PubChem Substructure Fingerprint

V1.3 http://pubchem.ncbi.nlm.nih.gov

The PubChem System generates a binary substructure fingerprint for chemical structures. These fingerprints are used by PubChem for similarity neighboring and similarity searching.

A substructure is a fragment of a chemical structure. A fingerprint is an ordered list of binary (1/0) bits. Each bit represents a Boolean determination of, or test for, the presence of, for example, an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure.

The native format of the PubChem Substructure Fingerprint property is binary data with a four byte integer prefix, where this integer prefix indicates the length of the bit list. For the ASN.1 and XML formatted data, this property is stored in a PC-InfoData container, as described by the PCSubstance ASN.1 definition or XML schema: ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/

PC-InfoData is able to handle various types of data. Each PC-InfoData has a PC-Urn object (urn = universal resource name). Each property has a unique trio of "label", "name", and "datatype" definition (e.g., for PubChem Substructure Fingerprint, this is "Fingerprint", "SubStructure Keys", and "fingerprint", respectively). The fingerprint binary data is hex-encoded, when provided in the XML or textual ASN.1 formats.

When exporting fingerprint information in the SD file format, the SD tag for the PubChem Substructure Fingerprint property is "PUBCHEM_CACTVS_SUBGRAPHKEYS". The PubChem Substructure Fingerprint is Base64 encoded to provide a textual representation of the binary data. For a description of the Base64 encoding and decoding algorithm specification, go to: http://www.faqs.org/rfcs/rfc3548.html

Below is the description of each bit represented in the PubChem Substructure Fingerprint. Some fingerprint bit descriptions are written in SMILES or SMARTS notation. For additional information on SMARTS and SMILES, please go to: http://en.wikipedia.org/wiki/Simplified_molecular_input_line_entry_specification

PubChem Substructure Fingerprint Description

Section 1: Hierarchic Element Counts - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.

Bit Position	Bit Substructure
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
6	>= 1 B
7	>= 2 B
8	>= 4 B
9	>= 2 C
10	>= 4 C
11	>= 8 C
12	>= 16 C
13	>= 32 C
14	>= 1 N
15	>= 2 N
16	>= 4 N
17	>= 8 N
18	>= 1 O
19	>= 2 O
20	>= 4 O
21	>= 8 O
22	>= 16 O
23	>= 1 F
24	>= 2 F
25	>= 4 F
26	>= 1 Na
27	>= 2 Na
28	>= 1 Si
29	>= 2 Si
30	>= 1 P
31	>= 2 P
32	>= 4 P
33	>= 1 S
34	>= 2 S
35	>= 4 S
36	>= 8 S
37	>= 1 Cl
38	>= 2 Cl
39	>= 4 Cl
40	>= 8 Cl
41	>= 1 K
42	>= 2 K
43	>= 1 Br
44	>= 2 Br
45	>= 4 Br
46	>= 1 I
47	>= 2 I
48	>= 4 I

PubChem Substructure Fingerprint Description (cont.)

Section 1: Hierarchic Element Counts (cont.)

Bit Position	Bit Substructure
49	>= 1 Be
50	>= 1 Mg
51	>= 1 Al
52	>= 1 Ca
53	>= 1 Sc
54	>= 1 Ti
55	>= 1 V
56	>= 1 Cr
57	>= 1 Mn
58	>= 1 Fe
59	>= 1 Co
60	>= 1 Ni
61	>= 1 Cu
62	>= 1 Zn
63	>= 1 Ga
64	>= 1 Ge
65	>= 1 As
66	>= 1 Se
67	>= 1 Kr
68	>= 1 Rb
69	>= 1 Sr
70	>= 1 Y
71	>= 1 Zr
72	>= 1 Nb
73	>= 1 Mo
74	>= 1 Ru
75	>= 1 Rh
76	>= 1 Pd
77	>= 1 Ag
78	>= 1 Cd
79	>= 1 In
80	>= 1 Sn
81	>= 1 Sb
82	>= 1 Te
83	>= 1 Xe
84	>= 1 Cs
85	>= 1 Ba
86	>= 1 Lu
87	>= 1 Hf
88	>= 1 Ta
89	>= 1 W
90	>= 1 Re
91	>= 1 Os
92	>= 1 Ir
93	>= 1 Pt
94	>= 1 Au
95	>= 1 Hg
96	>= 1 Tl
97	>= 1 Pb
98	>= 1 Bi

PubChem Substructure Fingerprint Description (cont.)

Section 1: Hierarchic Element Counts (cont.)

