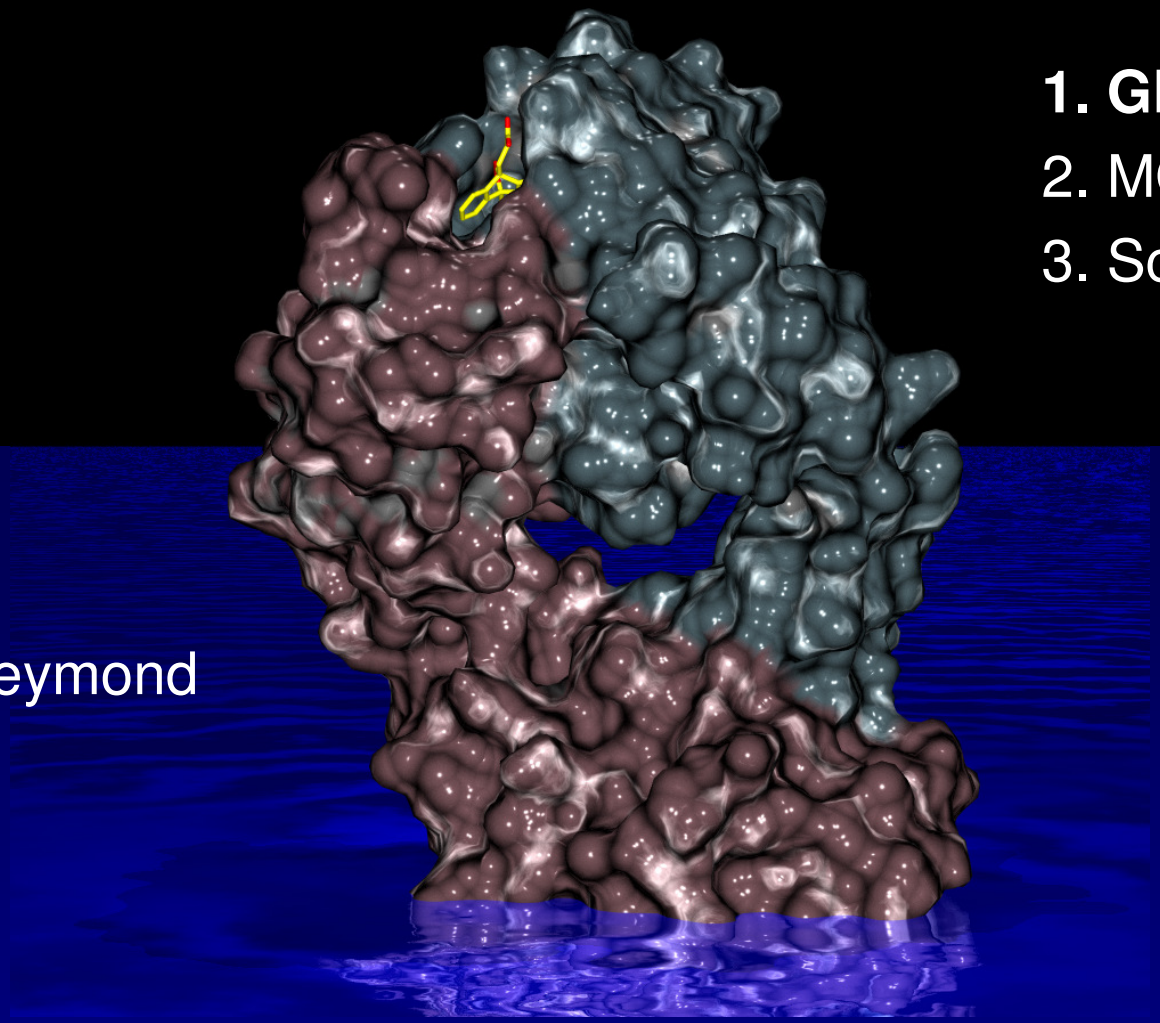
Enumerating, Mapping and Scoring of Chemical Space for Drug Discovery

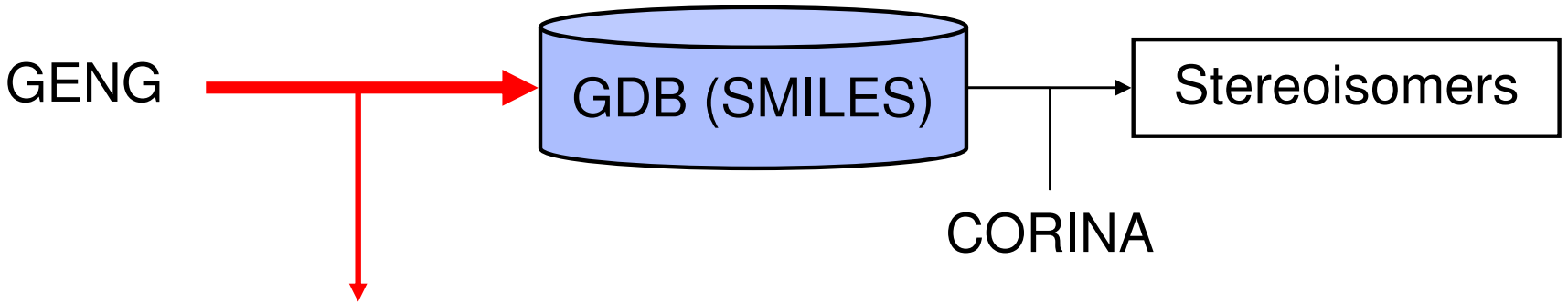


1. **GDB, CST**
2. MQN
3. Scoring

Jean-Louis Reymond
IPAM/UCLA
14 April 2011

J.-L. Reymond et al.. Chemical space as a source for new drugs. *Med. Chem. Commun.* **2010**, *1*, 30-38

GDB Assembly



I. Select hydrocarbon graphs (ring strain, topology)

II. C-C to C=C, C#C following valency rules, no allenes, no DB in bridgehead or 3- and 4-rings, no TB in <9-rings, no DB torsion

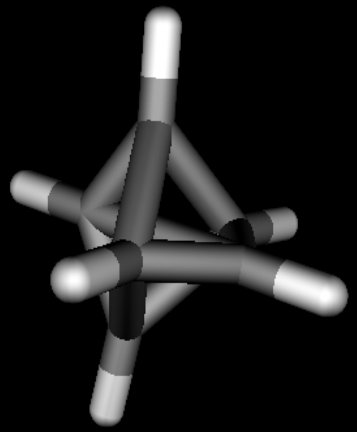
III. C to N or O following valency rules

IV. Filter bad functional groups, select tautomers

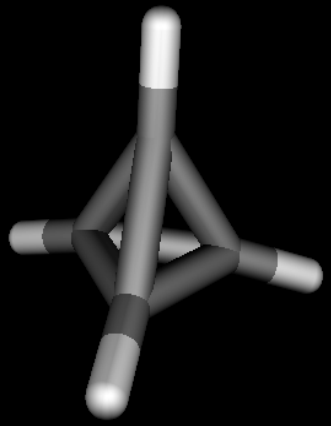
V. Post-processing: Halogens, S, etc.

T. Fink et al. *Angew. Chem. Int. Ed.* **2005**, *44*, 1504-1508, *J. Chem. Inf. Model.* **2007**, *47*, 342-353 (GDB-11)
L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732-3 (GDB-13)
Lars Ruddigkeit, Ruud van Deursen, Lorenz Blum (GDB-17)

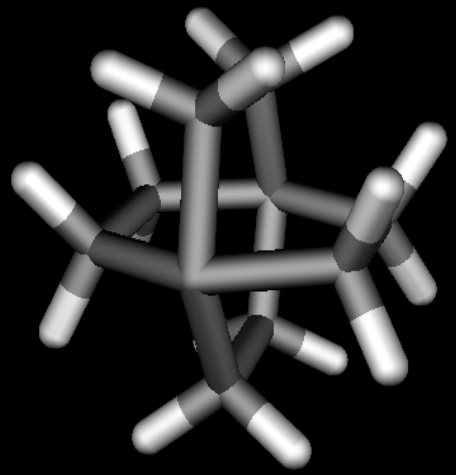
I. Graph Selection



Claus Benzol
$K_{3,3}$ Graph

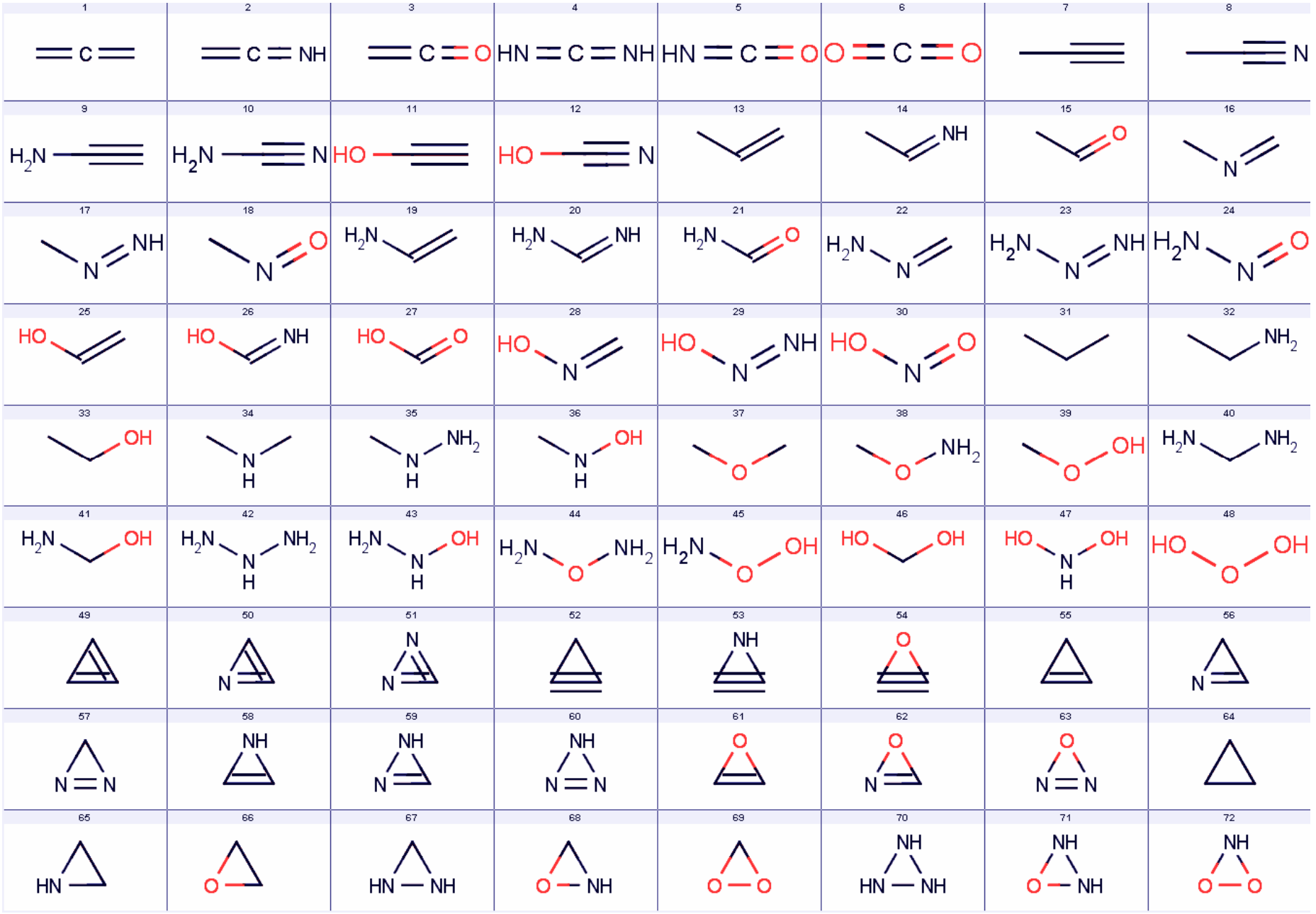


Tetrahedrane



Tricyclo[2.2.2.2]decane

IV. Filter Bad Functional Groups



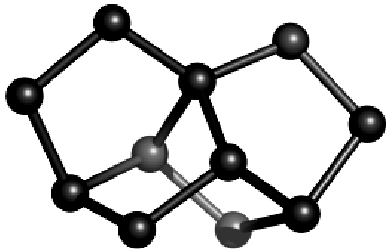
GDB-11 (CNOF)

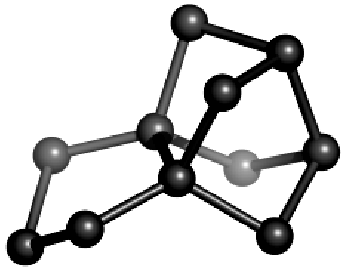
Table 2. Overview of the structure generation process.

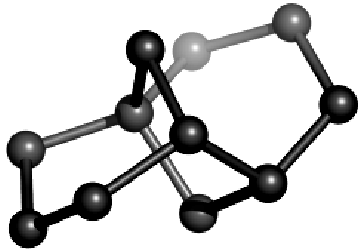
Nodes	Graphs[a]	Generated[b]	Accepted[c]	Unique Tautomers (GDB)[d]	All Tautomers	Stereoisomers[e]
1	1	4	4	4	4	4
2	1	10	9	9	9	9
3	2	52	20	20	21	20
4	4	332	80	80	88	87
5	8	2'294	357	352	397	469
6	20	18'066	1'906	1'850	2'135	2'911
7	57	154'542	10'953	10'568	12'438	19'904
8	194	1'445'073	69'563	66'706	79'899	153'601
9	705	14'213'741	464'402	444'313	540'002	1'258'963
10	2'822	146'004'340	3'259'036	3'114'041	3'827'907	10'898'065
11	11'912	1'558'491'448	23'875'101	22'796'628	28'240'425	98'645'474
Total	**15'726**	**1'720'329'902**	**27'681'431**	**26'434'571**	**32'703'325**	**110'979'507**

99.8 % are unknown

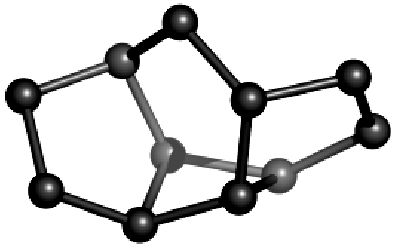
T. Fink et al., *J. Chem. Inf. Model.* **2007**, *47*, 342-353

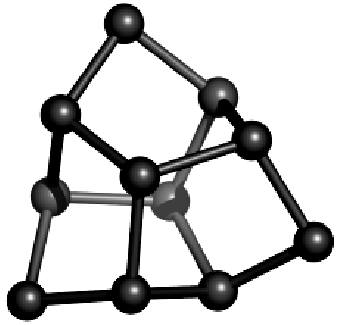






	Number of 4-membered rings			
Number of 3-membered rings	0	1	2	**Total**
0	124 [3]	189 [60]	103 [67]	**416 [130]**
1	225 [50]	238 [177]	20 [19]	**483 [246]**
2	201 [88]	55 [48]	-	**256 [136]**
3	53 [26]	-	-	**53 [26]**
Total	**603 [167]**	**482 [285]**	**123 [86]**	**1'208 [538]**

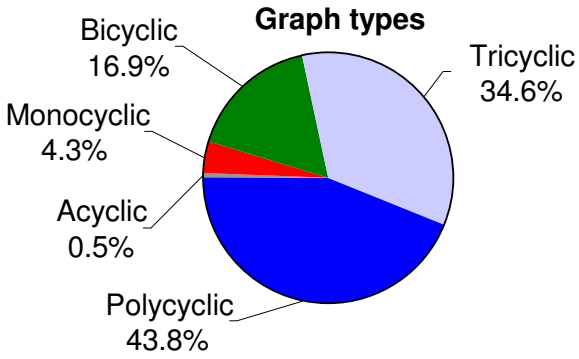


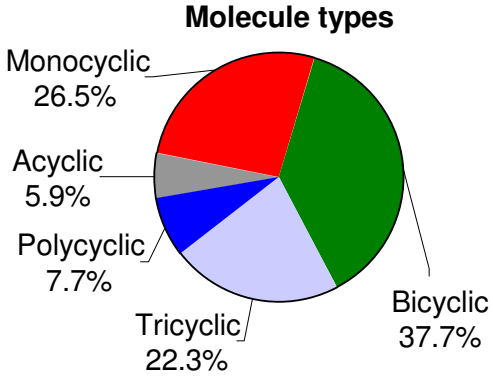


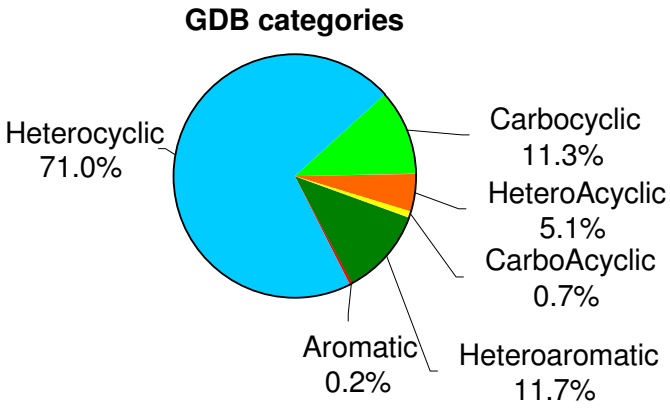
GDB-13 (CNOSCl, max. heteroatom ratio)

