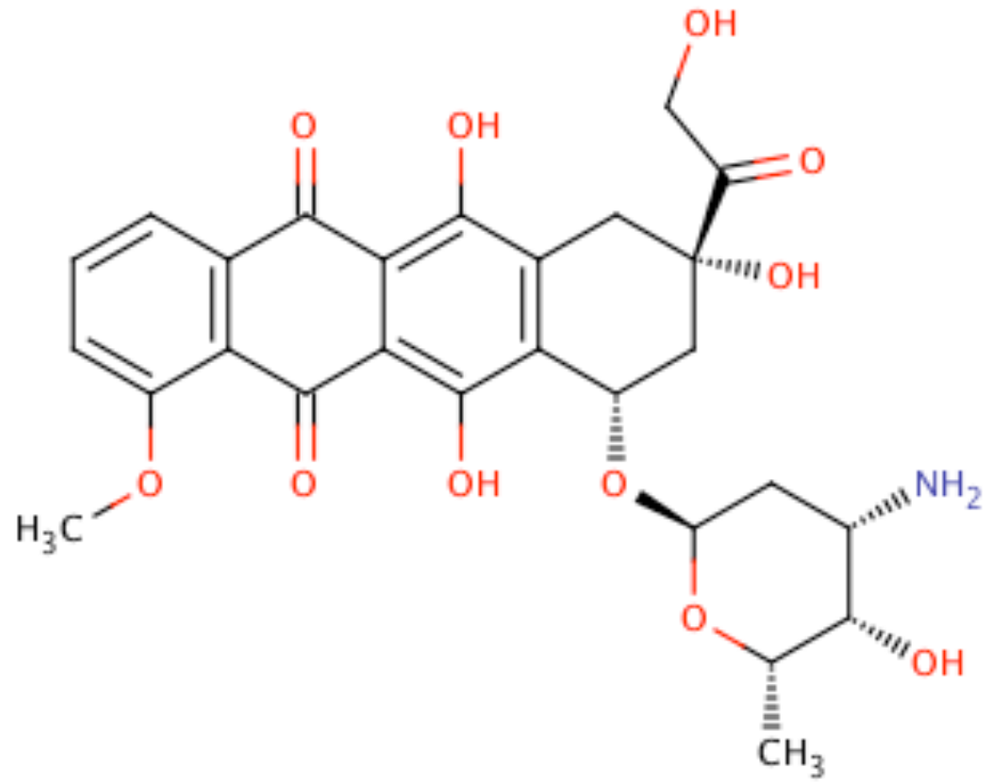








LINGO searches work based on SMILES strings



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

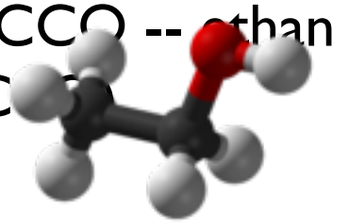
SMILES strings are a powerful, informative way to represent molecules

- "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

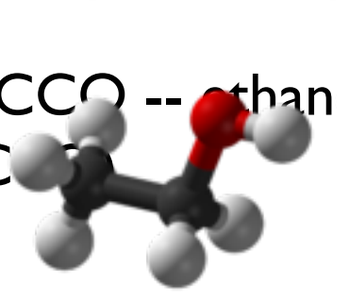
- "Simplified molecular-input line-entry specification"
- Simpler than reading/writing IUPAC names
- Human readable
- Element names and types (hydrogens implied)
  - C, B, N, O, P, S, F, Cl, Br, I. Other elements must be in brackets. Lowercase for aromatic
  - Hydrogens are assumed (except in brackets)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

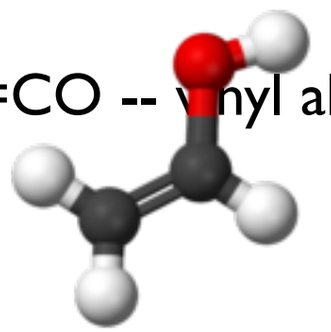


Images from Wikipedia

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

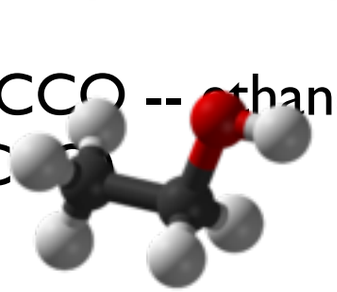


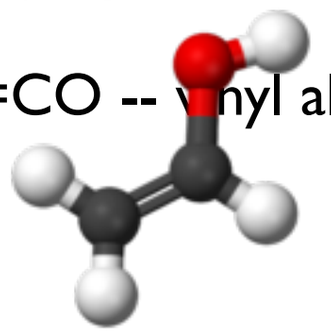


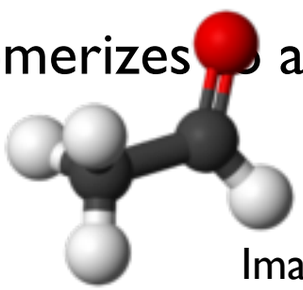
Images from Wikipedia

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)