Bit Position	Bit Substructure
99	>= 1 La
100	>= 1 Ce
101	>= 1 Pr
102	>= 1 Nd
103	>= 1 Pm
104	>= 1 Sm
105	>= 1 Eu
106	>= 1 Gd
107	>= 1 Tb
108	>= 1 Dy
109	>= 1 Ho
110	>= 1 Er
111	>= 1 Tm
112	>= 1 Yb
113	>= 1 Tc
114	>= 1 U

Section 2: Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set - These bits test for the presence or count of the described chemical ring system. An ESSSR ring is any ring which does not share three consecutive atoms with any other ring in the chemical structure. For example, naphthalene has three ESSSR rings (two phenyl fragments and the 10-membered envelope), while biphenyl will yield a count of only two ESSSR rings.

Bit Position	Bit Substructure
115	>= 1 any ring size 3
116	>= 1 saturated or aromatic carbon-only ring size 3
117	>= 1 saturated or aromatic nitrogen-containing ring size 3
118	>= 1 saturated or aromatic heteroatom-containing ring size 3
119	>= 1 unsaturated non-aromatic carbon-only ring size 3
120	>= 1 unsaturated non-aromatic nitrogen-containing ring size 3
121	>= 1 unsaturated non-aromatic heteroatom-containing ring size 3
122	>= 2 any ring size 3
123	>= 2 saturated or aromatic carbon-only ring size 3
124	>= 2 saturated or aromatic nitrogen-containing ring size 3
125	>= 2 saturated or aromatic heteroatom-containing ring size 3
126	>= 2 unsaturated non-aromatic carbon-only ring size 3
127	>= 2 unsaturated non-aromatic nitrogen-containing ring size 3
128	>= 2 unsaturated non-aromatic heteroatom-containing ring size 3
129	>= 1 any ring size 4
130	>= 1 saturated or aromatic carbon-only ring size 4
131	>= 1 saturated or aromatic nitrogen-containing ring size 4
132	>= 1 saturated or aromatic heteroatom-containing ring size 4
133	>= 1 unsaturated non-aromatic carbon-only ring size 4

PubChem Substructure Fingerprint Description (cont.)

Section 2: Rings in a canonic ESSR ring set (cont.)

Bit Position	Bit Substructure
134	>= 1 unsaturated non-aromatic nitrogen-containing ring size 4
135	>= 1 unsaturated non-aromatic heteroatom-containing ring size 4
136	>= 2 any ring size 4
137	>= 2 saturated or aromatic carbon-only ring size 4
138	>= 2 saturated or aromatic nitrogen-containing ring size 4
139	>= 2 saturated or aromatic heteroatom-containing ring size 4
140	>= 2 unsaturated non-aromatic carbon-only ring size 4
141	>= 2 unsaturated non-aromatic nitrogen-containing ring size 4
142	>= 2 unsaturated non-aromatic heteroatom-containing ring size 4
143	>= 1 any ring size 5
144	>= 1 saturated or aromatic carbon-only ring size 5
145	>= 1 saturated or aromatic nitrogen-containing ring size 5
146	>= 1 saturated or aromatic heteroatom-containing ring size 5
147	>= 1 unsaturated non-aromatic carbon-only ring size 5
148	>= 1 unsaturated non-aromatic nitrogen-containing ring size 5
149	>= 1 unsaturated non-aromatic heteroatom-containing ring size 5
150	>= 2 any ring size 5
151	>= 2 saturated or aromatic carbon-only ring size 5
152	>= 2 saturated or aromatic nitrogen-containing ring size 5
153	>= 2 saturated or aromatic heteroatom-containing ring size 5
154	>= 2 unsaturated non-aromatic carbon-only ring size 5
155	>= 2 unsaturated non-aromatic nitrogen-containing ring size 5
156	>= 2 unsaturated non-aromatic heteroatom-containing ring size 5
157	>= 3 any ring size 5
158	>= 3 saturated or aromatic carbon-only ring size 5
159	>= 3 saturated or aromatic nitrogen-containing ring size 5
160	>= 3 saturated or aromatic heteroatom-containing ring size 5
161	>= 3 unsaturated non-aromatic carbon-only ring size 5
162	>= 3 unsaturated non-aromatic nitrogen-containing ring size 5
163	>= 3 unsaturated non-aromatic heteroatom-containing ring size 5
164	>= 4 any ring size 5
165	>= 4 saturated or aromatic carbon-only ring size 5
166	>= 4 saturated or aromatic nitrogen-containing ring size 5
167	>= 4 saturated or aromatic heteroatom-containing ring size 5
168	>= 4 unsaturated non-aromatic carbon-only ring size 5
169	>= 4 unsaturated non-aromatic nitrogen-containing ring size 5
170	>= 4 unsaturated non-aromatic heteroatom-containing ring size 5
171	>= 5 any ring size 5
172	>= 5 saturated or aromatic carbon-only ring size 5
173	>= 5 saturated or aromatic nitrogen-containing ring size 5
174	>= 5 saturated or aromatic heteroatom-containing ring size 5
175	>= 5 unsaturated non-aromatic carbon-only ring size 5
176	>= 5 unsaturated non-aromatic nitrogen-containing ring size 5
177	>= 5 unsaturated non-aromatic heteroatom-containing ring size 5
178	>= 1 any ring size 6
179	>= 1 saturated or aromatic carbon-only ring size 6
180	>= 1 saturated or aromatic nitrogen-containing ring size 6
181	>= 1 saturated or aromatic heteroatom-containing ring size 6
182	>= 1 unsaturated non-aromatic carbon-only ring size 6

PubChem Substructure Fingerprint Description (cont.)