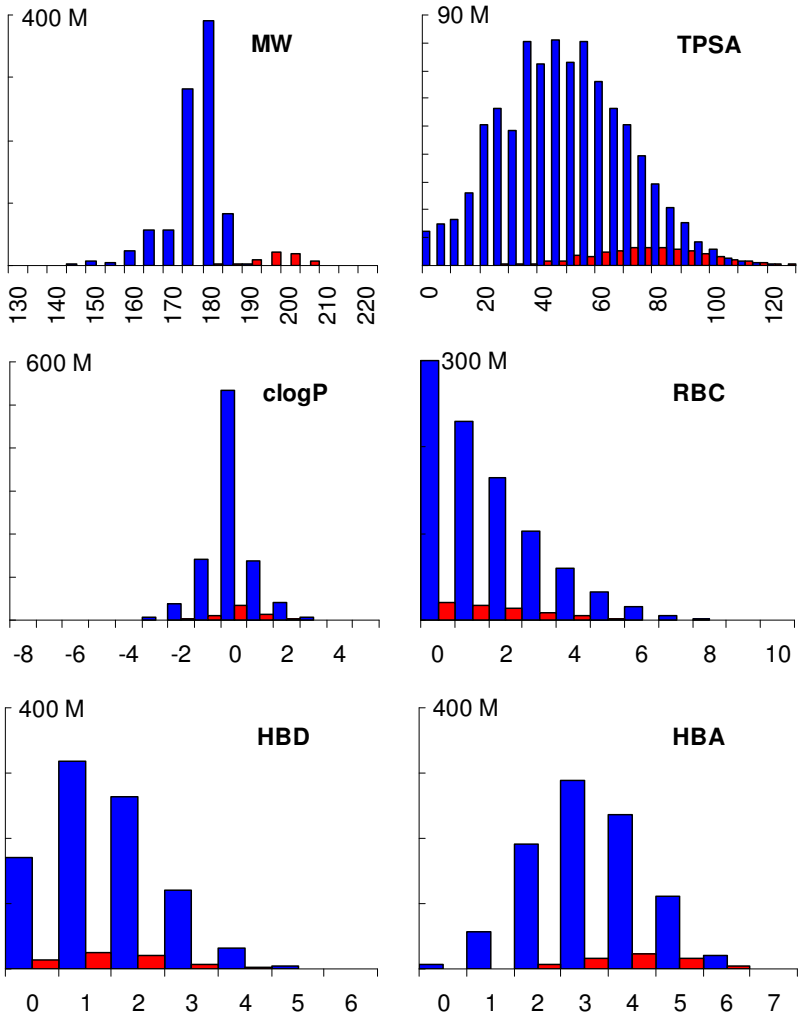
nodes[a]	graphs[b]	GDB[c]	Cl/S[d]
1	1	1	0
2	1	3	0
3	2	12	0
4	4	43	0
5	8	155	3
6	20	934	19
7	57	5 726	315
8	194	37 151	2 438
9	706	255 542	17 056
10	2 831	1 784 626	130 465
11	12 011	12 961 686	938 704
12	53 789	99 821 343	7 240 108
13	250 268	795 244 451	59 027 533
Total	**319 892**	**910 111 673**	**67 356 641**

L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732-3


Graph types
Bicyclic
16.9%
Tricyclic
34.6%
Monocyclic
4.3%
Acyclic
0.5%
Polycyclic
43.8%

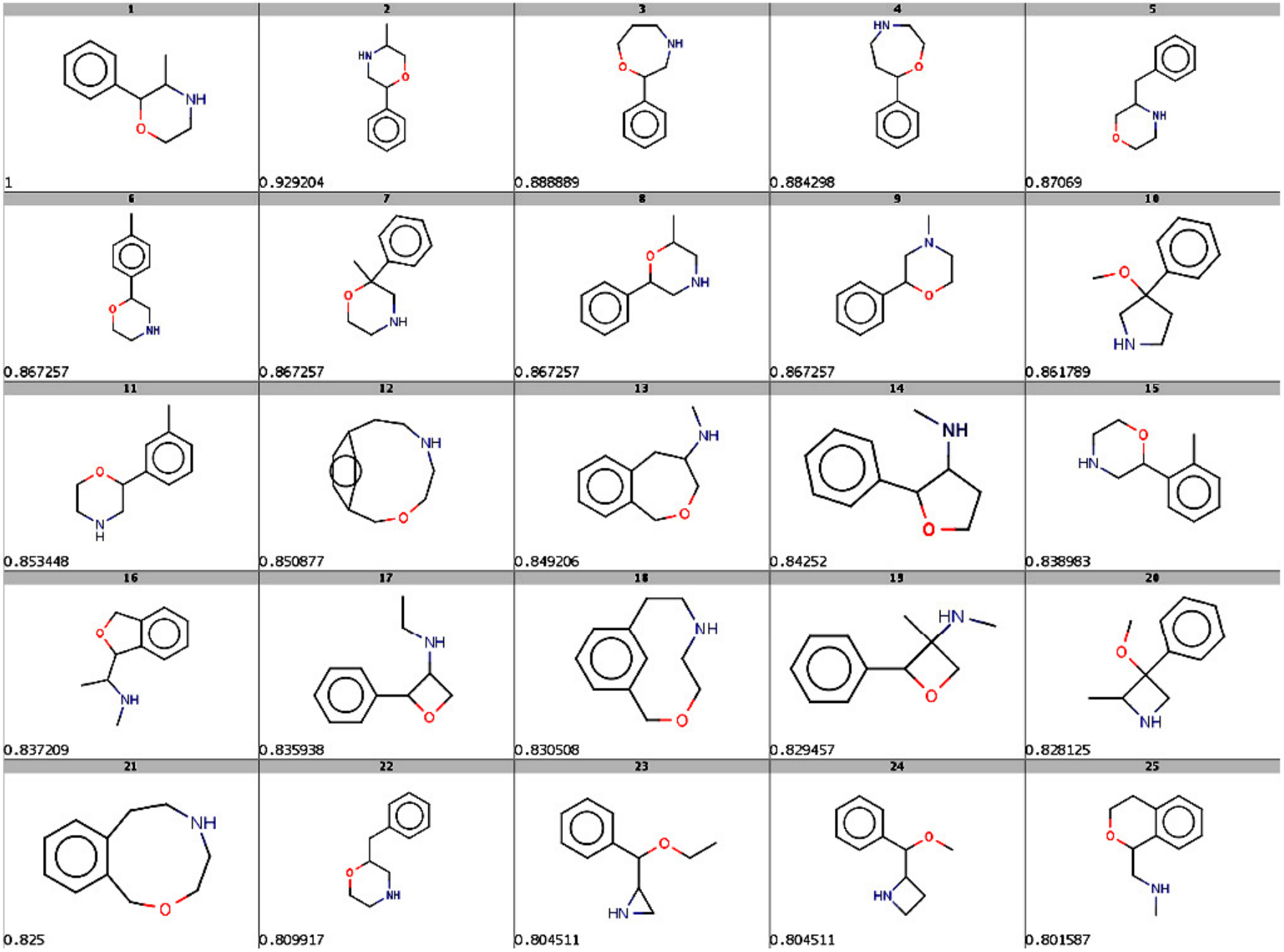

Molecule types
Monocyclic
26.5%
Acyclic
5.9%
Polycyclic
7.7%
Tricyclic
22.3%
Bicyclic
37.7%


GDB categories
Heterocyclic
71.0%
Carbocyclic
11.3%
HeteroAcyclic
5.1%
CarboAcyclic
0.7%
Aromatic
0.2%
Heteroaromatic
11.7%


400 M
MW
130 140 150 160 170 180 190 200 210 220
90 M
TPSA
0 20 40 60 80 100 120
600 M
clogP
-8 -6 -4 -2 0 2 4
300 M
RBC
0 2 4 6 8 10
400 M
HBD
0 1 2 3 4 5 6
400 M
HBA
0 1 2 3 4 5 6 7

Drugs and isomers in GDB-13

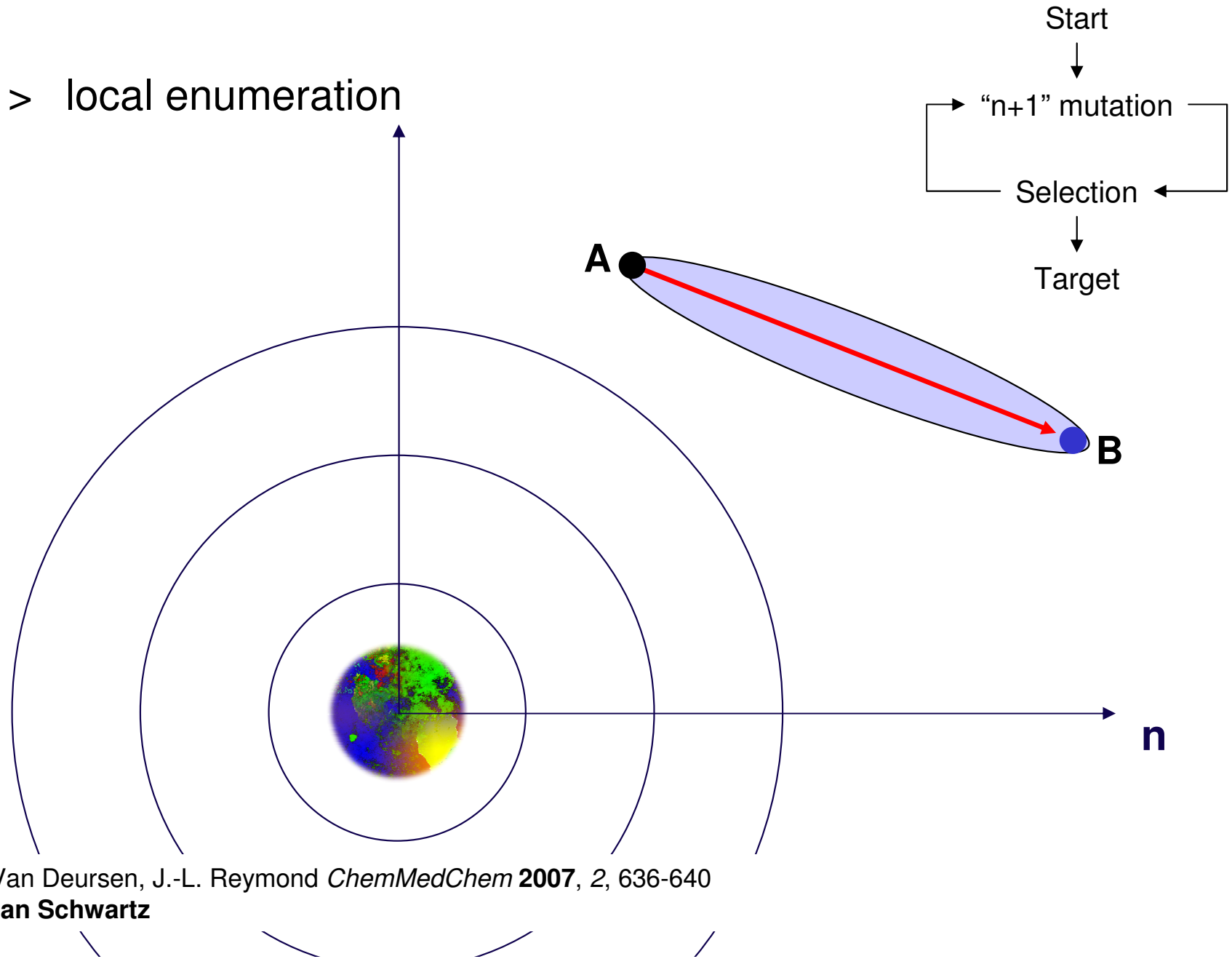
			T_{SF}	
Name	**Formula**	**Same Formula**	**AVG**	**> 0.7**
Aspirin	$C_9H_8O_4$	804 153	0.23	178
Benzocaine	$C_9H_{11}NO_2$	1 846 579	0.24	74
L-Tyrosine	$C_9H_{11}NO_3$	9 276 529	0.46	24 952
Levetiracetam	$C_8H_{14}N_2O_2$	2 154 955	0.28	35
Memantine	$C_{12}H_{21}N$	2 872 586	0.31	10 912
Menadione	$C_{11}H_8O_2$	233 715	0.44	112 186
Metaraminol	$C_9H_{13}NO_2$	2 920 516	0.26	30
Mexiletine	$C_{11}H_{17}NO$	18 371 393	0.25	119
Propofol	$C_{12}H_{18}O$	5 263 227	0.25	240
Rasagiline	$C_{12}H_{13}N$	1 323 525	0.13	411
Rimantadine	$C_{12}H_{21}N$	2 872 586	0.26	173



Phenmetrazine isomers

Chemical Space Travel

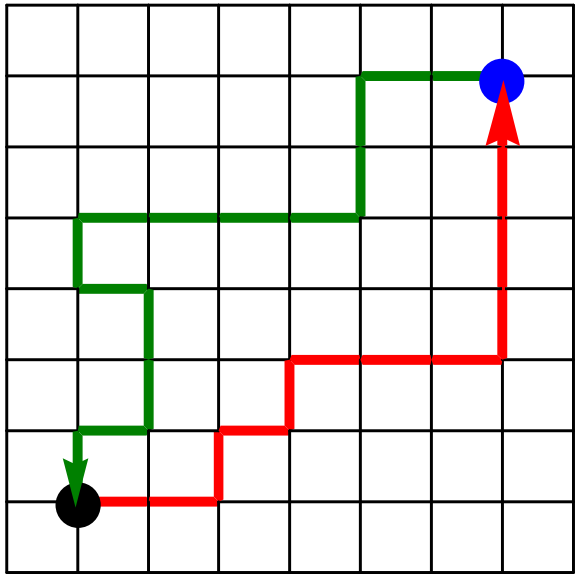
> local enumeration



R. Van Deursen, J.-L. Reymond *ChemMedChem* **2007**, *2*, 636-640

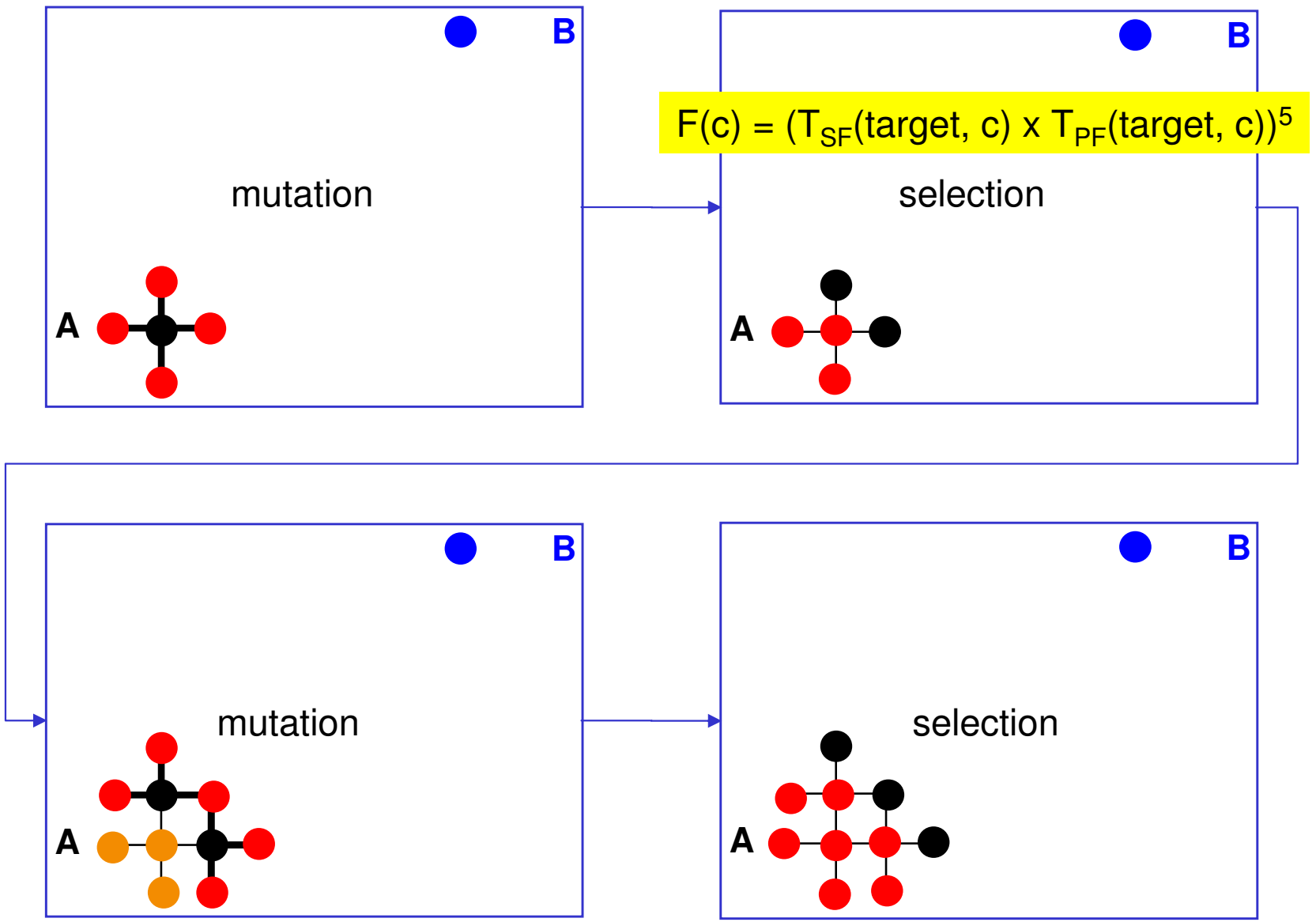
Julian Schwartz

Chemical Space as a Graph

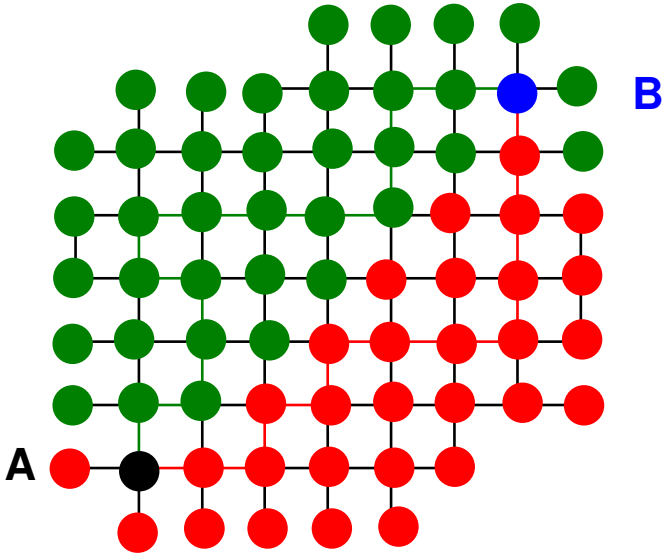


Nearest neighbour mutations

Atom type exchange
Atom inversion
Atom removal
Atom addition
Bond saturation
Bond unsaturation
Bond rearrangement