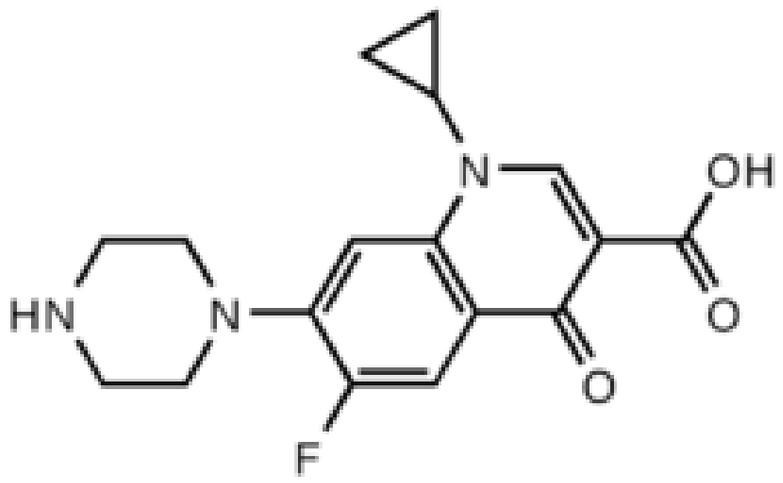




Images from Wikipedia

SMILES strings are a powerful, informative way to represent molecules

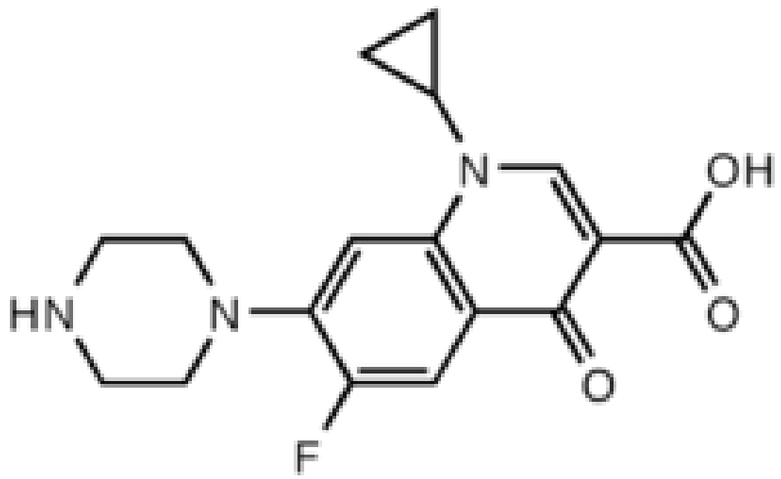
- Branching is indicated using parentheses
- Loop closure by numerically labeling atoms
  - Cyclohexane C1CCCCC1; dioxane O1CCCOCC1
- Generation is done by breaking cycles and writing as branches off a main backbone
- Ex:

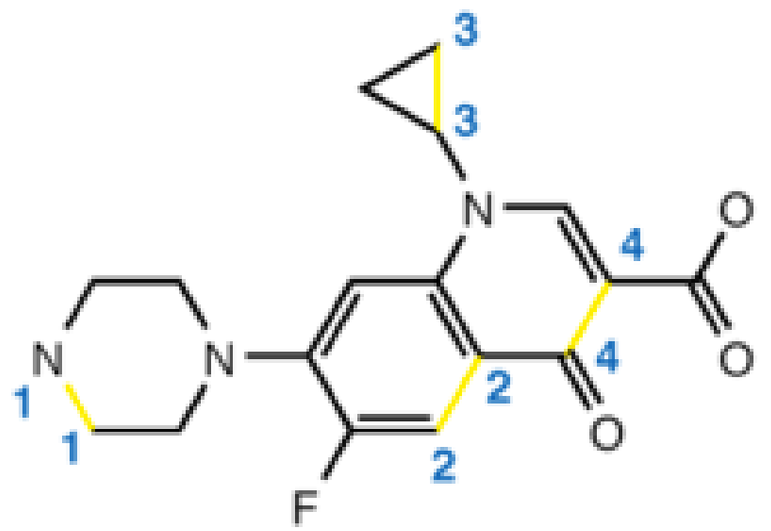


Some material, images from Wikipedia

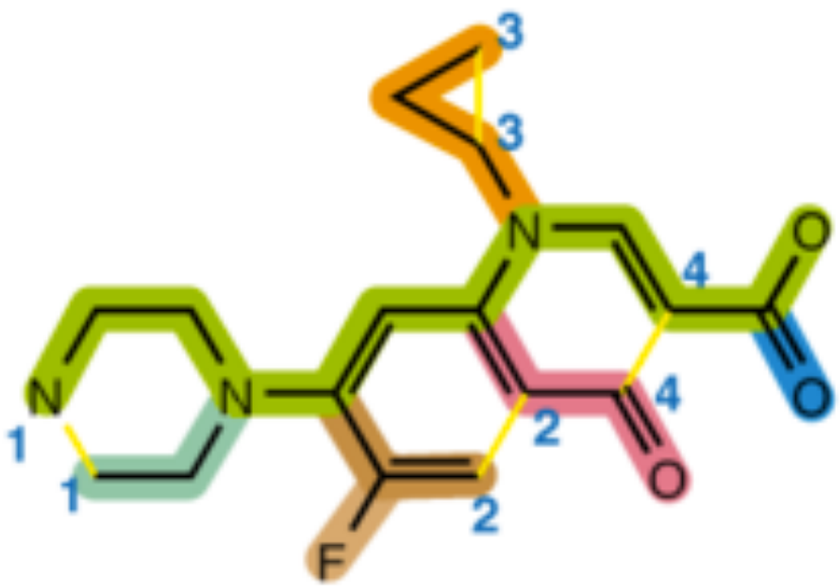
Example: Ciprofloxacin, an antibiotic

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

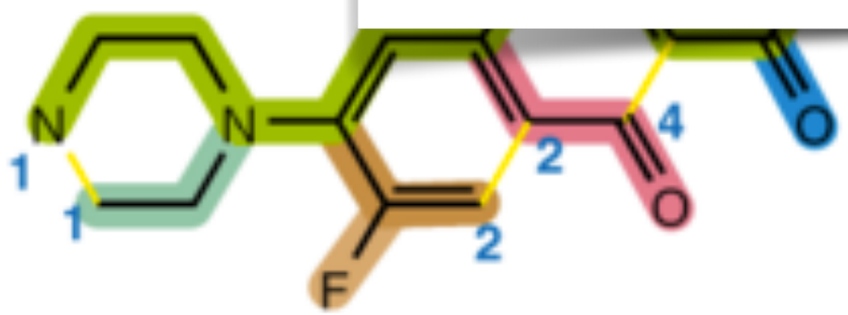
Break rings and identify backbone and places to close rings:

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

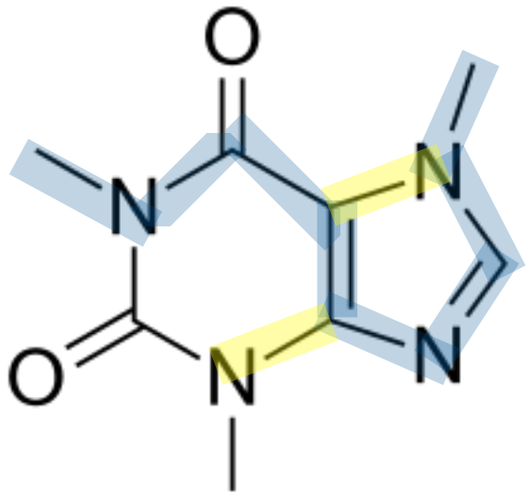
Write SM

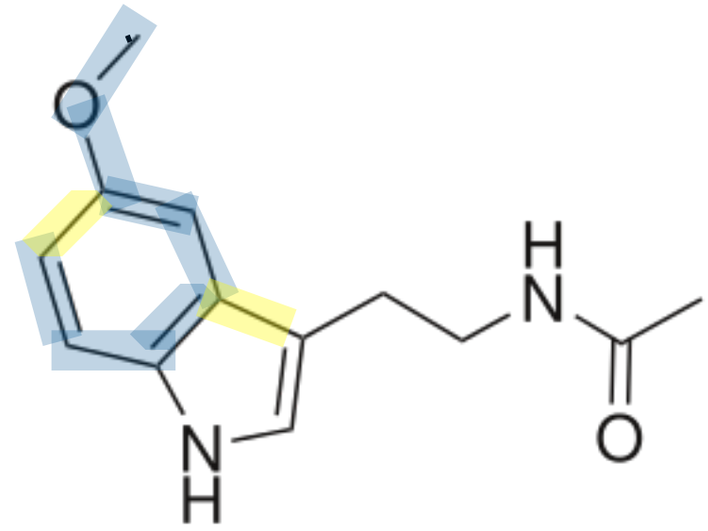
3CC3)C=C4C(=O)O



Some material, images from Wikipedia

Let’s try a couple examples together

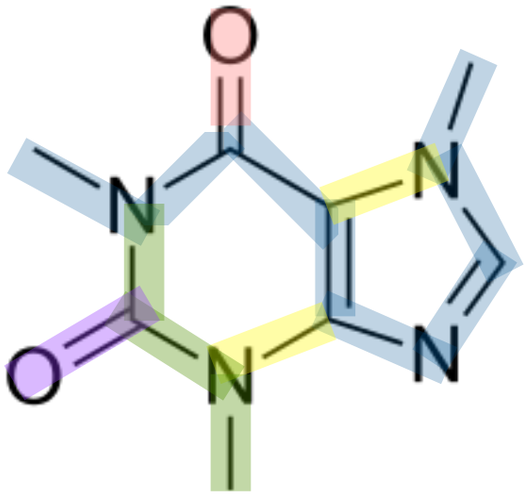


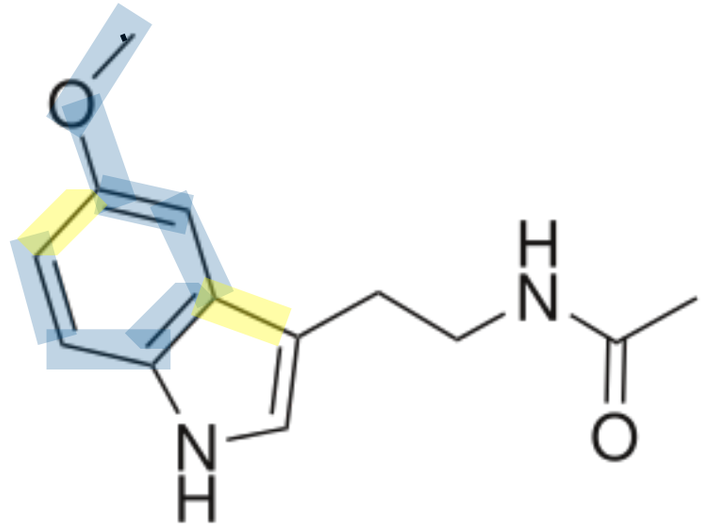


(remember: hydrogens implied)

Attrb: creative commons

Let’s try a couple examples together



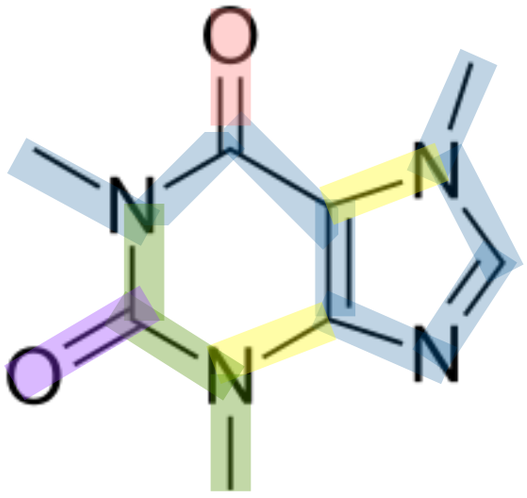


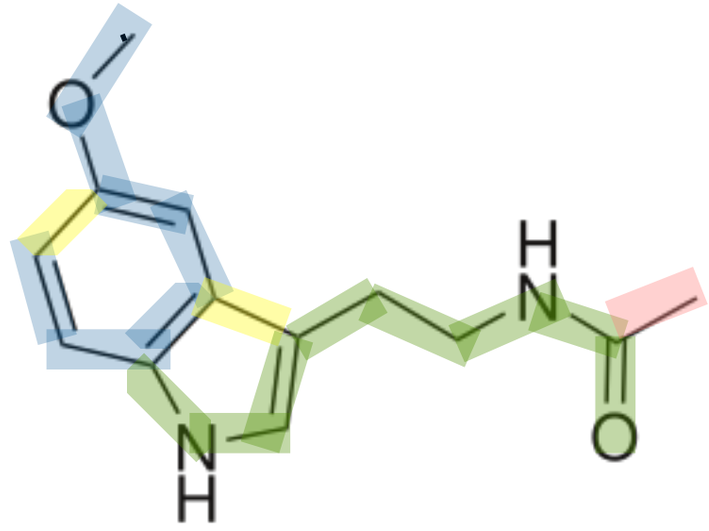
(remember: hydrogens implied)

