Analysis of Structure-Activity Relationships of Food-Derived DPP IV-Inhibitory Di- and Tripeptides Using Interpretable Descriptors

Monika Hrynkiewicz[1,†], Anna Iwaniak[1,*,†], Piotr Minkiewicz[1], Małgorzata Darewicz[1], Wojciech Płonka[2]

SUPPLEMENT

SCIGRESS-extracted report for QSAR model of dipeptide DPP IV inhibitors

Part 1: Summary

a. Property predicted

The property used to develop the QSAR is named *log10ic50 Extract from SDF (from A)* and its values were obtained from a training set of 46 chemical samples.

b. Best QSAR equation

Using the *Complete Topological QSAR* option, the following regression equation, which is the best of 336 descriptors selected with Enhanced Replacement Method, gave the highest correlation coefficient (r^2=0.7824).

> *log10ic50* = 1794.9070**hydrogen donor partial surface area/total accessible surface area/MW* + 364.4572**guanidine count/MW* - 0.0843**Nitrogen count*^{2} - 0.0635**methyl count*^{2} + 0.3563**ring count 5 member*^{2} + 2.2731

c. Quality of the best QSAR equation

The cross-validated correlation coefficient (cvr^2= 0.7373) suggests that the stability of the equation on addition of similar training data is likely to be reasonable as it is above 0.70. A more detailed analysis is provided in Part 2.

The average error for the training set is 0.2546 and the standard deviation is 0.3431.

The F-ratio is 28.7726. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05 (1 in 20), there is at least one significant descriptor in the model, i.e. this is a valid and stable equation. A probability above 0.05 indicates that the equation might be a chance correlation and not stable.

Based on the partial-F value of each descriptor, there is a greater than 99% probability that all descriptors are significant.

The training data and QSAR predictions were checked for the following and warnings were noted and are discussed in more detail in Part 2:

1. There are enough observed data values per descriptor.
2. The data is distributed evenly enough.
3. The training set had these notes:

   EK:
   Only sample in set with a carboxylate count.
   MP:
   Nonbonded atoms C13 and O32 are too close.
   SP:
   Nonbonded atoms C1 and O26 are too close.
   WC:
   Only sample in set with a thiol count.
   YP:
   Nonbonded atoms H9 and H35 are too close.

4. No outliers or problems with the predicted values were found.

An independent set of chemical samples should be used to test this equation.

d. Applicable prediction range and chemical space

This equation should be used to estimate only data values that fall within the training range from 1.577 to 3.986. Predictions that fall outside this range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that have properties in these ranges:

1. *hydrogen donor partial surface area/total accessible surface area/MW* from 0.000 to 0.001.
2. *guanidine count/MW* from 0.000 to 0.004.
3. *Nitrogen count^2* from 4.000 to 36.000.
4. *methyl count^2* from 0.000 to 16.000.
5. *ring count 5 member^2* from 0.000 to 4.000.

Predictions for chemical samples with properties that fall outside the training set property range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that are chemically similar to the training set or share a common mode of action. The 46 samples in the training set included the following elements (min:max:# samples): H(12:25:46) C(6:22:46) N(2:6:46) O(3:5:46) S(0:2:6). The 46 samples in the training set included the following functional groups: *thiol*, *ring*, *ring 5 member*, *sec-amine*, *ring aromatic*, *ring 6 member*, *carboxylate*, *H-bond acceptor*, *hydroxyl*, *guanidine*, *amide*, *ring aromatic 6*, *phenol*, *H-bond donor*, *methyl*, *sulfide*, *ring aromatic 5*, *methylene*, *amine*, *tertiary-amine*.

e. Mechanistic interpretation

The descriptors and their relative importance are listed below:

Descriptor	Relative importance
hydrogen donor partial surface area/total accessible surface area/MW	0.7559
guanidine count/MW	0.5086
Nitrogen count^2	-1.0000
methyl count^2	-0.3414
ring count 5 member^2	0.6070

Part 2: Detailed Analysis

Data distribution

There are 9.200 data values per descriptor in the QSAR model.

The QSAR equation was derived using a training set of 46 chemical samples with a three-sigma range for log10ic50 Extract from SDF (from A) from 0.594 to 4.96. The average was 2.777 and the standard deviation was 0.728 with a minimum data value of 1.577 and a maximum of 3.986. The data skewness measure is -0.023. The data skewness is between -2.0 and 2.0 which indicates that the data is not skewed. Partitioning the data into equal thirds from lowest to highest data values gives three bins with these counts: 15:14:17.