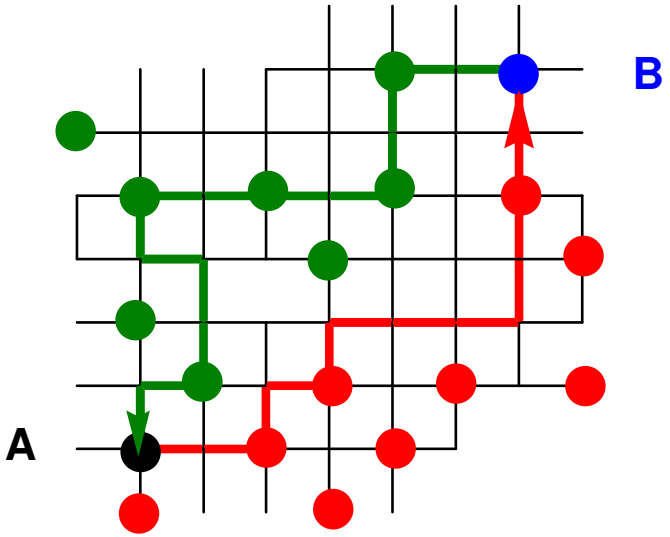

B
mutation
A
B
$F(c) = (T_{SF}(target, c) \times T_{PF}(target, c))^5$
selection
A
B
mutation
A
B
selection
A

Trajectory Libraries



"crude"

Filters



possible

Trajectory Examples (1)

Compound	Formula	mass	n[a]	Steps from CH_4[b] Nearest neighbours	Steps from CH_4[b] With aromatic[c]	N[d]	Steps to MeOH[b]	N[d]
Cubane	C_8H_8	104	8	12	-	6'638	7	994
Fluorouracil	$C_4H_3FN_2O_2$	130	9	16	9*	2'456	7*	560
Metheneamine	$C_6H_{12}N_4$	140	10	12	-	6'157	9	1'768
3-Tetrazene-2-carboximidamide	$C_2H_6N_{10}$	170	12	12	-	4'685	11*	2'007
Aspirine	$C_9H_8O_4$	180	13	15	8(1)	2'567	12	2'582
9-Ethyl-carbazole	$C_{14}H_{13}N$	197	15	n.f.[e]	20(2)	20'501	16	5'357
Vitamin H	$C_{10}H_{16}N_2O_3S$	244	16	18*	-	27'161	14*	6'304
VX (van)	$C_{11}H_{26}NO_2PS$	267	16	21	-	29'460	14*	3'954
Adenosine	$C_{10}H_{13}N_5O_4$	267	19	n.f.[e]	25(2)	23'680	19*	13'639
β-estradiol	$C_{18}H_{24}O_2$	272	20	23	15(2)	43'089	20	19'067
Retinal	$C_{20}H_{28}O$	284	21	23	-	45'176	19*	15'100
Morphine	$C_{17}H_{19}NO_3$	285	21	26	18(2)	69'113	20*	16'247
Aspartame	$C_{14}H_{18}N_2O_5$	294	21	303	16(2)	34'172	20*	11'430
Cocaine	$C_{17}H_{21}NO_4$	303	22	n.f.[e]	20(2)*	70'807	22	17'993

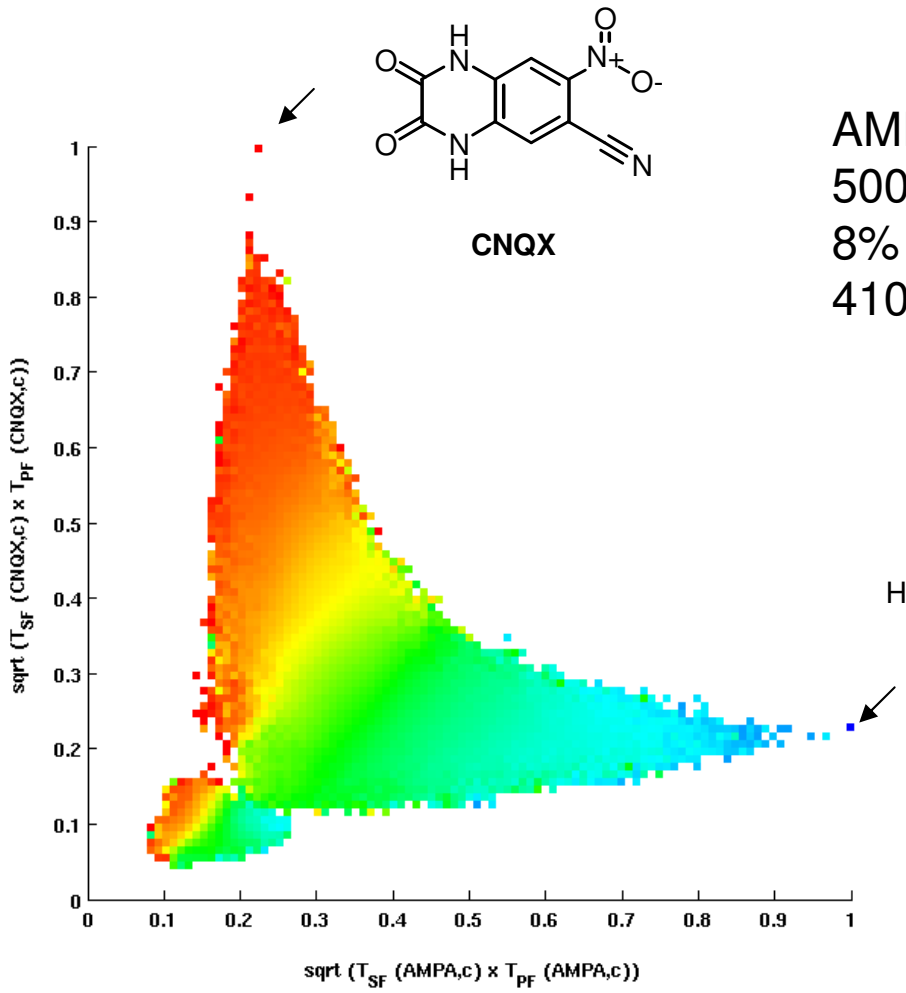
Trajectory Examples (2)

Compound	Formula	mass	n[a]	Steps from CH_4[b] Nearest neighbours	With aromatic[c]	N[d]	Steps to MeOH[b]	N[d]
Tetrodotoxine	$C_{11}H_{17}N_3O_8$	319	22	28	-	106'158	20*	16'757
Sucrose	$C_{12}H_{22}O_{11}$	342	23	25*	-	67'052	21	19'552
Penicillin G	$C_{16}H_{18}N_2O_4S$	334	23	n.f.[e]	20(2)	70'497	23*	15'748
Strychnine	$C_{21}H_{22}N_2O_2$	332	25	n.f.[e]	26(2)	176'721	25	32'479
Papaverin	$C_{20}H_{21}NO_4$	339	25	n.f.[e]	25(3)	53'099	25	28'449
Colchicine	$C_{22}H_{25}NO_6$	399	29	37	32(3)	136'519	28	33'592
Calcitriol	$C_{27}H_{44}O_3$	417	30	37*	-	298'327	28*	65'595
Dipicrylamine	$C_{12}H_5N_7O_{12}$	439	31	n.f.[e]	21(2)	21'015	26	13'950
Tetracycline	$C_{22}H_{24}N_2O_8$	428	31	36	30(1)	173'734	30	34'883
Vitamin K	$C_{31}H_{46}O_2$	451	33	55	42(3)	411'107	32*	77'337
Epothilone	$C_{27}H_{41}NO_6S$	508	35	n.f.[e]	62(4)	709'250	34*	75'219
Vitamin E	$C_{29}H_{50}O_2$	531	38	71	40(2)	443'477	37*	140'017
Reserpine	$C_{33}H_{40}N_2O_9$	609	44	n.f.[e]	68(5)	286'342	62	230'646
Taxotere	$C_{45}H_{55}NO_{15}$	808	58	n.f.[e]	74(4)	1'128'960	57*	304'172

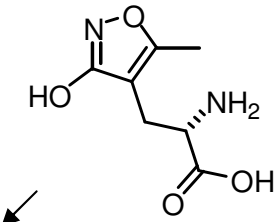
Cross-Trajectories

From: \ To:	Cubane	Aspirine	VX	Adenosine	Sucrose	Penicillin G	Strychnine	Colchicine	Tetracycline	Vitamin K
Cubane	-	10	18	23 (1)	19	18 (1)	18 (1)	22 (1)	24 (1)	26 (1)
Aspirine	10*	-	14	21	15	16	24	22	22	33
VX	13	17 (1)	-	31 (1)	18	15 (1)	21 (1)	20 (2)	24* (1)	25* (1)
Adenosine	17*	27	18*	-	14	15	24	23	27*	29
Sucrose	18*	22 (1)	22*	29 (1)	-	25	26 (1)	31 (1)	25 (1)	25 (1)
Penicillin G	19*	13*	14*	23	19*	-	20	19*	21*	29
Strychnine	21*	17*	20	26	22	16*	-	30*	17*	22*
Colchicine	27	22*	21	26	18	22	23	-	22*	21*
Tetracycline	28*	20	25*	49	19	19*	16	28	-	17
Vitamin K	30*	24*	30*	34*	28*	27*	19*	30*	22*	-

AMPA ↔ CNQX

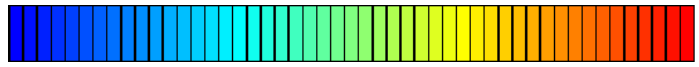


CNQX

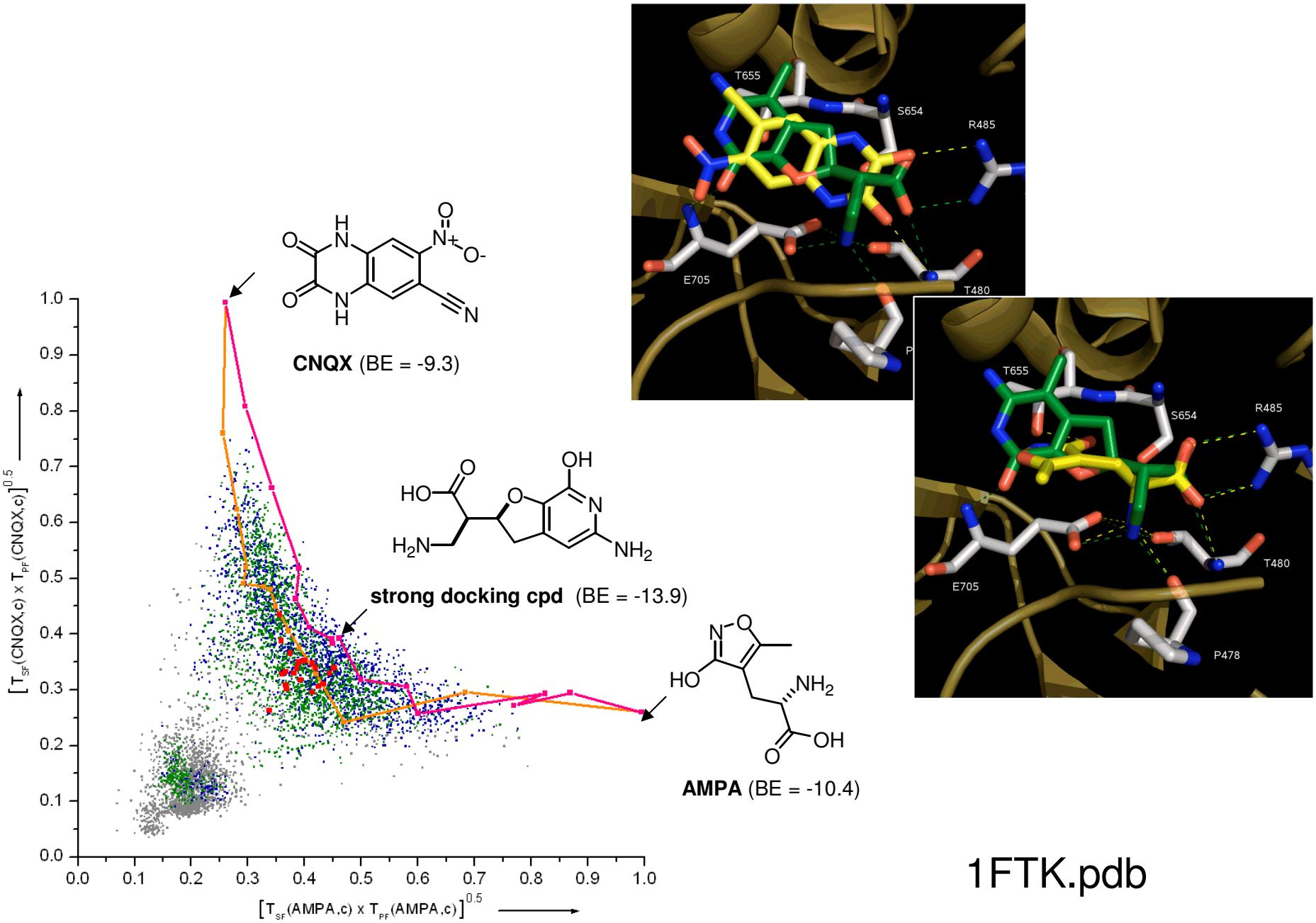


AMPA

AMPA ↔ CNQX
500 runs, 559,658 cpds
8% from 6.9 million generated
410 cpds (0.1%) also in ZINC

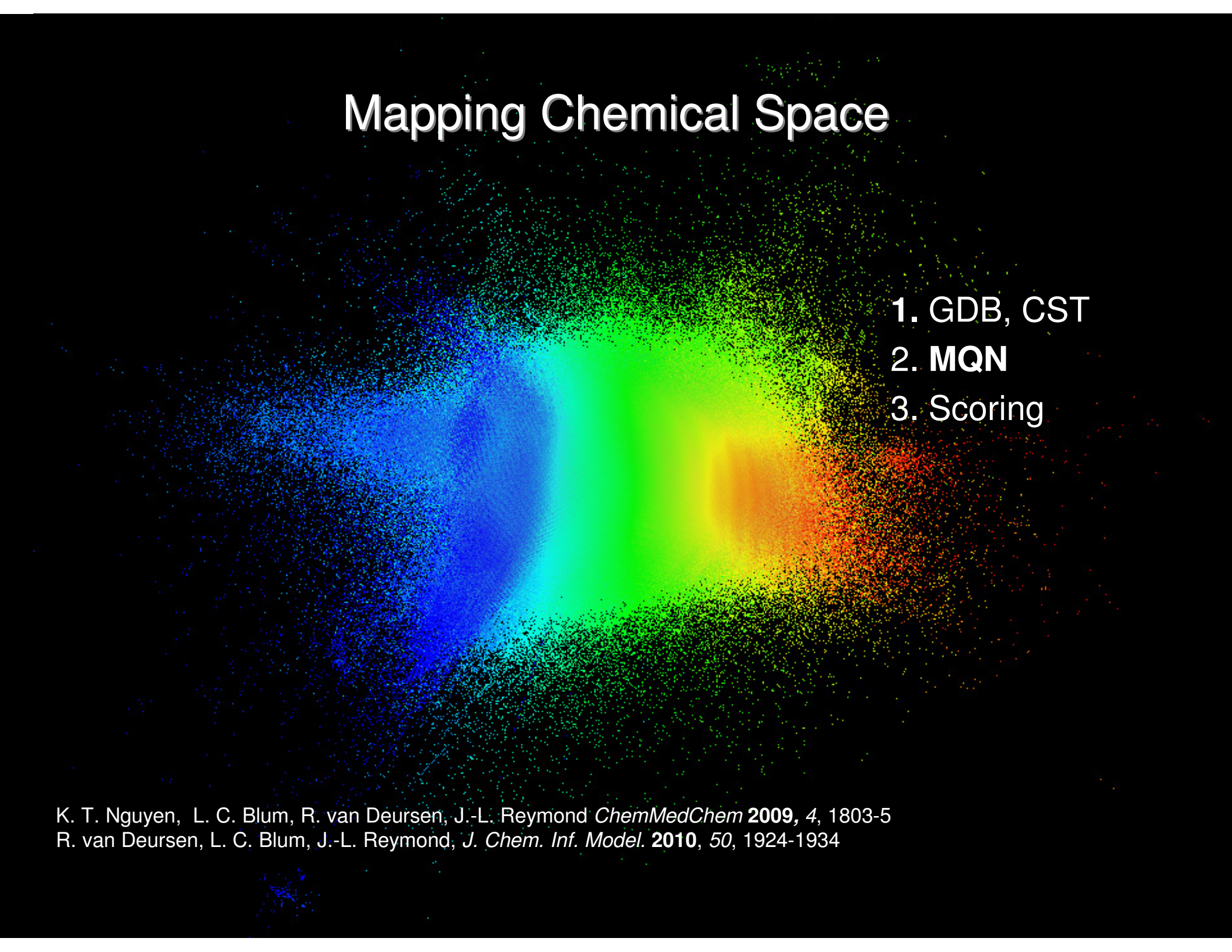


0 **0.50** **1.0**


T655
S654
R485
E705
T480
P478
CNQX (BE = -9.3)
strong docking cpd (BE = -13.9)
AMPA (BE = -10.4)
[T_SF(CNQX,c) x T_PF(CNQX,c)]^0.5
[T_SF(AMPA,c) x T_PF(AMPA,c)]^0.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