Attrb: creative commons

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

Let’s try a couple examples together





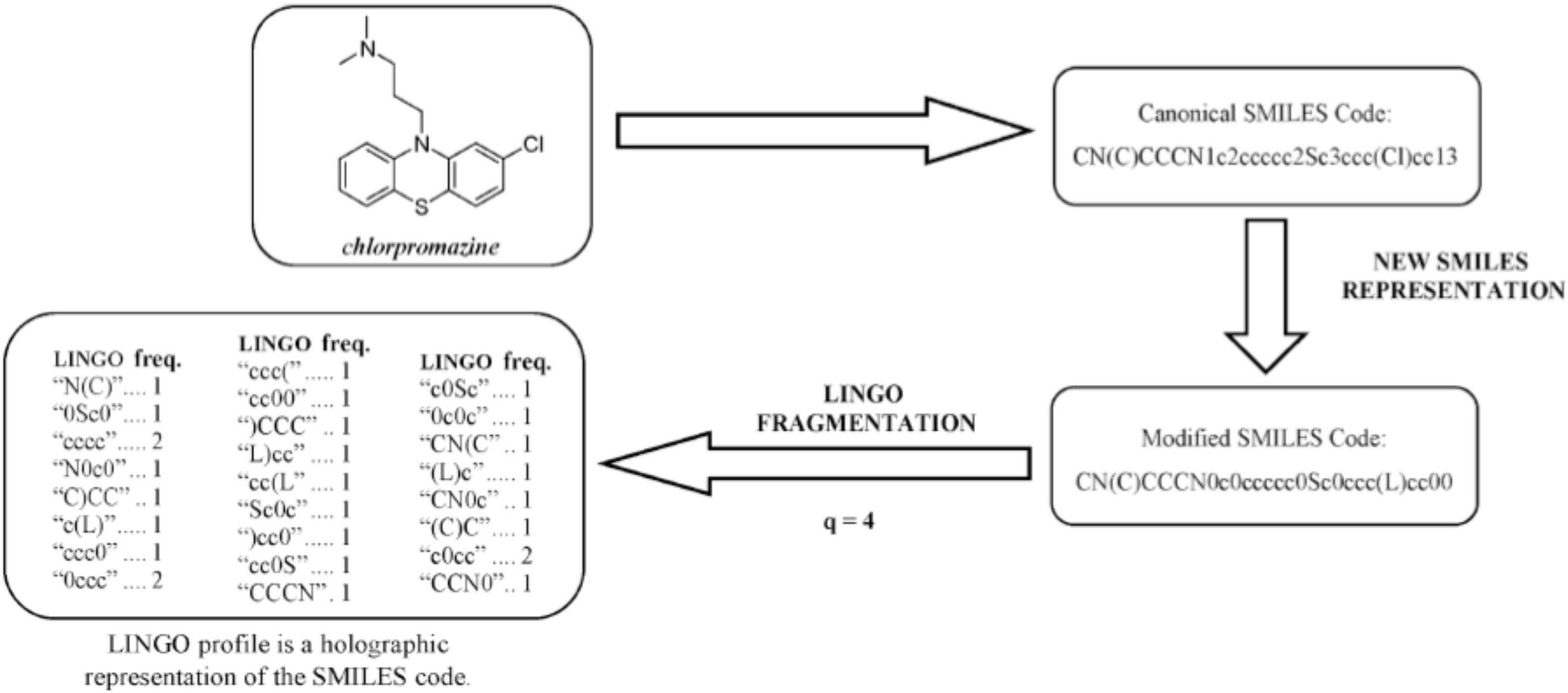
(remember: hydrogens implied)