Section 2: Rings in a canonic ESSR ring set (cont.)

Bit Position	Bit Substructure
183	>= 1 unsaturated non-aromatic nitrogen-containing ring size 6
184	>= 1 unsaturated non-aromatic heteroatom-containing ring size 6
185	>= 2 any ring size 6
186	>= 2 saturated or aromatic carbon-only ring size 6
187	>= 2 saturated or aromatic nitrogen-containing ring size 6
188	>= 2 saturated or aromatic heteroatom-containing ring size 6
189	>= 2 unsaturated non-aromatic carbon-only ring size 6
190	>= 2 unsaturated non-aromatic nitrogen-containing ring size 6
191	>= 2 unsaturated non-aromatic heteroatom-containing ring size 6
192	>= 3 any ring size 6
193	>= 3 saturated or aromatic carbon-only ring size 6
194	>= 3 saturated or aromatic nitrogen-containing ring size 6
195	>= 3 saturated or aromatic heteroatom-containing ring size 6
196	>= 3 unsaturated non-aromatic carbon-only ring size 6
197	>= 3 unsaturated non-aromatic nitrogen-containing ring size 6
198	>= 3 unsaturated non-aromatic heteroatom-containing ring size 6
199	>= 4 any ring size 6
200	>= 4 saturated or aromatic carbon-only ring size 6
201	>= 4 saturated or aromatic nitrogen-containing ring size 6
202	>= 4 saturated or aromatic heteroatom-containing ring size 6
203	>= 4 unsaturated non-aromatic carbon-only ring size 6
204	>= 4 unsaturated non-aromatic nitrogen-containing ring size 6
205	>= 4 unsaturated non-aromatic heteroatom-containing ring size 6
206	>= 5 any ring size 6
207	>= 5 saturated or aromatic carbon-only ring size 6
208	>= 5 saturated or aromatic nitrogen-containing ring size 6
209	>= 5 saturated or aromatic heteroatom-containing ring size 6
210	>= 5 unsaturated non-aromatic carbon-only ring size 6
211	>= 5 unsaturated non-aromatic nitrogen-containing ring size 6
212	>= 5 unsaturated non-aromatic heteroatom-containing ring size 6
213	>= 1 any ring size 7
214	>= 1 saturated or aromatic carbon-only ring size 7
215	>= 1 saturated or aromatic nitrogen-containing ring size 7
216	>= 1 saturated or aromatic heteroatom-containing ring size 7
217	>= 1 unsaturated non-aromatic carbon-only ring size 7
218	>= 1 unsaturated non-aromatic nitrogen-containing ring size 7
219	>= 1 unsaturated non-aromatic heteroatom-containing ring size 7
220	>= 2 any ring size 7
221	>= 2 saturated or aromatic carbon-only ring size 7
222	>= 2 saturated or aromatic nitrogen-containing ring size 7
223	>= 2 saturated or aromatic heteroatom-containing ring size 7
224	>= 2 unsaturated non-aromatic carbon-only ring size 7
225	>= 2 unsaturated non-aromatic nitrogen-containing ring size 7
226	>= 2 unsaturated non-aromatic heteroatom-containing ring size 7
227	>= 1 any ring size 8
228	>= 1 saturated or aromatic carbon-only ring size 8
229	>= 1 saturated or aromatic nitrogen-containing ring size 8
230	>= 1 saturated or aromatic heteroatom-containing ring size 8
231	>= 1 unsaturated non-aromatic carbon-only ring size 8

PubChem Substructure Fingerprint Description (cont.)

Section 2: Rings in a canonic ESSR ring set (cont.)