1FTK.pdb

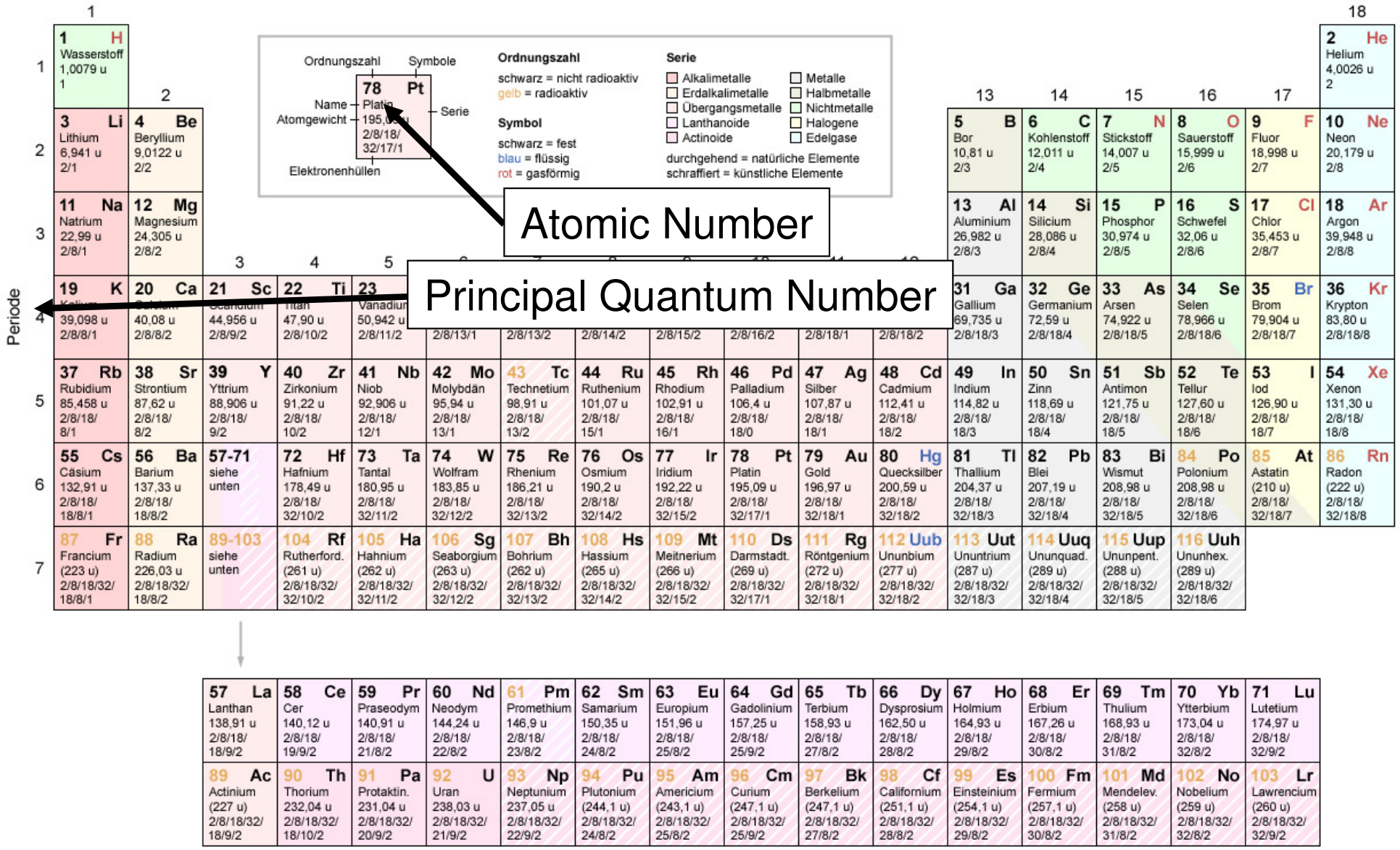
Mapping Chemical Space



1. GDB, CST
2. **MQN**
3. Scoring

K. T. Nguyen, L. C. Blum, R. van Deursen, J.-L. Reymond *ChemMedChem* **2009**, *4*, 1803-5
R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924-1934

The Periodic System of the Elements



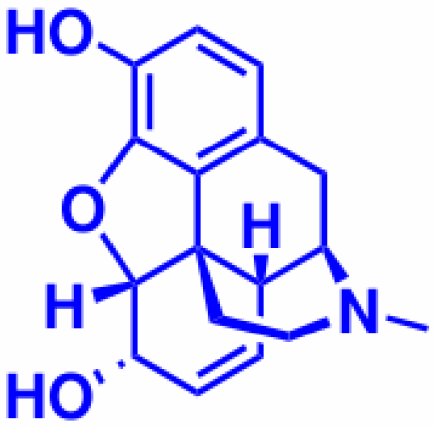
Molecular Quantum Numbers

Atom counts		
Carbon	17	14
Fluorine	0	0
Chlorine	0	0
Bromine	0	0
Iodine	0	0
Sulphur	0	1
Phosphorus	0	0
Acyclic nitrogen	0	1
Cyclic nitrogen	1	1
Acyclic oxygen	2	4
Cyclic oxygen	1	1
Heavy atom count	21	21

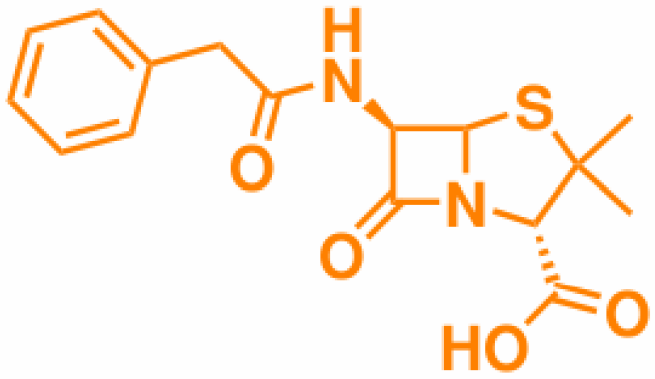
Polarity counts		
H-Bond donor atoms	3	1
H-Bond donor sites	3	1
H-Bond acc. atoms	3	4
H-Bond acc. sites	3	7
Positive charges	1	0
Negative charges	0	1

Bond counts		
Acyclic single bonds	3	6
Acyclic double bonds	0	3
Acyclic triple bonds	0	0
Cyclic single bonds	18	11
Cyclic double bonds	4	3
Cyclic triple bonds	0	0
Rotatable bonds	0	4

Topology counts		
Acyclic monovalent nodes	3	4
Acyclic divalent nodes	0	2
Acyclic trivalent nodes	0	2
Acyclic tetravalent nodes	0	0
Cyclic divalent nodes	8	7
Cyclic trivalent nodes	9	6
Cyclic tetravalent nodes	1	0
3-Membered rings	0	0
4-Membered rings	0	1
5-Membered rings	1	1
6-Membered rings	4	1
7-Membered rings	0	0
8-Membered rings	0	0
9-Membered rings	0	0
≥10-Membered rings	0	0
Atoms shared by fused rings	7	2
Bonds shared by fused rings	6	1

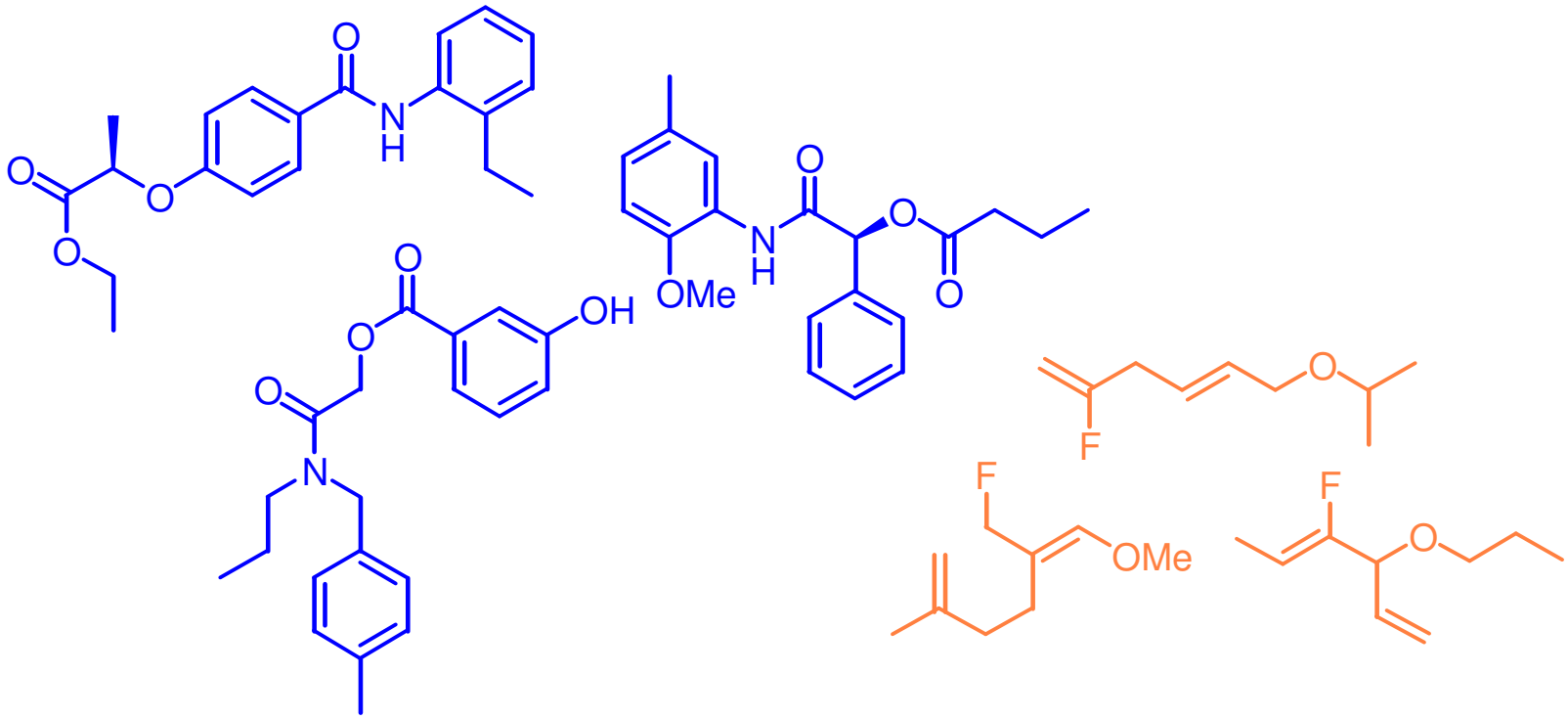


Morphine

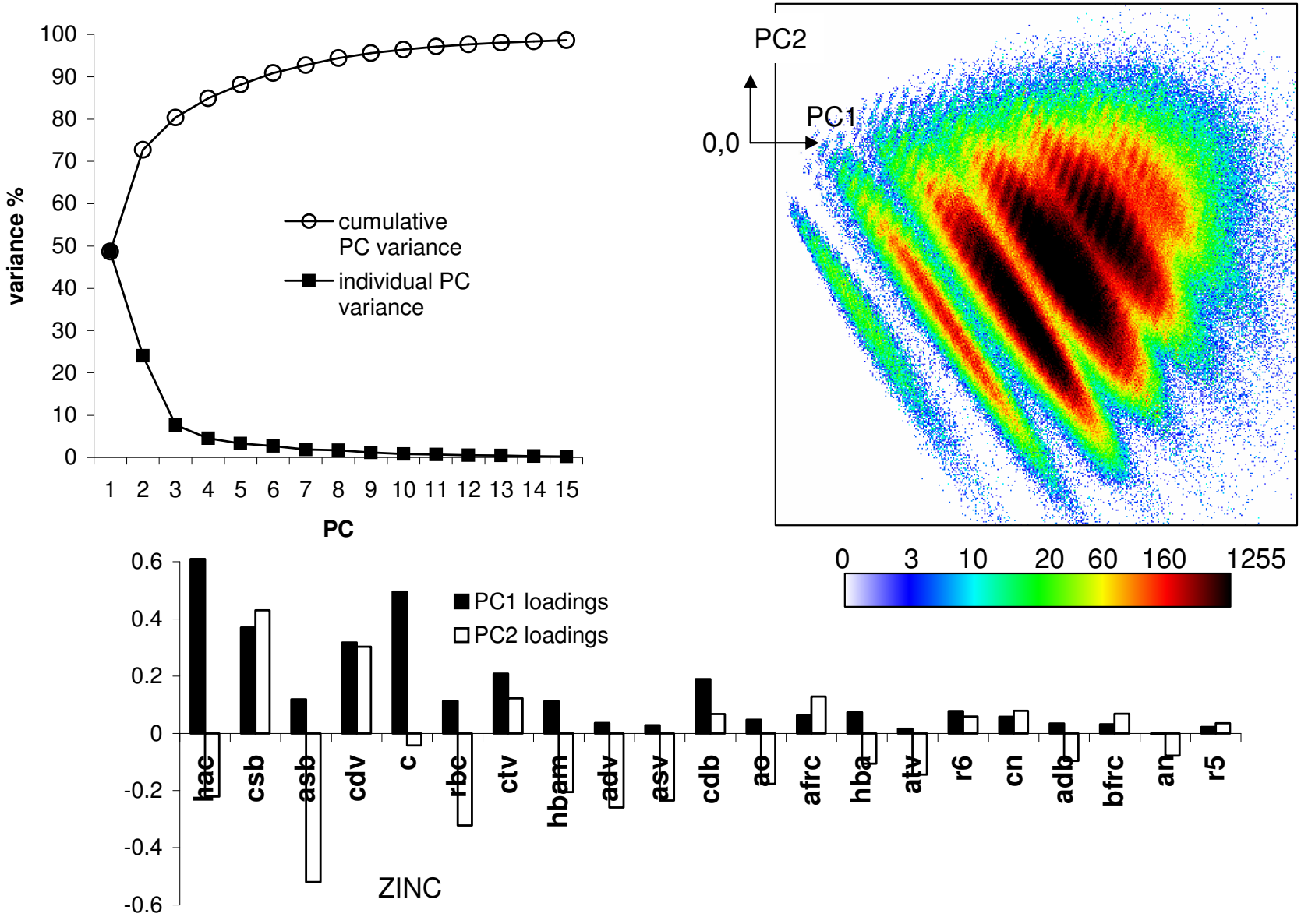


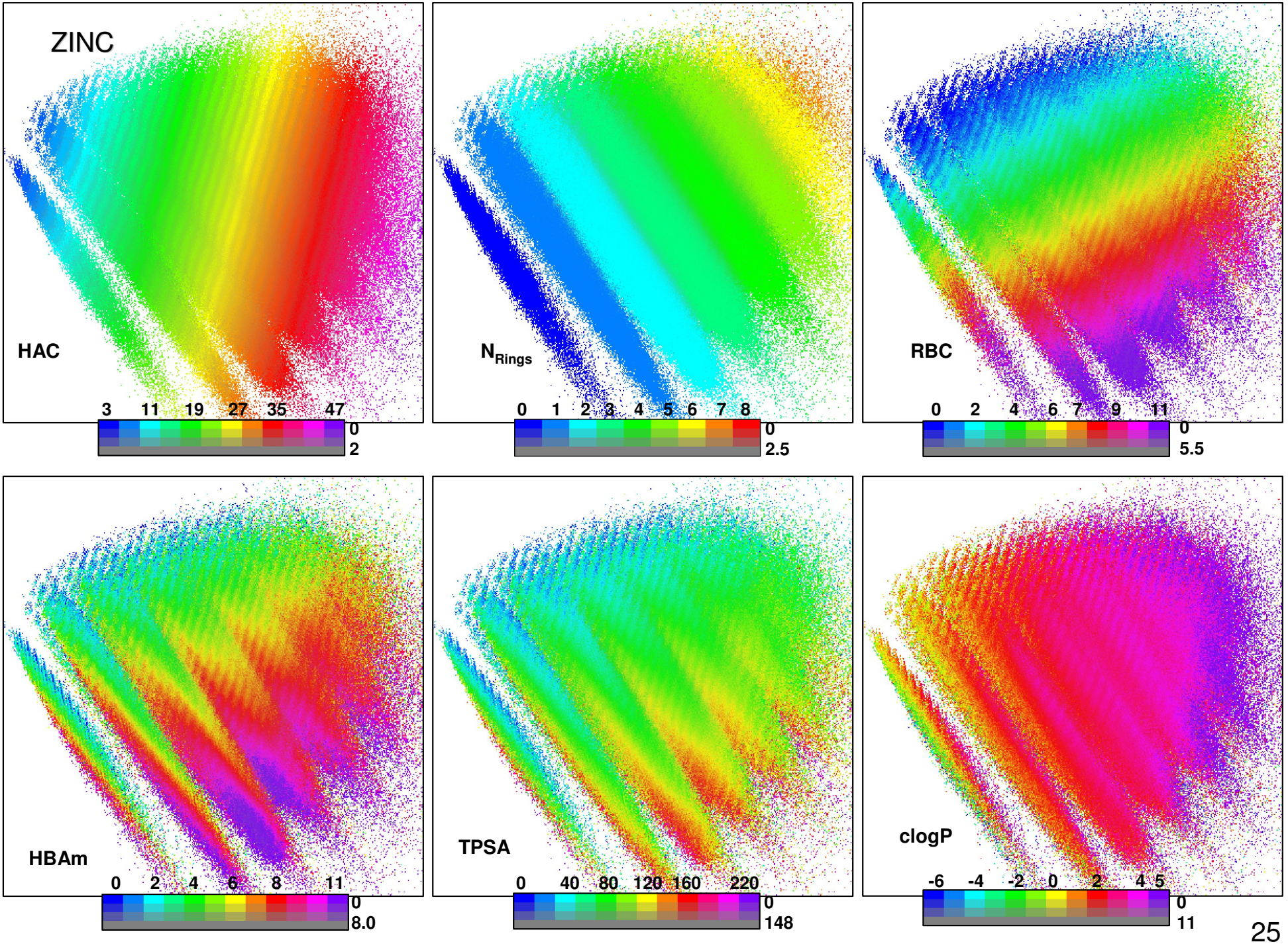
Penicillin G

	ZINC	GDB-11
no. of cpds	8 436 272	26 434 567
no. of MQN-bins	3 654 836	2 859 938
no. of single occupied MQN-bins	1 832 566	660 851
no. of cpds in most occupied MQN-bin	300	1 982
no. of shared MQN-bins	13 769	13 769
no. of cpds in shared MQN-bins	30 779	254 604