Attrb: creative commons

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

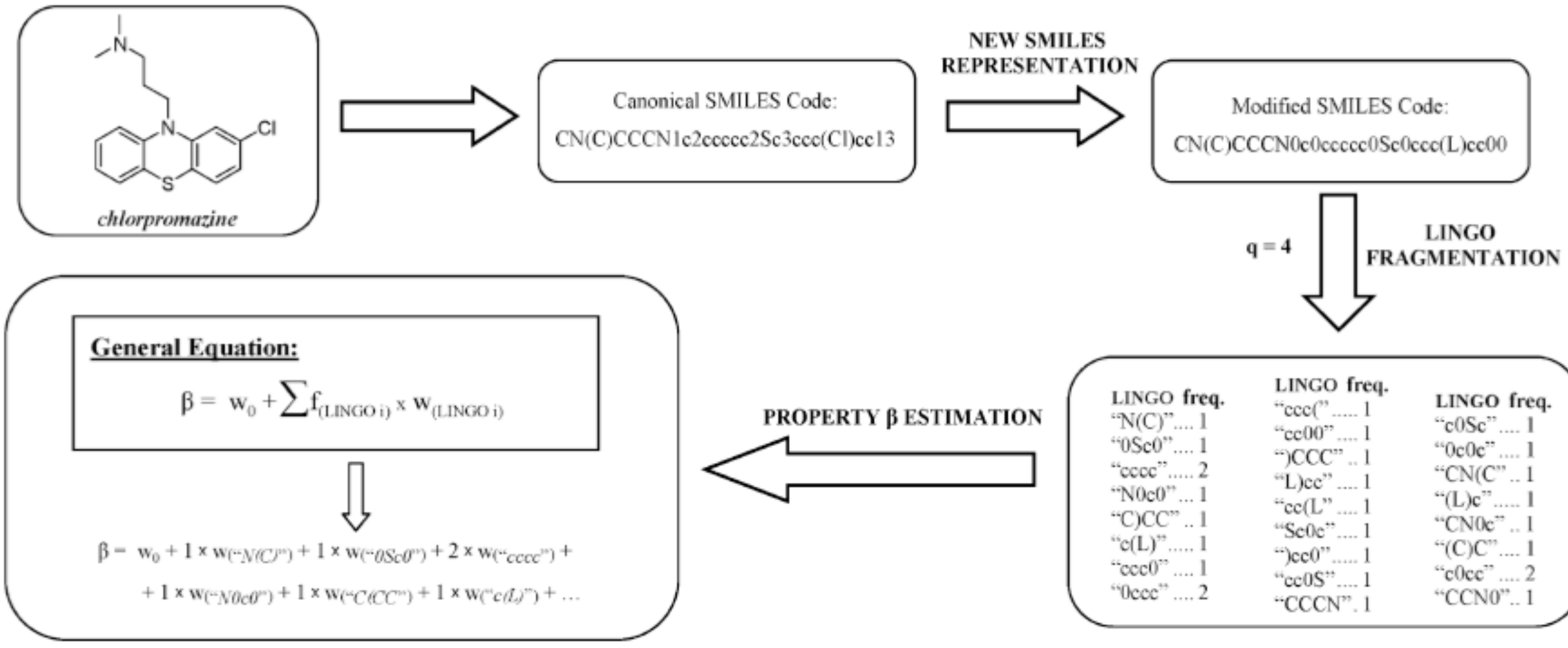
COC1=CC2=C(NC=C2CCNC(C)=O)C=C1

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

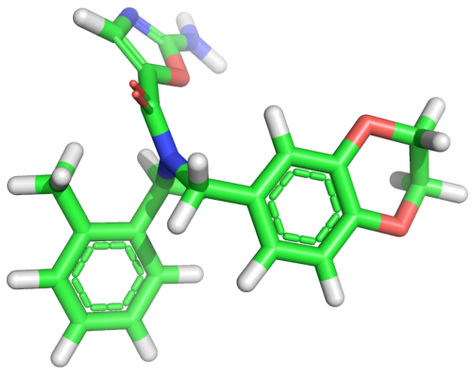


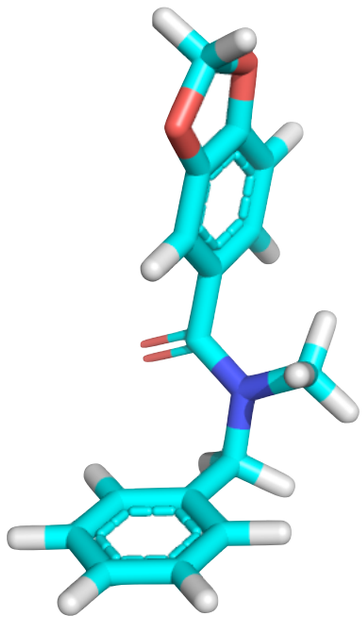
LINGO profile is a holographic representation of the SMILES code.