Bit Position	Bit Substructure
232	>= 1 unsaturated non-aromatic nitrogen-containing ring size 8
233	>= 1 unsaturated non-aromatic heteroatom-containing ring size 8
234	>= 2 any ring size 8
235	>= 2 saturated or aromatic carbon-only ring size 8
236	>= 2 saturated or aromatic nitrogen-containing ring size 8
237	>= 2 saturated or aromatic heteroatom-containing ring size 8
238	>= 2 unsaturated non-aromatic carbon-only ring size 8
239	>= 2 unsaturated non-aromatic nitrogen-containing ring size 8
240	>= 2 unsaturated non-aromatic heteroatom-containing ring size 8
241	>= 1 any ring size 9
242	>= 1 saturated or aromatic carbon-only ring size 9
243	>= 1 saturated or aromatic nitrogen-containing ring size 9
244	>= 1 saturated or aromatic heteroatom-containing ring size 9
245	>= 1 unsaturated non-aromatic carbon-only ring size 9
246	>= 1 unsaturated non-aromatic nitrogen-containing ring size 9
247	>= 1 unsaturated non-aromatic heteroatom-containing ring size 9
248	>= 1 any ring size 10
249	>= 1 saturated or aromatic carbon-only ring size 10
250	>= 1 saturated or aromatic nitrogen-containing ring size 10
251	>= 1 saturated or aromatic heteroatom-containing ring size 10
252	>= 1 unsaturated non-aromatic carbon-only ring size 10
253	>= 1 unsaturated non-aromatic nitrogen-containing ring size 10
254	>= 1 unsaturated non-aromatic heteroatom-containing ring size 10
255	>= 1 aromatic ring
256	>= 1 hetero-aromatic ring
257	>= 2 aromatic rings
258	>= 2 hetero-aromatic rings
259	>= 3 aromatic rings
260	>= 3 hetero-aromatic rings
261	>= 4 aromatic rings
262	>= 4 hetero-aromatic rings

Section 3: Simple atom pairs – These bits test for the presence of patterns of bonded atom pairs, regardless of bond order or count.

Bit Position	Bit Substructure
263	Li-H
264	Li-Li
265	Li-B
266	Li-C
267	Li-O
268	Li-F
269	Li-P
270	Li-S

PubChem Substructure Fingerprint Description (cont.)

Section 3: Simple atom pairs (cont.)

Bit Position	Bit Substructure
271	Li-Cl
272	B-H
273	B-B
274	B-C
275	B-N
276	B-O
277	B-F
278	B-Si
279	B-P
280	B-S
281	B-Cl
282	B-Br
283	C-H
284	C-C
285	C-N
286	C-O
287	C-F
288	C-Na
289	C-Mg
290	C-Al
291	C-Si
292	C-P
293	C-S
294	C-Cl
295	C-As
296	C-Se
297	C-Br
298	C-I
299	N-H
300	N-N
301	N-O
302	N-F
303	N-Si
304	N-P
305	N-S
306	N-Cl
307	N-Br
308	O-H
309	O-O
310	O-Mg
311	O-Na
312	O-Al
313	O-Si
314	O-P
315	O-K
316	F-P
317	F-S
318	Al-H
319	Al-Cl
320	Si-H

PubChem Substructure Fingerprint Description (cont.)

Section 3: Simple atom pairs (cont.)

Bit Position	Bit Substructure
321	Si-Si
322	Si-Cl
323	P-H
324	P-P
325	As-H
326	As-As

Section 4: Simple atom nearest neighbors – These bits test for the presence of atom nearest neighbor patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.

Bit Position	Bit Substructure
327	C(~Br)(~C)
328	C(~Br)(~C)(~C)
329	C(~Br)(~H)
330	C(~Br)(:C)
331	C(~Br)(:N)
332	C(~C)(~C)
333	C(~C)(~C)(~C)
334	C(~C)(~C)(~C)(~C)
335	C(~C)(~C)(~C)(~H)
336	C(~C)(~C)(~C)(~N)
337	C(~C)(~C)(~C)(~O)
338	C(~C)(~C)(~H)(~N)
339	C(~C)(~C)(~H)(~O)
340	C(~C)(~C)(~N)
341	C(~C)(~C)(~O)
342	C(~C)(~Cl)
343	C(~C)(~Cl)(~H)
344	C(~C)(~H)
345	C(~C)(~H)(~N)
346	C(~C)(~H)(~O)
347	C(~C)(~H)(~O)(~O)
348	C(~C)(~H)(~P)
349	C(~C)(~H)(~S)
350	C(~C)(~I)
351	C(~C)(~N)
352	C(~C)(~O)
353	C(~C)(~S)
354	C(~C)(~Si)
355	C(~C)(:C)
356	C(~C)(:C)(:C)
357	C(~C)(:C)(:N)
358	C(~C)(:N)
359	C(~C)(:N)(:N)
360	C(~Cl)(~Cl)

PubChem Substructure Fingerprint Description (cont.)

Section 4: Simple atom nearest neighbors (cont.)