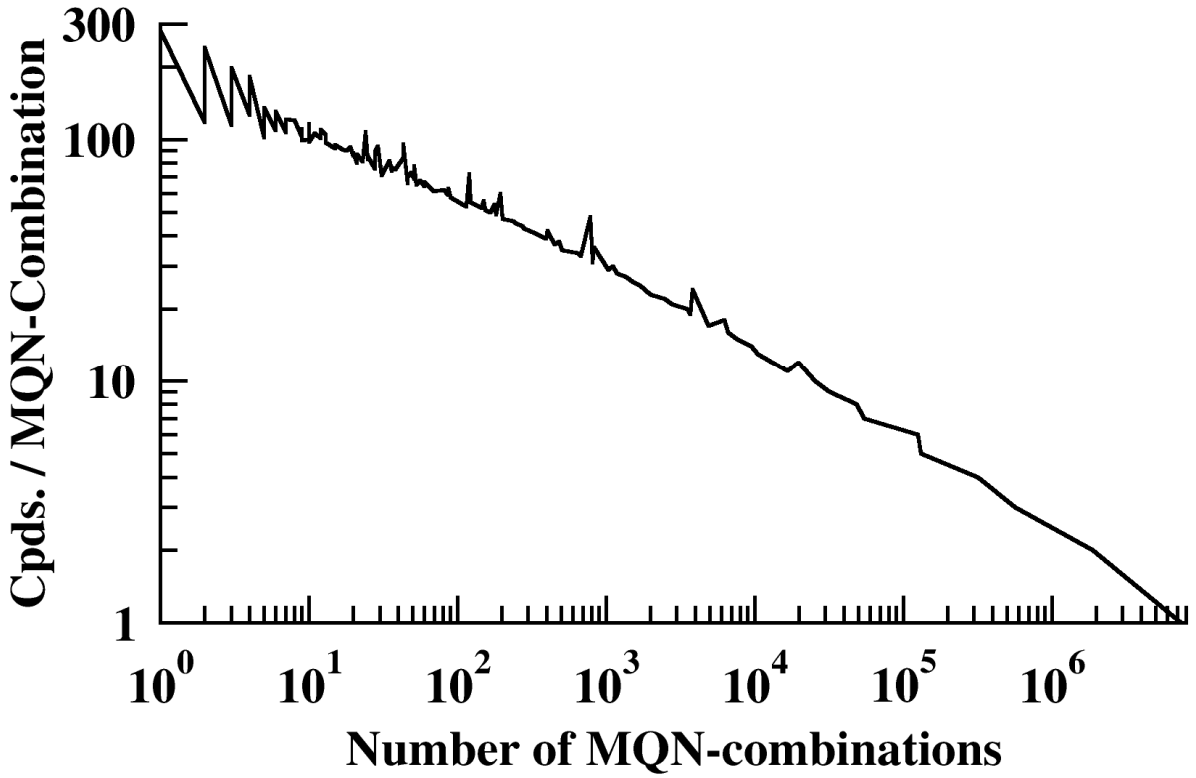


ZINC (8.4 million SMILES)

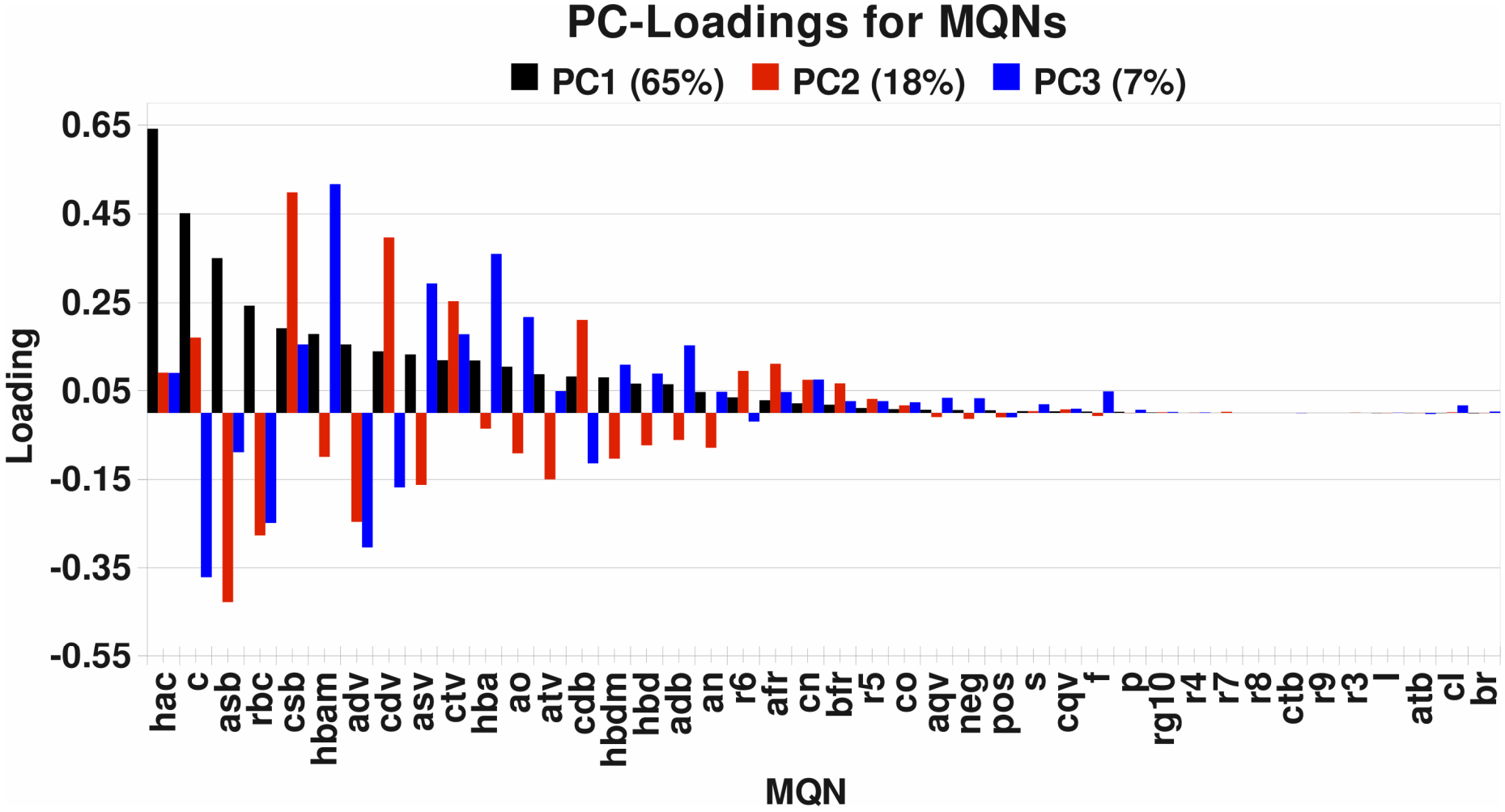



ZINC
HAC
3 11 19 27 35 47
0
2
N Rings
0 1 2 3 4 5 6 7 8
0
2.5
RBC
0 2 4 6 7 9 11
0
5.5
HBAm
0 2 4 6 8 11
0
8.0
TPSA
0 40 80 120 160 220
0
148
clogP
-6 -4 -2 0 2 4 5
0
11

PubChem (19.2 million SMILES)

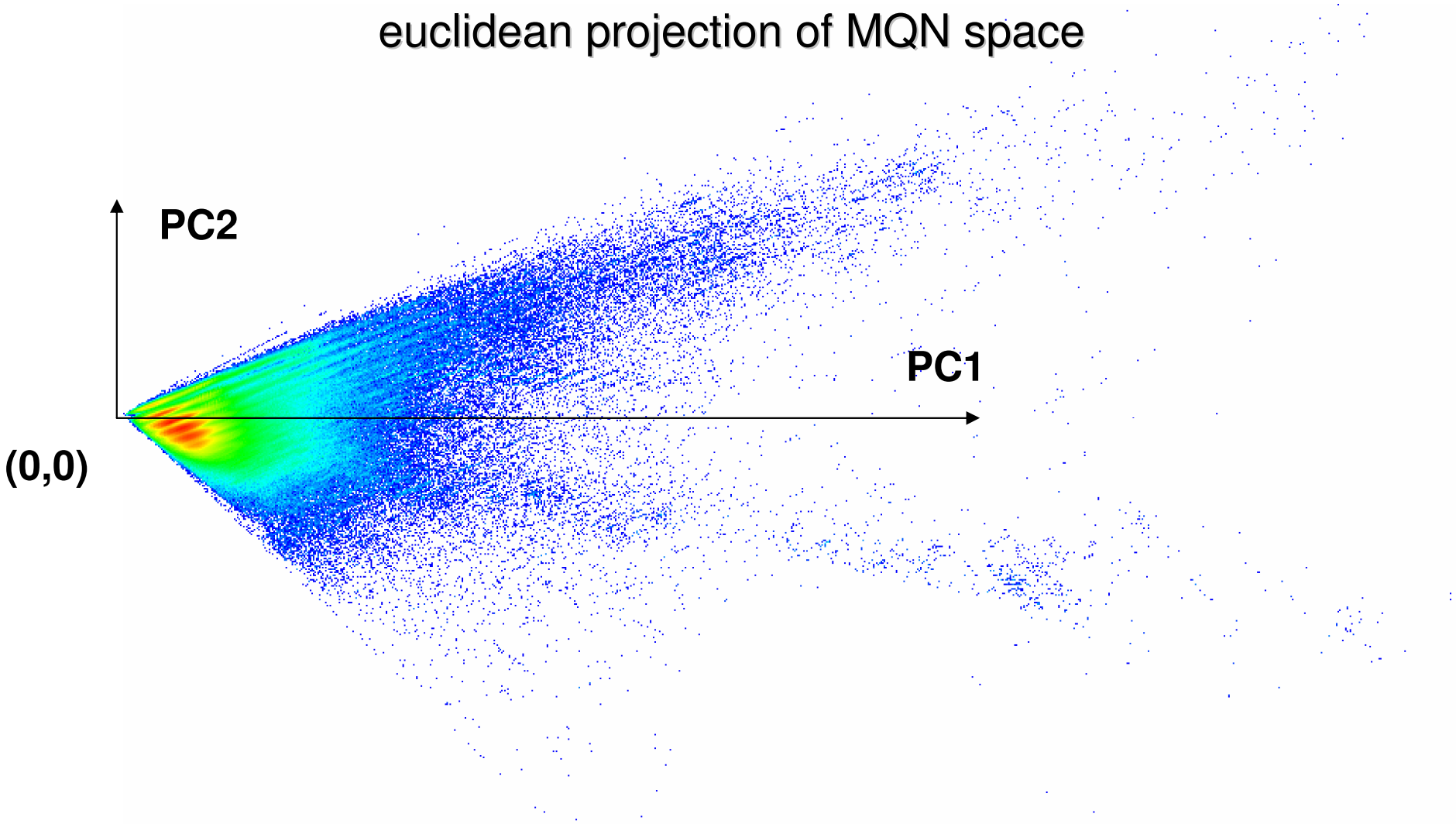


R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924-1934


PC-Loadings for MQNs
PC1 (65%)
PC2 (18%)
PC3 (7%)
Loading
0.65
0.45
0.25
0.05
-0.15
-0.35
-0.55
hac
c
asb
rbc
csb
hbam
adv
cdv
asv
ctv
hba
ao
atv
cdb
hbdm
hbd
adb
an
r6
afr
cn
bfr
r5
co
aqv
neg
pos
s
cqv
f
p
rg10
r4
r7
r8
ctb
r9
r3
l
atb
cl
br
MQN