Chemical samples

Chemical samples in the training set had molecular weights from 160.171 to 390.435 and these elements and counts:

Element	Lowest	Highest
Hydrogen	12	25
Carbon	6	22
Nitrogen	2	6
Oxygen	3	5
Sulfur	0	2

The chemical samples had these charges (and counts): +0(36) +1(10).

The chemical samples had these groups and counts:

Group	Lowest	Highest
thiol	0	1
ring	0	4
ring 5 member	0	2
sec-amine	0	2
ring aromatic	0	4
ring 6 member	0	2
carboxylate	1	2

H-bond acceptor	4	7
hydroxyl	0	1
guanidine	0	1
amide	0	1
ring aromatic 6	0	2
phenol	0	1
H-bond donor	2	10
methyl	0	4
sulfide	0	2
ring aromatic 5	0	2
methylene	0	7
amine	0	2
tertiary-amine	0	1

Chemical samples were used as-is without preconditioning the geometry.

The training set samples had these notes:

EK:
Only sample in set with a carboxylate count.
MP:
Nonbonded atoms C13 and O32 are too close.
SP:
Nonbonded atoms C1 and O26 are too close.
WC:
Only sample in set with a thiol count.
YP:
Nonbonded atoms H9 and H35 are too close.

Analysis of QSAR equation

The following equation predicts log10ic50 Extract from SDF (from A):

> *log10ic50* = 1794.9070**hydrogen donor partial surface area/total accessible surface area/MW* + 364.4572**guanidine count/MW* - 0.0843**Nitrogen count*^{2} - 0.0635**methyl count*^{2} + 0.3563**ring count 5 member*^{2} + 2.2731

The average error for the training set is 0.2546. The standard deviation of the error is 0.3431.

r^2 is 0.7824; the degrees-of-freedom adjusted r^2 is 0.7553; and the leave-one-out cross-validated r^2 is 0.7373. The standard deviation in the error predicted by leave-one-out cross-validation is 0.3770. The F-ratio is 28.7726. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05, there is at least one significant descriptor in the model.

Use the normalized coefficients in the following analysis section to interpret the relative importance of each descriptor.

78.2% of the variability in log10ic50 Extract from SDF (from A) is explained by this equation.

The relative weight of each normalized contribution is:

Descriptor	Coefficient	Normalized coefficient	Descriptor standard deviation	Partial-F	Probability of greater F-ratio
hydrogen donor partial surface area/total accessible surface area/MW	1794.9070	0.7559	0.0003	51.3355	0.0000
guanidine count/MW	364.4572	0.5086	0.0008	14.5636	0.0005
Nitrogen count^2	-0.0843	-1.0000	7.1918	53.3368	0.0000
methyl count^2	-0.0635	-0.3414	3.2607	9.8481	0.0032
ring count 5 member^2	0.3563	0.6070	1.0328	28.3351	0.0000
Constant	2.2731				

The equation is the best of 336 descriptors selected with Enhanced Replacement Method. The correlation between any pair of descriptors that appear in the equation is less than 0.9500

A plot of predicted values against the original data may be found in the 3. Best QSAR plot. Other equations with r^2 from 0.7824 to 0.7824 may be found in the 3. Other QSARs table.

This equation should be used only to predict values of similar chemicals that have molecular weights, elements, groups and charges within the lowest to highest ranges of the training set.

SCIGRESS-extracted report of the list of 336 descriptors for QSAR modeling of dipeptide DPP IV inhibitors

1. 1.0/Carbon count inverse (from A)
2. 1.0/Csp^2 bonded to 1 C inverse (from A)
3. 1.0/Csp^3 bonded to 2 C inverse (from A)
4. 1.0/H-bond acceptor count inverse (from A)
5. 1.0/H-bond donor count inverse (from A)
6. 1.0/Hydrogen count inverse (from A)
7. 1.0/Nitrogen count inverse (from A)
8. 1.0/Oxygen count inverse (from A)
9. 1.0/all bond count inverse (from A)
10. 1.0/all count inverse (from A)
11. 1.0/box area inverse (from A)
12. 1.0/box cross section inverse (from A)
13. 1.0/box volume inverse (from A)
14. 1.0/depth inverse (from A)
15. 1.0/double bond count inverse (from A)
16. 1.0/hydrogen donor partial surface area inverse (from A)
17. 1.0/hydrogen donor partial surface area/total accessible surface area inverse (from A)
18. 1.0/hydrophobic dipole inverse (from A)
19. 1.0/hydrophobicity weighted area inverse (from A)
20. 1.0/hydrophobicity weighted negative area inverse (from A)