Bit Position	Bit Substructure
361	C(~Cl)(~H)
362	C(~Cl)(:C)
363	C(~F)(~F)
364	C(~F)(:C)
365	C(~H)(~N)
366	C(~H)(~O)
367	C(~H)(~O)(~O)
368	C(~H)(~S)
369	C(~H)(~Si)
370	C(~H)(:C)
371	C(~H)(:C)(:C)
372	C(~H)(:C)(:N)
373	C(~H)(:N)
374	C(~H)(~H)(~H)
375	C(~N)(~N)
376	C(~N)(:C)
377	C(~N)(:C)(:C)
378	C(~N)(:C)(:N)
379	C(~N)(:N)
380	C(~O)(~O)
381	C(~O)(:C)
382	C(~O)(:C)(:C)
383	C(~S)(:C)
384	C(:C)(:C)
385	C(:C)(:C)(:C)
386	C(:C)(:C)(:N)
387	C(:C)(:N)
388	C(:C)(:N)(:N)
389	C(:N)(:N)
390	N(~C)(~C)
391	N(~C)(~C)(~C)
392	N(~C)(~C)(~H)
393	N(~C)(~H)
394	N(~C)(~H)(~N)
395	N(~C)(~O)
396	N(~C)(:C)
397	N(~C)(:C)(:C)
398	N(~H)(~N)
399	N(~H)(:C)
400	N(~H)(:C)(:C)
401	N(~O)(~O)
402	N(~O)(:O)
403	N(:C)(:C)
404	N(:C)(:C)(:C)
405	O(~C)(~C)
406	O(~C)(~H)
407	O(~C)(~P)
408	O(~H)(~S)
409	O(:C)(:C)
410	P(~C)(~C)

PubChem Substructure Fingerprint Description (cont.)

Section 4: Simple atom nearest neighbors (cont.)

Bit Position	Bit Substructure
411	P(~O)(~O)
412	S(~C)(~C)
413	S(~C)(~H)
414	S(~C)(~O)
415	Si(~C)(~C)

Section 5: Detailed atom neighborhoods – These bits test for the presence of detailed atom neighborhood patterns, regardless of count, but where bond orders are specific, bond aromaticity matches both single and double bonds, and where "-", "=", and "#" matches a single bond, double bond, and triple bond order, respectively.

Bit Position	Bit Substructure
416	C=C
417	C#C
418	C=N
419	C#N
420	C=O
421	C=S
422	N=N
423	N=O
424	N=P
425	P=O
426	P=P
427	C(#C)(-C)
428	C(#C)(-H)
429	C(#N)(-C)
430	C(-C)(-C)(=C)
431	C(-C)(-C)(=N)
432	C(-C)(-C)(=O)
433	C(-C)(-Cl)(=O)
434	C(-C)(-H)(=C)
435	C(-C)(-H)(=N)
436	C(-C)(-H)(=O)
437	C(-C)(-N)(=C)
438	C(-C)(-N)(=N)
439	C(-C)(-N)(=O)
440	C(-C)(-O)(=O)
441	C(-C)(=C)
442	C(-C)(=N)
443	C(-C)(=O)
444	C(-Cl)(=O)
445	C(-H)(-N)(=C)
446	C(-H)(=C)
447	C(-H)(=N)
448	C(-H)(=O)

PubChem Substructure Fingerprint Description (cont.)

Section 5: Detailed atom neighborhoods (cont.)

Bit Position	Bit Substructure
449	C(-N)(=C)
450	C(-N)(=N)
451	C(-N)(=O)
452	C(-O)(=O)
453	N(-C)(=C)
454	N(-C)(=O)
455	N(-O)(=O)
456	P(-O)(=O)
457	S(-C)(=O)
458	S(-O)(=O)
459	S(=O)(=O)

Section 6: Simple SMARTS patterns – These bits test for the presence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds.

Bit Position	Bit Substructure
460	C-C-C#C
461	O-C-C=N
462	O-C-C=O
463	N:C-S-[#1]
464	N-C-C=C
465	O=S-C-C
466	N#C-C=C
467	C=N-N-C
468	O=S-C-N
469	S-S-C:C
470	C:C-C=C
471	S:C:C:C
472	C:N:C-C
473	S-C:N:C
474	S:C:C:N
475	S-C=N-C
476	C-O-C=C
477	N-N-C:C
478	S-C=N-[#1]
479	S-C-S-C
480	C:S:C-C
481	O-S-C:C
482	C:N-C:C
483	N-S-C:C
484	N-C:N:C
485	N:C:C:N
486	N-C:N:N
487	N-C=N-C
488	N-C=N-[#1]

PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
489	N-C-S-C
490	C-C-C=C
491	C-N:C-[#1]
492	N-C:O:C
493	O=C-C:C
494	O=C-C:N
495	C-N-C:C
496	N:N-C-[#1]
497	O-C:C:N
498	O-C=C-C
499	N-C:C:N
500	C-S-C:C
501	Cl-C:C-C
502	N-C=C-[#1]
503	Cl-C:C-[#1]
504	N:C:N-C
505	Cl-C:C-O
506	C-C:N:C
507	C-C-S-C
508	S=C-N-C
509	Br-C:C-C
510	[#1]-N-N-[#1]
511	S=C-N-[#1]
512	C-[As]-O-[#1]
513	S:C:C-[#1]
514	O-N-C-C
515	N-N-C-C
516	[#1]-C=C-[#1]
517	N-N-C-N
518	O=C-N-N
519	N=C-N-C
520	C=C-C:C
521	C:N-C-[#1]
522	C-N-N-[#1]
523	N:C:C-C
524	C-C=C-C
525	[As]-C:C-[#1]
526	Cl-C:C-Cl
527	C:C:N-[#1]
528	[#1]-N-C-[#1]
529	Cl-C-C-Cl
530	N:C-C:C
531	S-C:C-C
532	S-C:C-[#1]
533	S-C:C-N
534	S-C:C-O
535	O=C-C-C
536	O=C-C-N
537	O=C-C-O
538	N=C-C-C

PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
539	N=C-C-[#1]
540	C-N-C-[#1]
541	O-C:C-C
542	O-C:C-[#1]
543	O-C:C-N
544	O-C:C-O
545	N-C:C-C
546	N-C:C-[#1]
547	N-C:C-N
548	O-C-C:C
549	N-C-C:C
550	Cl-C-C-C
551	Cl-C-C-O
552	C:C-C:C
553	O=C-C=C
554	Br-C-C-C
555	N=C-C=C
556	C=C-C-C
557	N:C-O-[#1]
558	O=N-C:C
559	O-C-N-[#1]
560	N-C-N-C
561	Cl-C-C=O
562	Br-C-C=O
563	O-C-O-C
564	C=C-C=C
565	C:C-O-C
566	O-C-C-N
567	O-C-C-O
568	N#C-C-C
569	N-C-C-N
570	C:C-C-C
571	[#1]-C-O-[#1]
572	N:C:N:C
573	O-C-C=C
574	O-C-C:C-C
575	O-C-C:C-O
576	N=C-C:C-[#1]
577	C:C-N-C:C
578	C-C:C-C:C
579	O=C-C-C-C
580	O=C-C-C-N
581	O=C-C-C-O
582	C-C-C-C-C
583	Cl-C:C-O-C
584	C:C-C=C-C
585	C-C:C-N-C
586	C-S-C-C-C
587	N-C:C-O-[#1]
588	O=C-C-C=O

PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
589	C-C:C-O-C
590	C-C:C-O-[#1]
591	Cl-C-C-C-C
592	N-C-C-C-C
593	N-C-C-C-N
594	C-O-C-C=C
595	C:C-C-C-C
596	N=C-N-C-C
597	O=C-C-C:C
598	Cl-C:C:C-C
599	[#1]-C-C=C-[#1]
600	N-C:C:C-C
601	N-C:C:C-N
602	O=C-C-N-C
603	C-C:C:C-C
604	C-O-C-C:C
605	O=C-C-O-C
606	O-C:C-C-C
607	N-C-C-C:C
608	C-C-C-C:C
609	Cl-C-C-N-C
610	C-O-C-O-C
611	N-C-C-N-C
612	N-C-O-C-C
613	C-N-C-C-C
614	C-C-O-C-C
615	N-C-C-O-C
616	C:C:N:N:C
617	C-C-C-O-[#1]
618	C:C-C-C:C
619	O-C-C=C-C
620	C:C-O-C-C
621	N-C:C:C:N
622	O=C-O-C:C
623	O=C-C:C-C
624	O=C-C:C-N
625	O=C-C:C-O
626	C-O-C:C-C
627	O=[As]-C:C:C
628	C-N-C-C:C
629	S-C:C:C-N
630	O-C:C-O-C
631	O-C:C-O-[#1]
632	C-C-O-C:C
633	N-C-C:C-C
634	C-C-C:C-C
635	N-N-C-N-[#1]
636	C-N-C-N-C
637	O-C-C-C-C
638	O-C-C-C-N

PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
639	O-C-C-C-O
640	C=C-C-C-C
641	O-C-C-C=C
642	O-C-C-C=O
643	[#1]-C-C-N-[#1]
644	C-C=N-N-C
645	O=C-N-C-C
646	O=C-N-C-[#1]
647	O=C-N-C-N
648	O=N-C:C-N
649	O=N-C:C-O
650	O=C-N-C=O
651	O-C:C:C-C
652	O-C:C:C-N
653	O-C:C:C-O
654	N-C-N-C-C
655	O-C-C-C:C
656	C-C-N-C-C
657	C-N-C:C-C
658	C-C-S-C-C
659	O-C-C-N-C
660	C-C=C-C-C
661	O-C-O-C-C
662	O-C-C-O-C
663	O-C-C-O-[#1]
664	C-C=C-C=C
665	N-C:C-C-C
666	C=C-C-O-C
667	C=C-C-O-[#1]
668	C-C:C-C-C
669	Cl-C:C-C=O
670	Br-C:C:C-C
671	O=C-C=C-C
672	O=C-C=C-[#1]
673	O=C-C=C-N
674	N-C-N-C:C
675	Br-C-C-C:C
676	N#C-C-C-C
677	C-C=C-C:C
678	C-C-C=C-C
679	C-C-C-C-C-C
680	O-C-C-C-C-C
681	O-C-C-C-C-O
682	O-C-C-C-C-N
683	N-C-C-C-C-C
684	O=C-C-C-C-C
685	O=C-C-C-C-N
686	O=C-C-C-C-O
687	O=C-C-C-C=O
688	C-C-C-C-C-C-C

PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
689	O-C-C-C-C-C-C
690	O-C-C-C-C-C-O
691	O-C-C-C-C-C-N
692	O=C-C-C-C-C-C
693	O=C-C-C-C-C-O
694	O=C-C-C-C-C=O
695	O=C-C-C-C-C-N
696	C-C-C-C-C-C-C-C
697	C-C-C-C-C-C(C)-C
698	O-C-C-C-C-C-C-C
699	O-C-C-C-C-C(C)-C
700	O-C-C-C-C-C-O-C
701	O-C-C-C-C-C(O)-C
702	O-C-C-C-C-C-N-C
703	O-C-C-C-C-C(N)-C
704	O=C-C-C-C-C-C-C
705	O=C-C-C-C-C(O)-C
706	O=C-C-C-C-C(=O)-C
707	O=C-C-C-C-C(N)-C
708	C-C(C)-C-C
709	C-C(C)-C-C-C
710	C-C-C(C)-C-C
711	C-C(C)(C)-C-C
712	C-C(C)-C(C)-C

Section 7: Complex SMARTS patterns – These bits test for the presence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific.

Bit Position	Bit Substructure
713	Cc1ccc(C)cc1
714	Cc1ccc(O)cc1
715	Cc1ccc(S)cc1
716	Cc1ccc(N)cc1
717	Cc1ccc(Cl)cc1
718	Cc1ccc(Br)cc1
719	Oc1ccc(O)cc1
720	Oc1ccc(S)cc1
721	Oc1ccc(N)cc1
722	Oc1ccc(Cl)cc1
723	Oc1ccc(Br)cc1
724	Sc1ccc(S)cc1
725	Sc1ccc(N)cc1
726	Sc1ccc(Cl)cc1
727	Sc1ccc(Br)cc1
728	Nc1ccc(N)cc1
729	Nc1ccc(Cl)cc1
730	Nc1ccc(Br)cc1

PubChem Substructure Fingerprint Description (cont.)

Section 7: Complex SMARTS patterns (cont.)

Bit Position	Bit Substructure
731	Clc1ccc(Cl)cc1
732	Clc1ccc(Br)cc1
733	Brc1ccc(Br)cc1
734	Cc1cc(C)ccc1
735	Cc1cc(O)ccc1
736	Cc1cc(S)ccc1
737	Cc1cc(N)ccc1
738	Cc1cc(Cl)ccc1
739	Cc1cc(Br)ccc1
740	Oc1cc(O)ccc1
741	Oc1cc(S)ccc1
742	Oc1cc(N)ccc1
743	Oc1cc(Cl)ccc1
744	Oc1cc(Br)ccc1
745	Sc1cc(S)ccc1
746	Sc1cc(N)ccc1
747	Sc1cc(Cl)ccc1
748	Sc1cc(Br)ccc1
749	Nc1cc(N)ccc1
750	Nc1cc(Cl)ccc1
751	Nc1cc(Br)ccc1
752	Clc1cc(Cl)ccc1
753	Clc1cc(Br)ccc1
754	Brc1cc(Br)ccc1
755	Cc1c(C)cccc1
756	Cc1c(O)cccc1
757	Cc1c(S)cccc1
758	Cc1c(N)cccc1
759	Cc1c(Cl)cccc1
760	Cc1c(Br)cccc1
761	Oc1c(O)cccc1
762	Oc1c(S)cccc1
763	Oc1c(N)cccc1
764	Oc1c(Cl)cccc1
765	Oc1c(Br)cccc1
766	Sc1c(S)cccc1
767	Sc1c(N)cccc1
768	Sc1c(Cl)cccc1
769	Sc1c(Br)cccc1
770	Nc1c(N)cccc1
771	Nc1c(Cl)cccc1
772	Nc1c(Br)cccc1
773	Clc1c(Cl)cccc1
774	Clc1c(Br)cccc1
775	Brc1c(Br)cccc1
776	CC1CCC(C)CC1
777	CC1CCC(O)CC1
778	CC1CCC(S)CC1
779	CC1CCC(N)CC1
780	CC1CCC(Cl)CC1

PubChem Substructure Fingerprint Description (cont.)

Section 7: Complex SMARTS patterns (cont.)