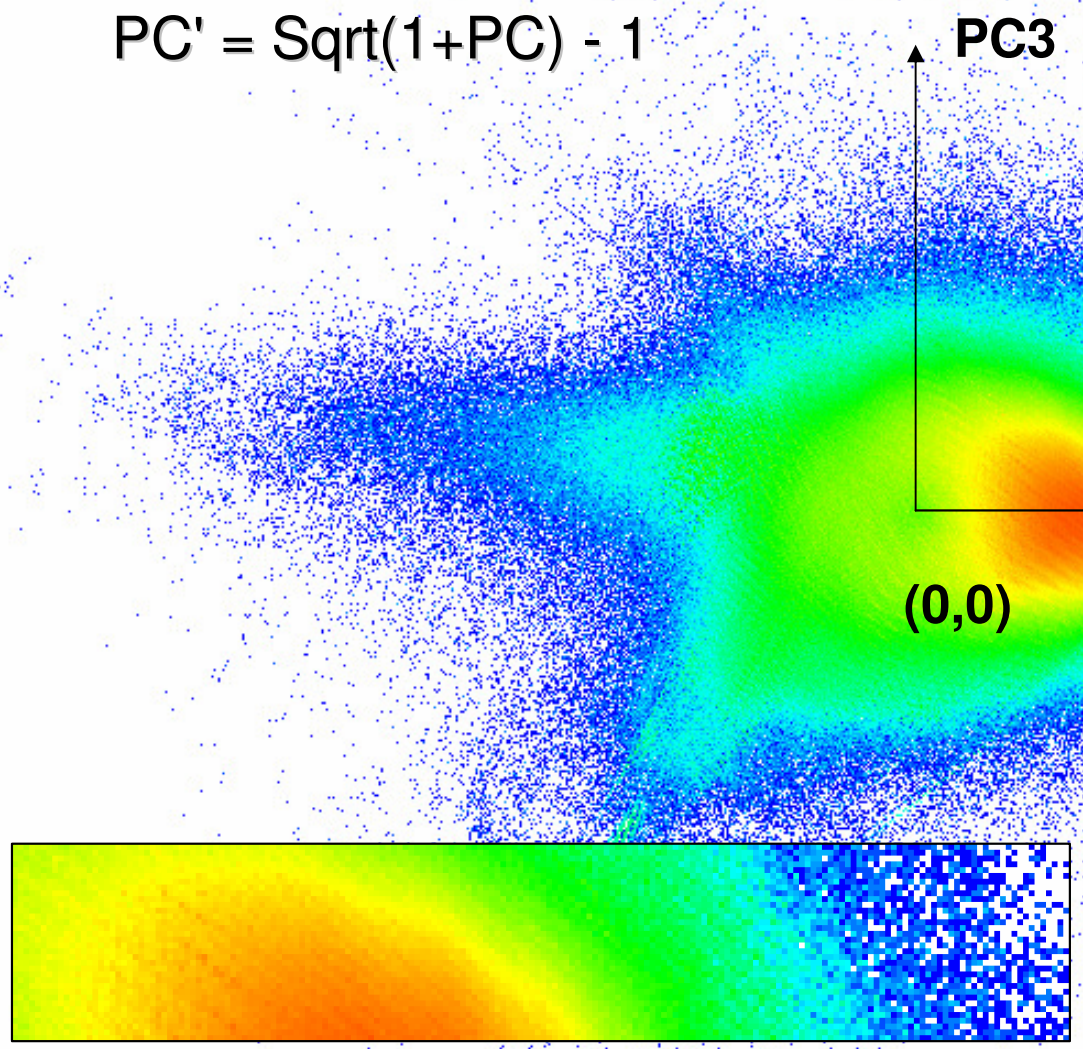
Frequency map

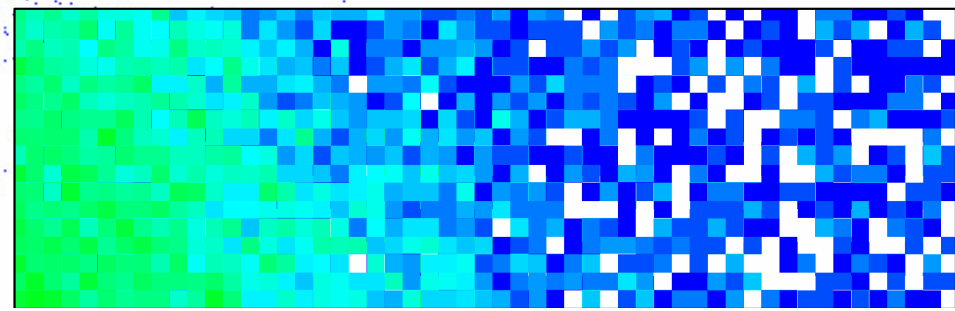
euclidean projection of MQN space

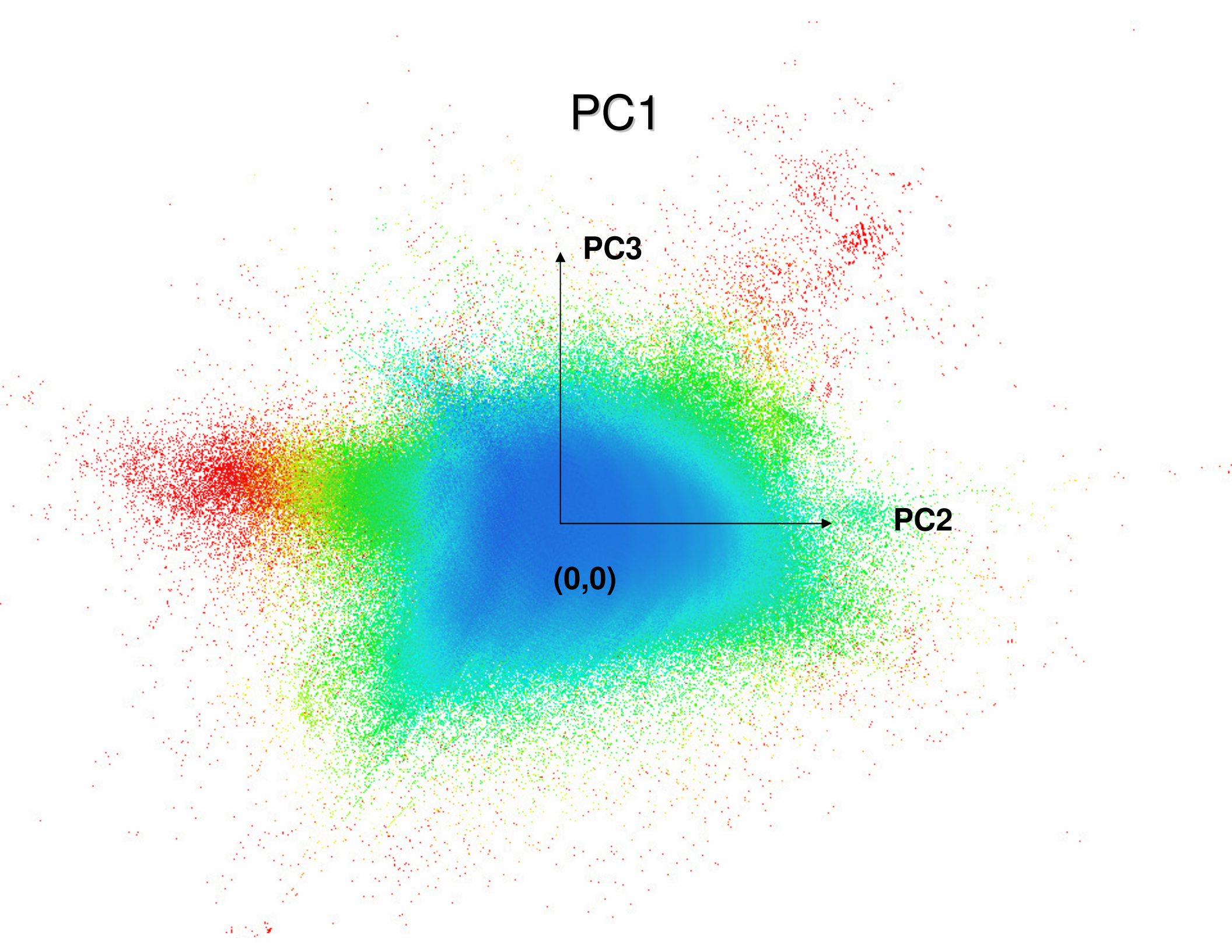


Frequency map

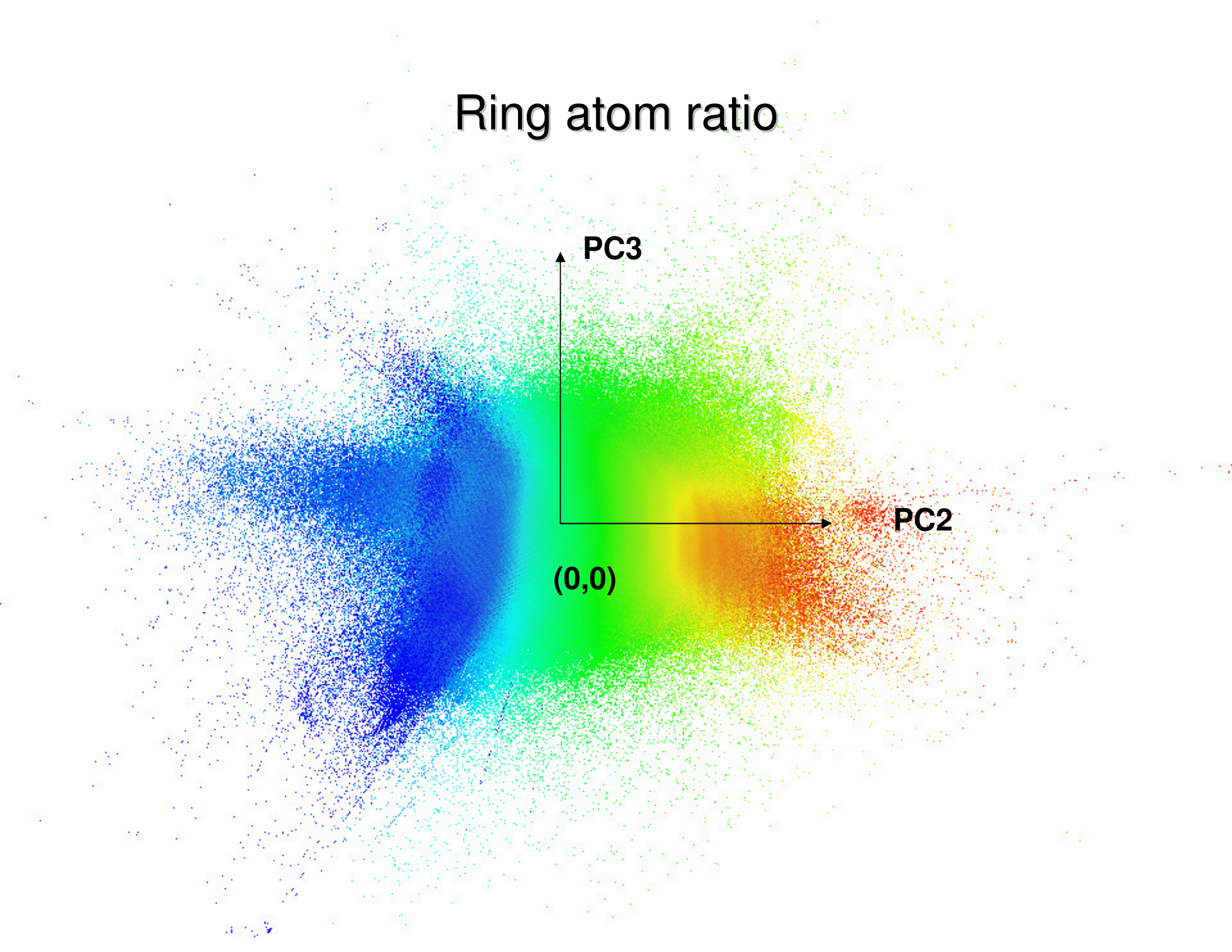
PC' = Sqrt(1+PC) - 1



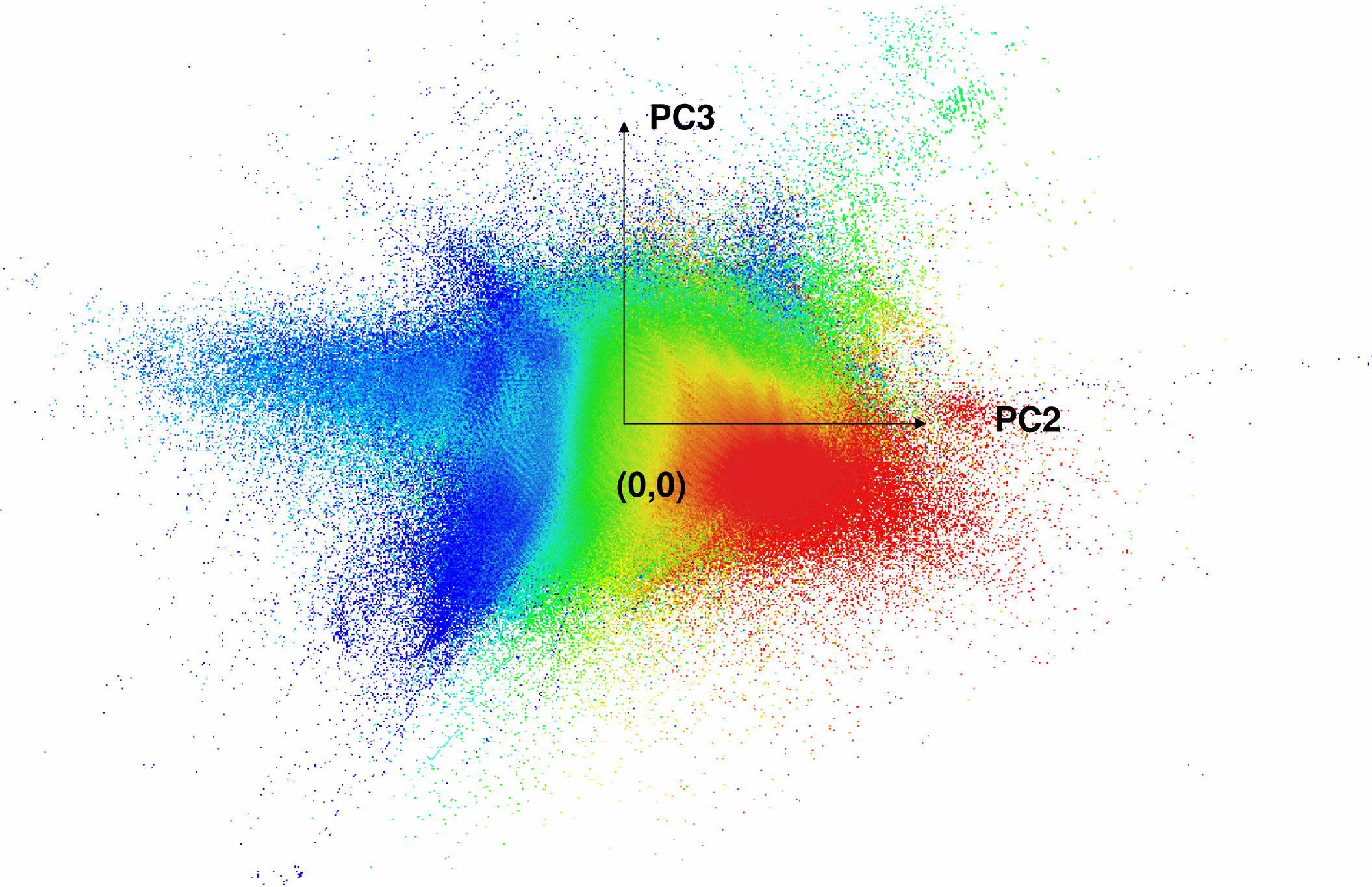



PC1
PC3
PC2
(0,0)