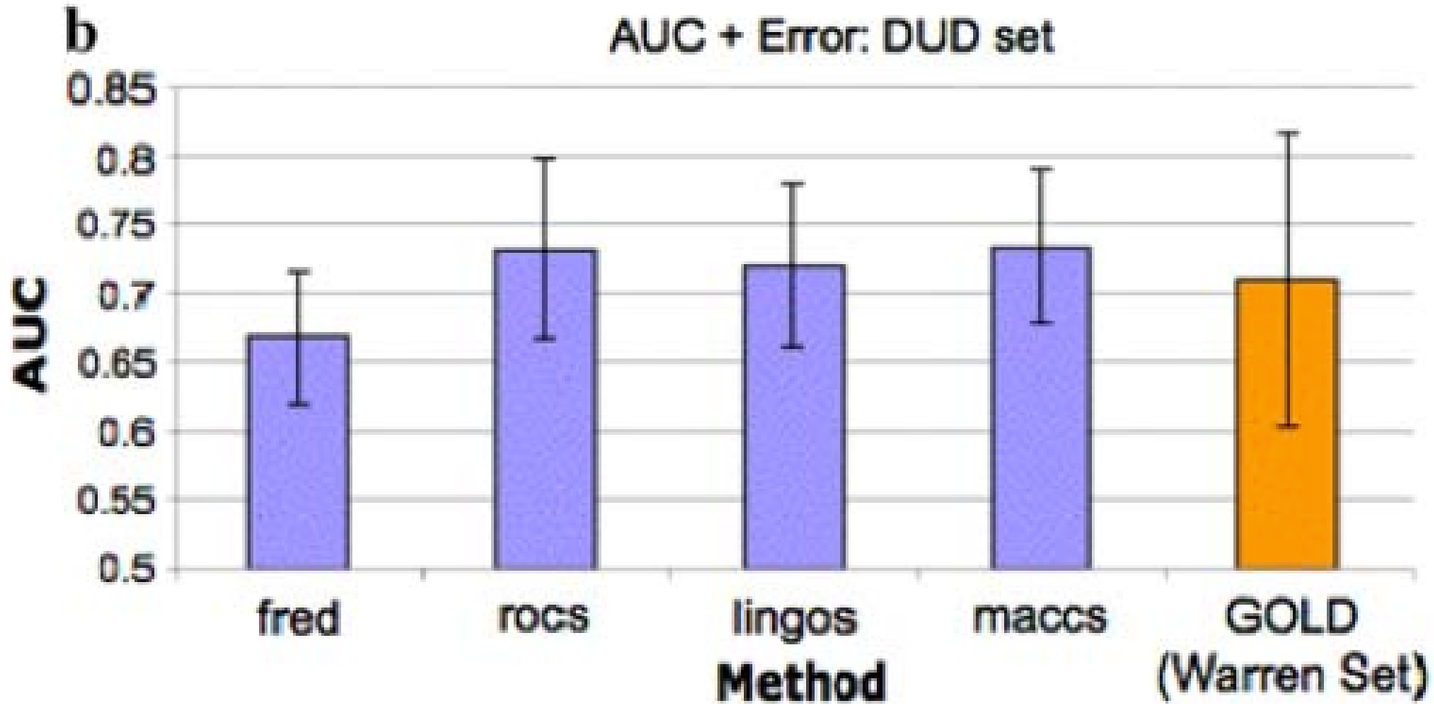
Some empirical models use LINGO to estimate properties based on functional groups