Bit Position	Bit Substructure
781	CC1CCC(Br)CC1
782	OC1CCC(O)CC1
783	OC1CCC(S)CC1
784	OC1CCC(N)CC1
785	OC1CCC(Cl)CC1
786	OC1CCC(Br)CC1
787	SC1CCC(S)CC1
788	SC1CCC(N)CC1
789	SC1CCC(Cl)CC1
790	SC1CCC(Br)CC1
791	NC1CCC(N)CC1
792	NC1CCC(Cl)CC1
793	NC1CCC(Br)CC1
794	ClC1CCC(Cl)CC1
795	ClC1CCC(Br)CC1
796	BrC1CCC(Br)CC1
797	CC1CC(C)CCC1
798	CC1CC(O)CCC1
799	CC1CC(S)CCC1
800	CC1CC(N)CCC1
801	CC1CC(Cl)CCC1
802	CC1CC(Br)CCC1
803	OC1CC(O)CCC1
804	OC1CC(S)CCC1
805	OC1CC(N)CCC1
806	OC1CC(Cl)CCC1
807	OC1CC(Br)CCC1
808	SC1CC(S)CCC1
809	SC1CC(N)CCC1
810	SC1CC(Cl)CCC1
811	SC1CC(Br)CCC1
812	NC1CC(N)CCC1
813	NC1CC(Cl)CCC1
814	NC1CC(Br)CCC1
815	ClC1CC(Cl)CCC1
816	ClC1CC(Br)CCC1
817	BrC1CC(Br)CCC1
818	CC1C(C)CCCC1
819	CC1C(O)CCCC1
820	CC1C(S)CCCC1
821	CC1C(N)CCCC1
822	CC1C(Cl)CCCC1
823	CC1C(Br)CCCC1
824	OC1C(O)CCCC1
825	OC1C(S)CCCC1
826	OC1C(N)CCCC1
827	OC1C(Cl)CCCC1
828	OC1C(Br)CCCC1
829	SC1C(S)CCCC1
830	SC1C(N)CCCC1

PubChem Substructure Fingerprint Description (cont.)

Section 7: Complex SMARTS patterns (cont.)

Bit Position	Bit Substructure
831	SC1C(Cl)CCCC1
832	SC1C(Br)CCCC1
833	NC1C(N)CCCC1
834	NC1C(Cl)CCCC1
835	NC1C(Br)CCCC1
836	ClC1C(Cl)CCCC1
837	ClC1C(Br)CCCC1
838	BrC1C(Br)CCCC1
839	CC1CC(C)CC1
840	CC1CC(O)CC1
841	CC1CC(S)CC1
842	CC1CC(N)CC1
843	CC1CC(Cl)CC1
844	CC1CC(Br)CC1
845	OC1CC(O)CC1
846	OC1CC(S)CC1
847	OC1CC(N)CC1
848	OC1CC(Cl)CC1
849	OC1CC(Br)CC1
850	SC1CC(S)CC1
851	SC1CC(N)CC1
852	SC1CC(Cl)CC1
853	SC1CC(Br)CC1
854	NC1CC(N)CC1
855	NC1CC(Cl)CC1
856	NC1CC(Br)CC1
857	ClC1CC(Cl)CC1
858	ClC1CC(Br)CC1
859	BrC1CC(Br)CC1
860	CC1C(C)CCC1
861	CC1C(O)CCC1
862	CC1C(S)CCC1
863	CC1C(N)CCC1
864	CC1C(Cl)CCC1
865	CC1C(Br)CCC1
866	OC1C(O)CCC1
867	OC1C(S)CCC1
868	OC1C(N)CCC1
869	OC1C(Cl)CCC1
870	OC1C(Br)CCC1
871	SC1C(S)CCC1
872	SC1C(N)CCC1
873	SC1C(Cl)CCC1
874	SC1C(Br)CCC1
875	NC1C(N)CCC1
876	NC1C(Cl)CC1
877	NC1C(Br)CCC1
878	ClC1C(Cl)CCC1
879	ClC1C(Br)CCC1
880	BrC1C(Br)CCC1

Decoding PubChem Fingerprints

PubChem fingerprints are currently 881 bits in length. Binary data is stored in one byte increments. The fingerprint is, therefore, 111 bytes in length (888 bits), which includes padding of seven bits at the end to complete the last byte. A four-byte prefix, containing the bit length of the fingerprint (881 bits), increases the stored PubChem fingerprint size to 115 bytes (920 bits).

When PubChem fingerprints are encoded in base64 format, the base64-encoded fingerprints are 156 bytes in length. The last two bytes are padding so that the base64 length is divisible by four (156 bytes – 2 bytes = 154 bytes). Each base64 byte encodes six binary bits (154 bytes * 6 bits/byte = 924 bits). The last four bits are padding to complete the last base64 byte (924 bits – 4 bits = 920 bits). The resulting 920 binary bits (115 bytes) are described in the previous paragraph.

Document Version History

V1.3 – 2009May01 – Updated introduction to describe how to identify the PubChem Substructure Fingerprint property in a PubChem Compound record.
V1.2 – 2007Aug30 – Added section on decoding PubChem fingerprints.
V1.1 – 2007Aug06 – Corrected and expanded documentation of bits with SMARTS patterns used.
V1.0 – 2005Dec02 - Initial release.