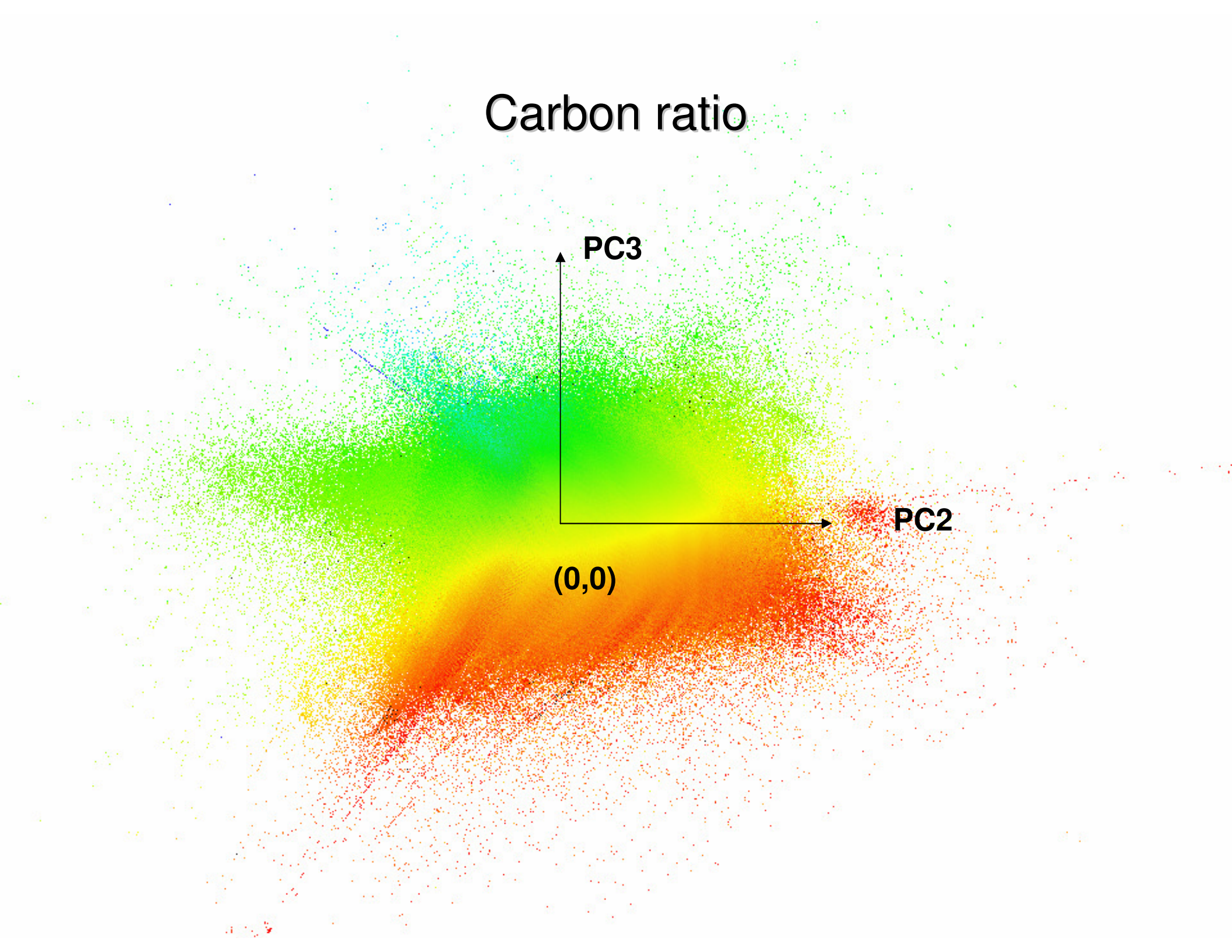
Ring atom ratio



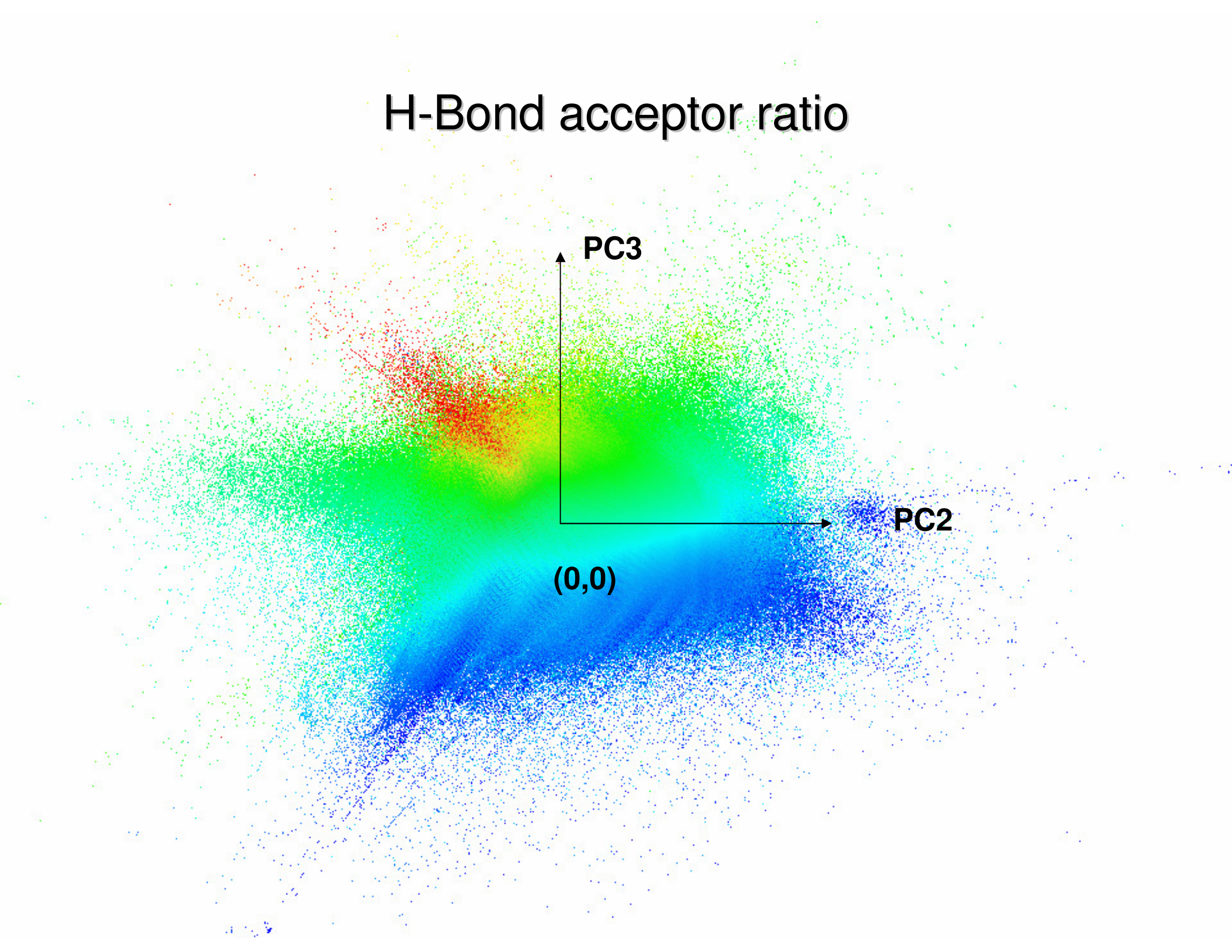
Cyclic double bond ratio



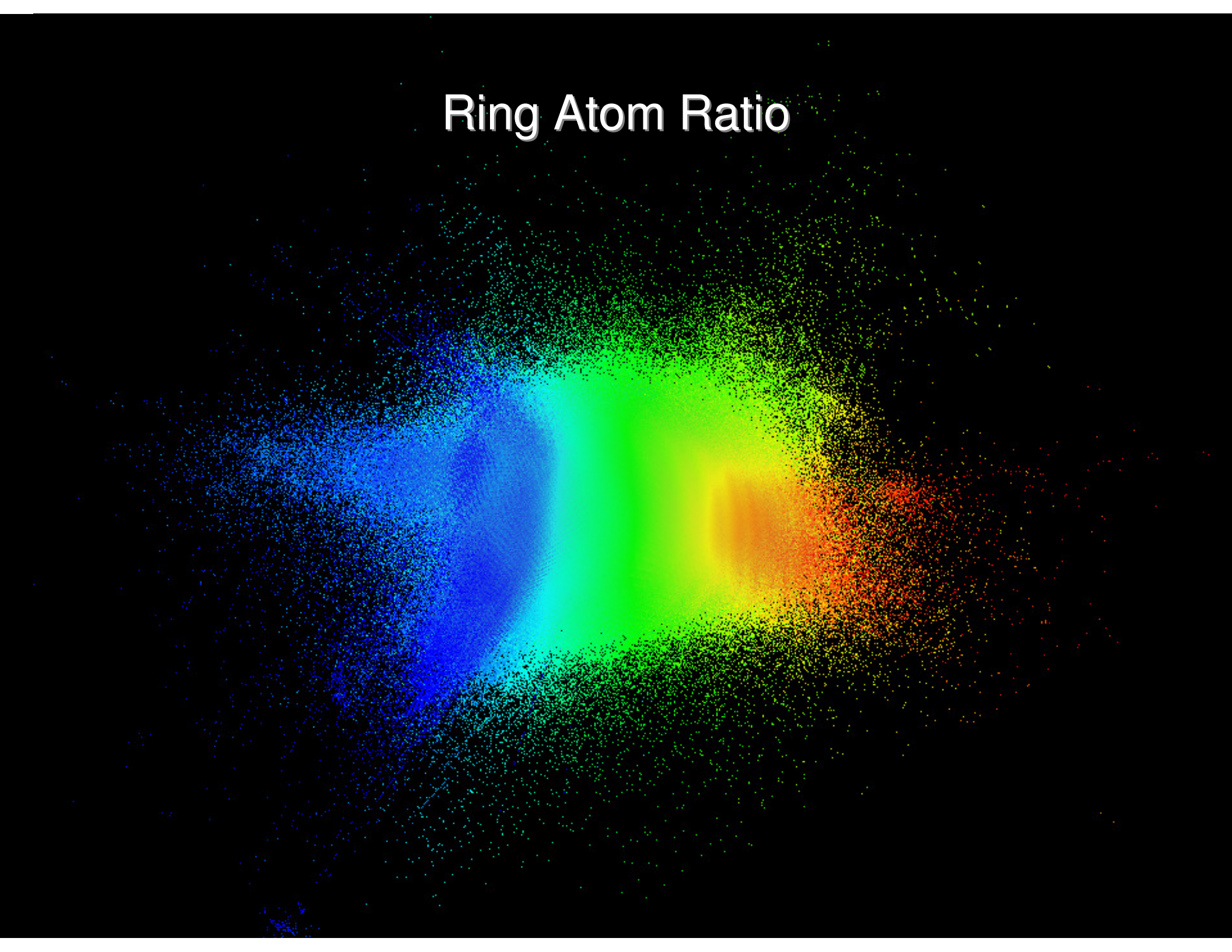
Carbon ratio

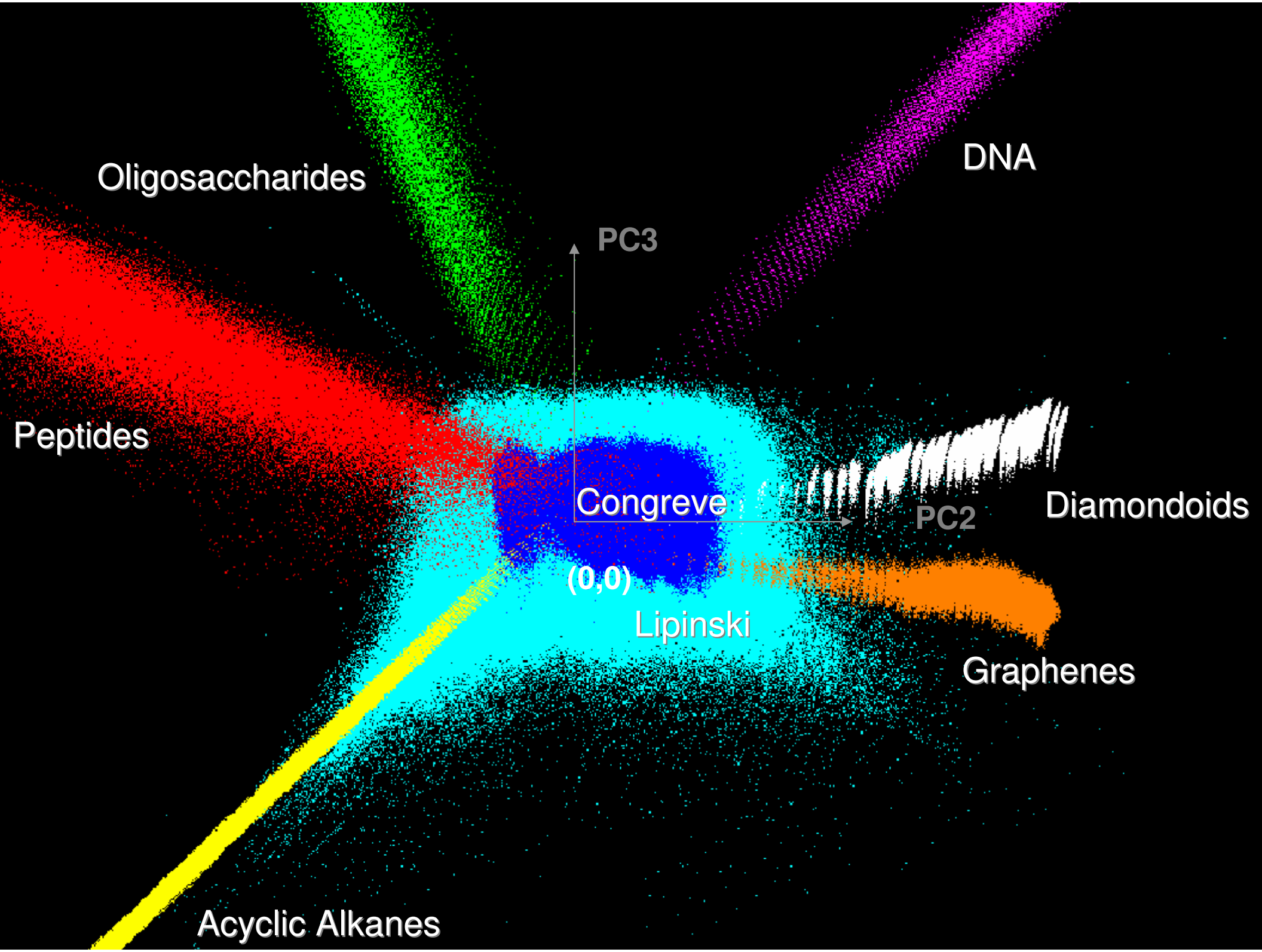


H-Bond acceptor ratio

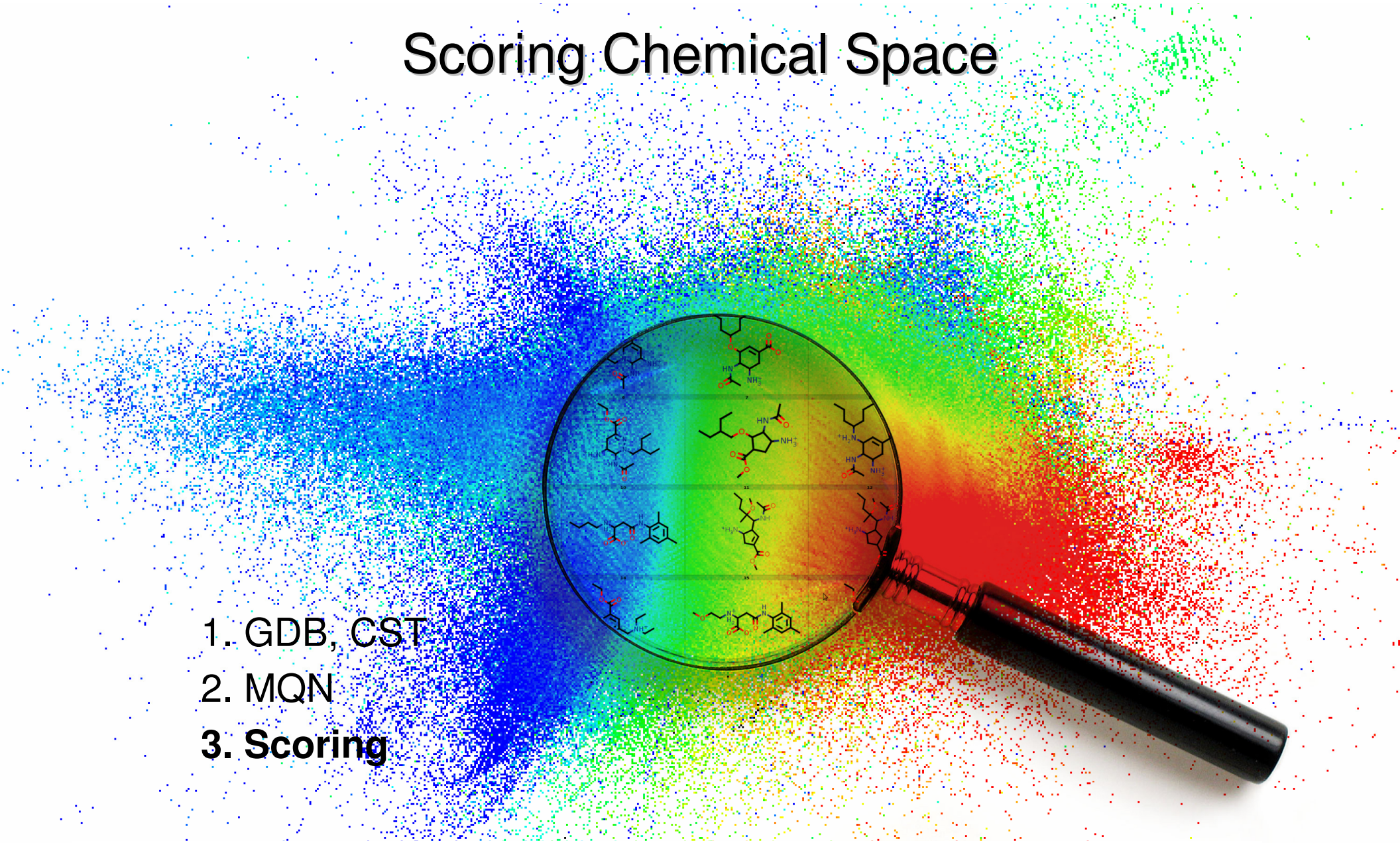


Ring Atom Ratio




Oligosaccharides
DNA
PC3
Peptides
Congreve
PC2
Diamondoids
(0,0)
Lipinski
Graphenes
Acyclic Alkanes

Scoring Chemical Space



1. GDB, CST
2. MQN
3. **Scoring**

R. van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924-1934

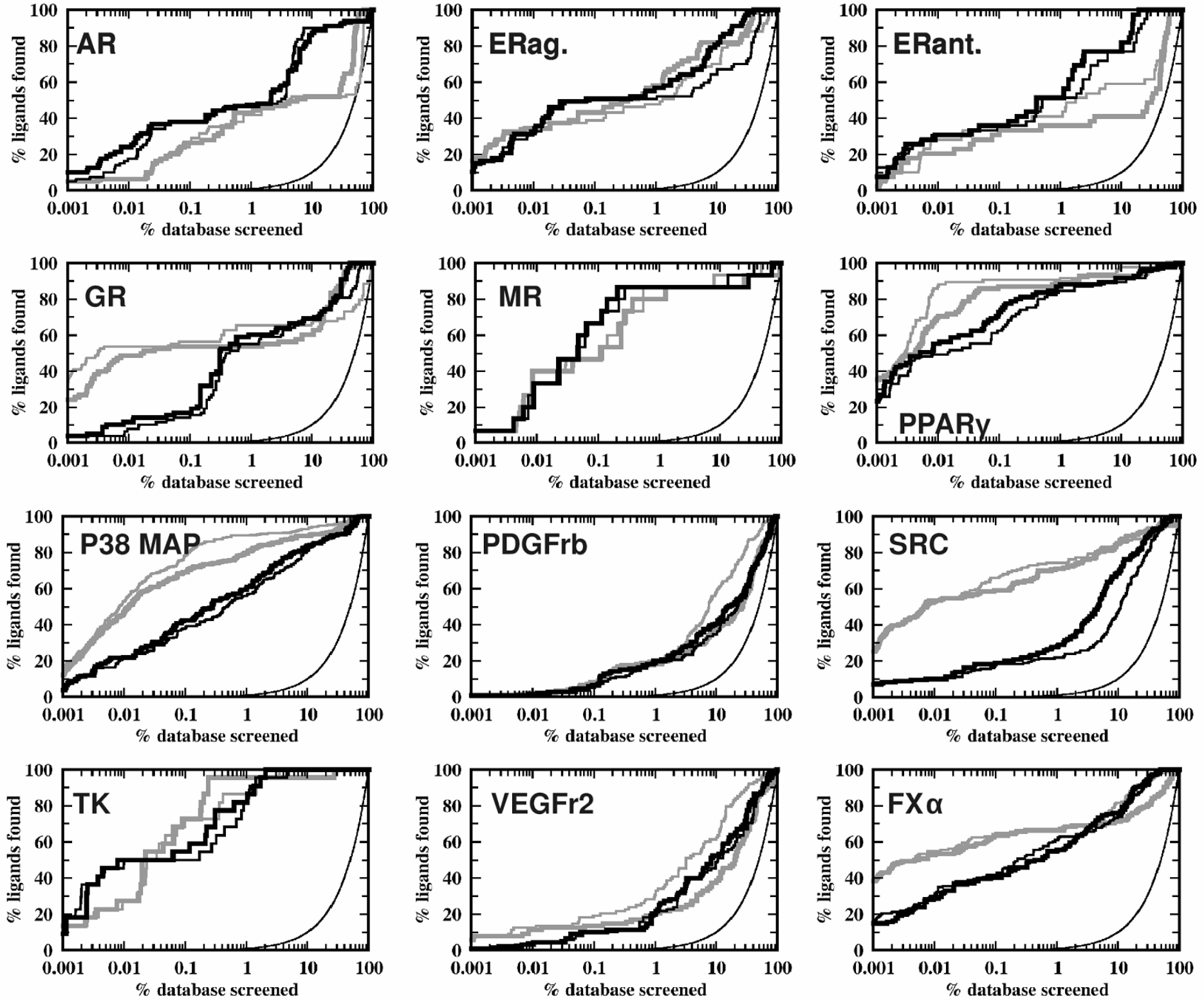
Lorenz Blum (GDB subsets)

E. Luethi et al., *J. Med. Chem.* **2010**, *53*, 7236, and N. Garcia-Delgado et al., *ACS Med. Chem. Lett.* **2010**, *1*, 422-426