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

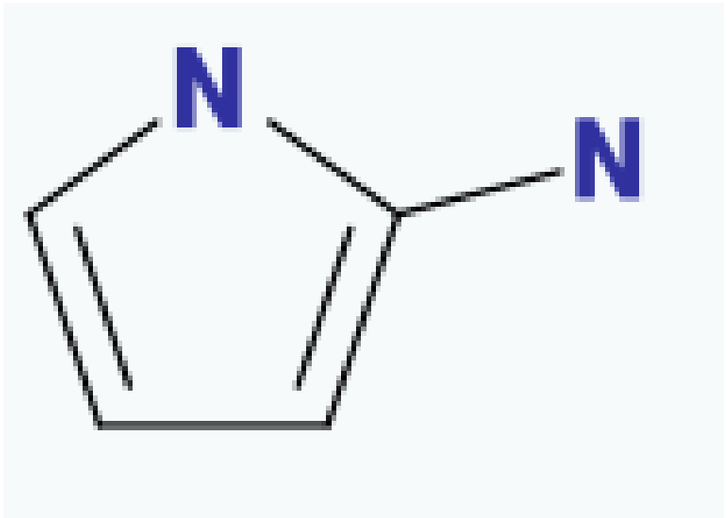


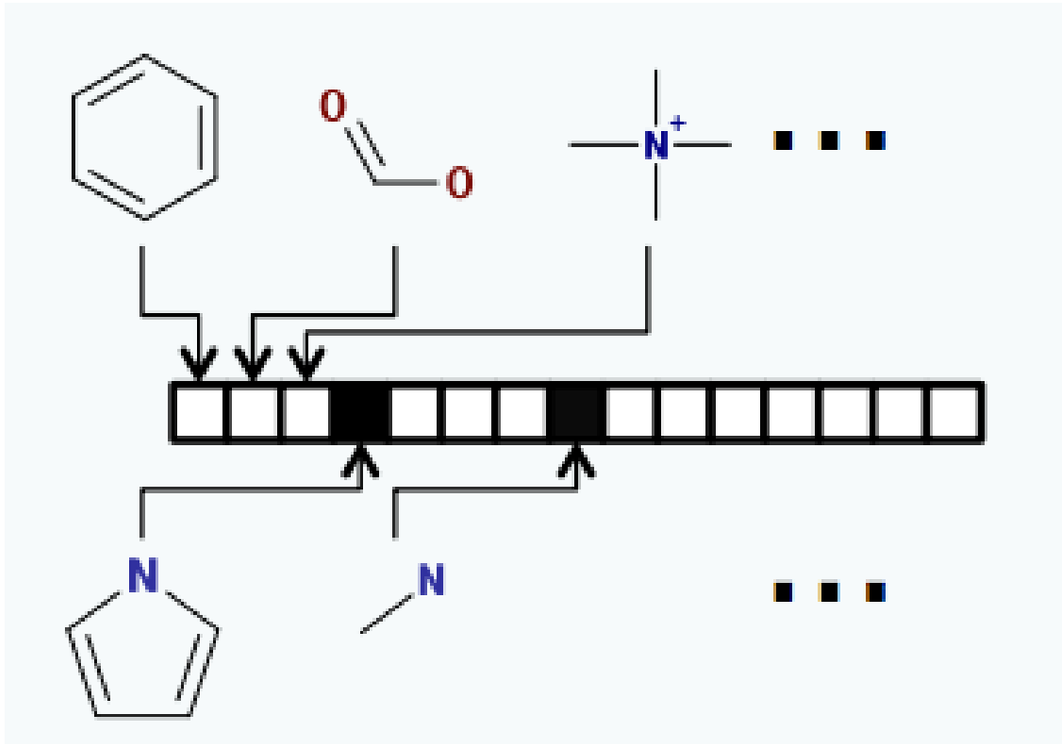




Nicholls, J. Comput. Aided Mol. Design 22:239 (2008)

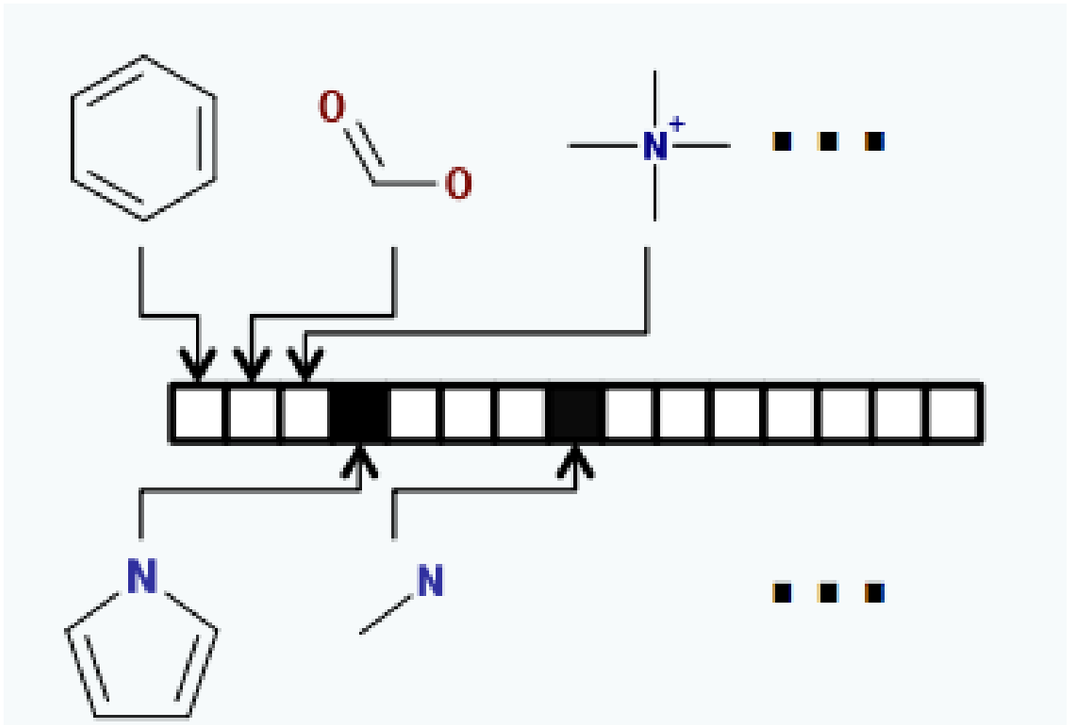
Fingerprint methods encode molecular descriptors in a fast way, often binary





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