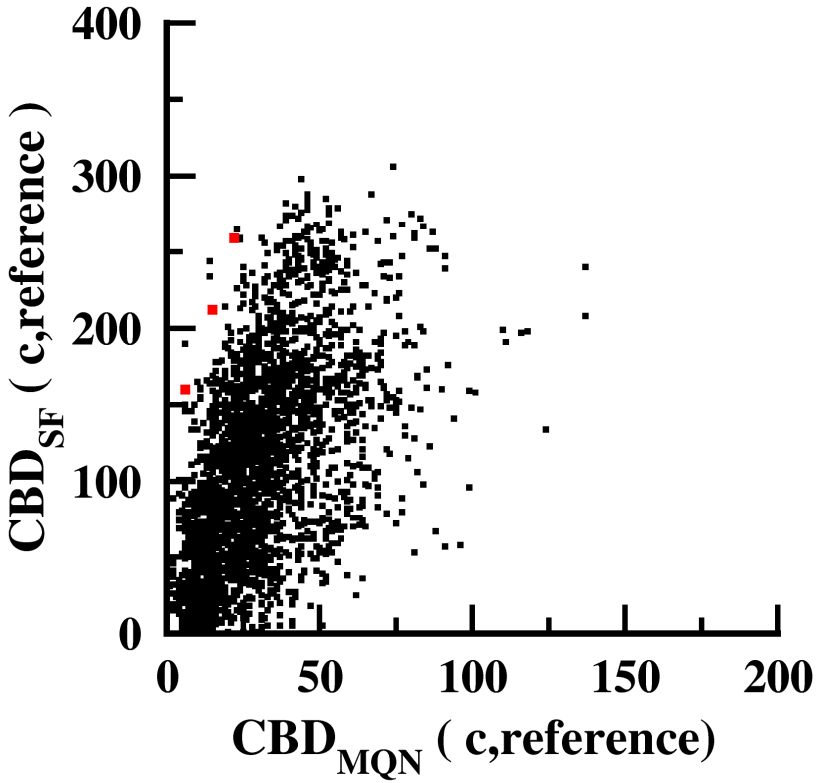
Enriching the DUD actives from Pubchem

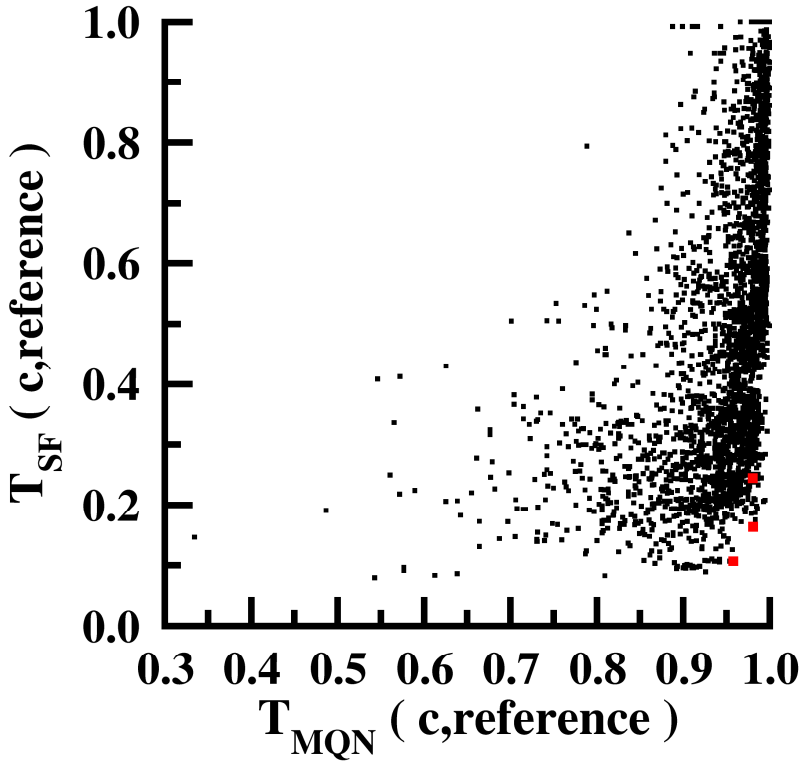
	nr. of actives[b)]	$EF_{0.1}$ CBD_{MQN}	T_{MQN}	CBD_{SF}	T_{SF}	EF_1 CBD_{MQN}	T_{MQN}	CBD_{SF}	T_{SF}
Nuclear Hormone Receptors									
AR	79	379.5	379.5	265.6	265.6	46.8	48.1	43.0	41.8
ERagonist	67	507.1	507.1	432.5	387.8	56.7	50.7	58.2	47.8
ERantagonist	39	358.7	358.7	333.1	333.1	51.3	51.3	35.9	41.0
GR	78	166.6	140.9	538.1	563.7	60.2	55.1	53.8	65.4
MR	15	666.2	666.2	466.3	466.3	86.7	86.7	80.0	86.7
PPARg	85	728.9	623.1	870.0	905.3	87.0	84.7	89.4	91.7
PR	27	592.2	555.2	592.2	629.2	59.2	59.2	59.2	70.4
RXRa	20	849.4	599.6	849.4	849.4	95.0	85.0	85.0	100.0
Kinases									
CDK2	72	111.0	83.3	138.8	138.8	20.8	18.1	15.3	16.7
EGFR	475	90.5	67.3	126.2	132.5	25.5	20.8	20.2	27.2
FGFr1	120	191.5	183.2	624.6	724.5	29.2	22.5	74.2	85.0
HSP90	37	378.1	378.1	648.2	702.2	54.0	37.8	70.3	70.3
P38 MAP	454	424.8	380.8	691.2	783.6	59.5	55.9	79.5	89.4
PDGFrb	170	64.7	52.9	82.3	82.3	20.0	18.2	18.2	20.0
SRC	159	188.5	182.3	590.8	659.9	27.7	22.0	71.1	74.2
TK	22	545.1	499.7	726.8	726.8	81.8	81.8	95.4	86.3
VEGFr2	88	102.2	102.2	136.3	193.0	20.5	19.3	20.5	30.7

Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006,** *49*, 6789-6801.

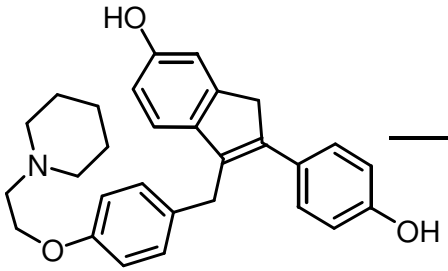

AR
ERag.
ERant.
GR
MR
PPARγ
P38 MAP
PDGFrb
SRC
TK
VEGFr2
FXα
% ligands found
% database screened
0
20
40
60
80
100
0.001
0.01
0.1
1
10
100

Lead Hop(p)ing

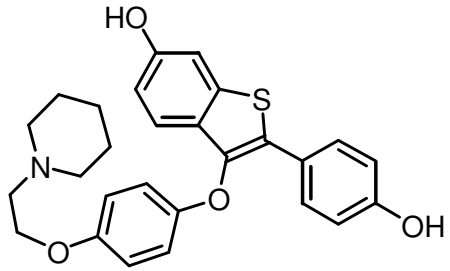




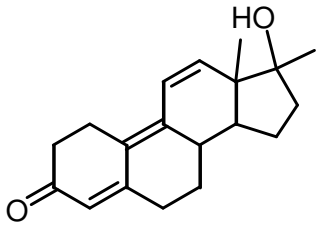
Lead Hop(p)ing



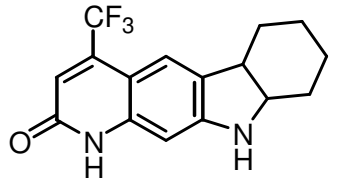
Lead ER antagonist



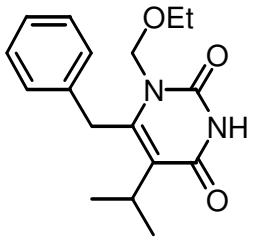
CBD_{MQN} = 6
(CBD_{SF}= 160)



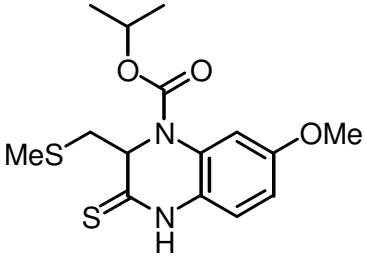
Lead AR inhibitor



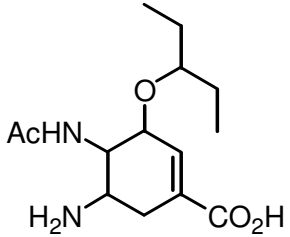
T_{MQN} = 0.958
(T_{SF} = 0.107)



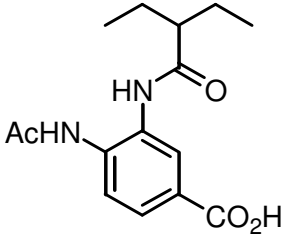
Lead HIV RT inhibitor



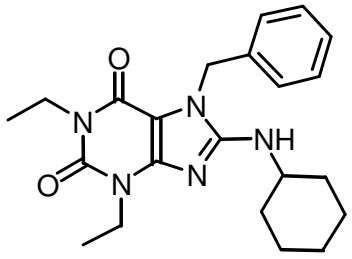
CBD_{MQN} = 15
(CBD_{SF}= 212)



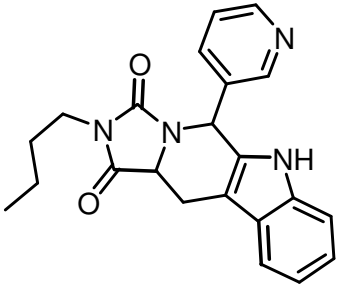
Lead NA inhibitor



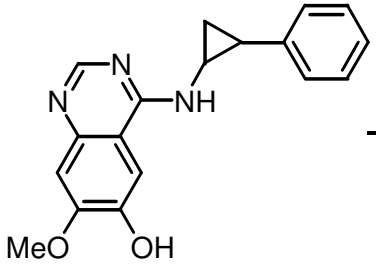
T_{MQN} = 0.981
(T_{SF} = 0.163)



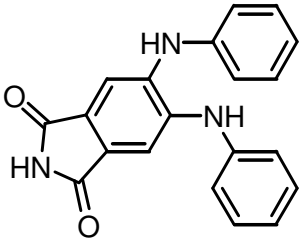
Lead pde5 inhibitor



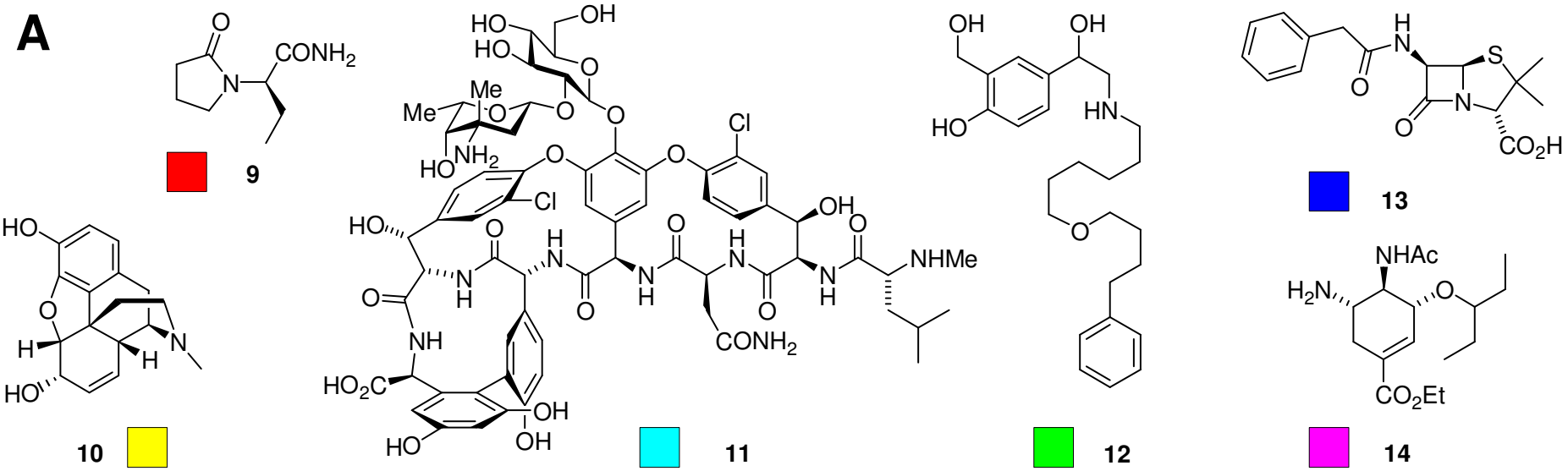
CBD_{MQN} = 22
(CBD_{SF} = 259)

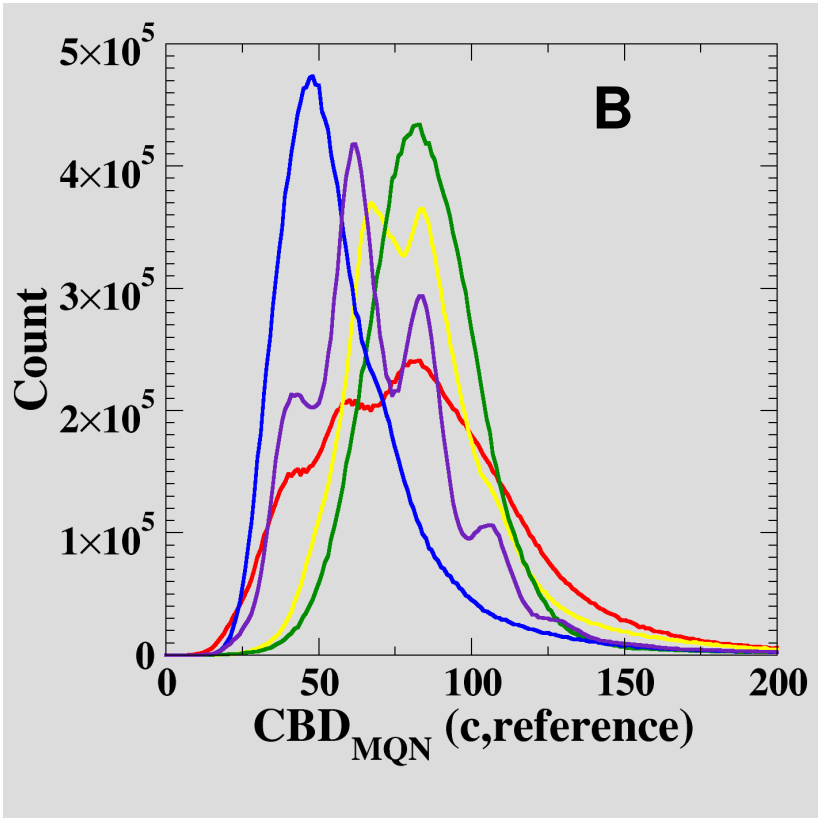


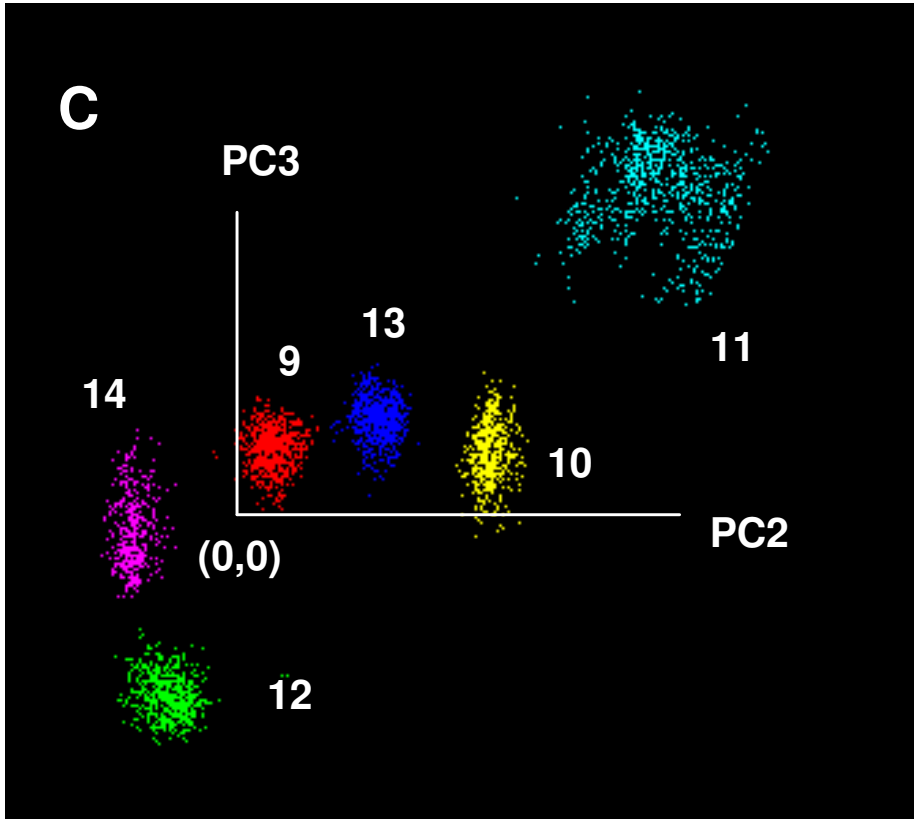
Lead EGFR inhibitor



T_{MQN} = 0.980
(T_{SF} = 0.244)


A
9
10
11
12
13
14


B
Count
5×10^5
4×10^5
3×10^5
2×10^5
1×10^5
0
0
50
100
150
200
CBD_MQN (c,reference)


C
PC3
PC2
(0,0)
9
10
11
12
13
14