21. 1.0/hydrophobicity weighted negative area/total accessible surface area inverse (from A)
22. 1.0/hydrophobicity weighted positive area inverse (from A)
23. 1.0/hydrophobicity weighted positive area/total accessible surface area inverse (from A)
24. 1.0/length inverse (from A)
25. 1.0/length/width inverse (from A)
26. 1.0/log P inverse (from A)
27. 1.0/log P/box area inverse (from A)
28. 1.0/log P/box cross section inverse (from A)
29. 1.0/log P/box volume inverse (from A)
30. 1.0/log P/width/depth inverse (from A)
31. 1.0/molecular weight inverse (from A)
32. 1.0/nonpolar area inverse (from A)
33. 1.0/rotatable bond count inverse (from A)
34. 1.0/rotatable bond count nonterminal inverse (from A)
35. 1.0/single bond count inverse (from A)
36. 1.0/total accessible surface area inverse (from A)
37. 1.0/total positive charges inverse (from A)
38. 1.0/width inverse (from A)
39. 1.0/width/depth inverse (from A)
40. Carbon count carbon (from A)
41. Carbon count/MW DivideByMW (from A)
42. Carbon count^2 square (from A)
43. Csp^2 bonded to 1 C standard procedure (from A)
44. Csp^2 bonded to 1 C/MW DivideByMW (from A)
45. Csp^2 bonded to 1 C^2 square (from A)
46. Csp^2 bonded to 2 C standard procedure (from A)
47. Csp^2 bonded to 2 C/MW DivideByMW (from A)
48. Csp^2 bonded to 2 C^2 square (from A)
49. Csp^2 bonded to 3 C standard procedure (from A)
50. Csp^2 bonded to 3 C/MW DivideByMW (from A)
51. Csp^2 bonded to 3 C^2 square (from A)
52. Csp^3 bonded to 1 C standard procedure (from A)
53. Csp^3 bonded to 1 C/MW DivideByMW (from A)
54. Csp^3 bonded to 1 C^2 square (from A)
55. Csp^3 bonded to 2 C standard procedure (from A)
56. Csp^3 bonded to 2 C/MW DivideByMW (from A)
57. Csp^3 bonded to 2 C^2 square (from A)
58. Csp^3 bonded to 3 C standard procedure (from A)
59. Csp^3 bonded to 3 C/MW DivideByMW (from A)
60. Csp^3 bonded to 3 C^2 square (from A)
61. H-bond acceptor count Lipinski's count (from A)
62. H-bond acceptor count/MW DivideByMW (from A)
63. H-bond acceptor count^2 square (from A)
64. H-bond donor count Lipinski's count (from A)
65. H-bond donor count/MW DivideByMW (from A)
66. H-bond donor count^2 square (from A)
67. Hydrogen count hydrogen (from A)
68. Hydrogen count/MW DivideByMW (from A)

69. Hydrogen count^2 square (from A)
70. Nitrogen count nitrogen (from A)
71. Nitrogen count/MW DivideByMW (from A)
72. Nitrogen count^2 square (from A)
73. Oxygen count oxygen (from A)
74. Oxygen count/MW DivideByMW (from A)
75. Oxygen count^2 square (from A)
76. Sulfur count sulfur (from A)
77. Sulfur count/MW DivideByMW (from A)
78. Sulfur count^2 square (from A)
79. all bond count all bonds (from A)
80. all bond count/MW DivideByMW (from A)
81. all bond count^2 square (from A)
82. all count all atoms (from A)
83. all count/MW DivideByMW (from A)
84. all count^2 square (from A)
85. amide count amide (from A)
86. amide count/MW DivideByMW (from A)
87. amine count amine (from A)
88. amine count/MW DivideByMW (from A)
89. amine count^2 square (from A)
90. box area extract from sample [angstrom^2] (from A)
91. box area/MW DivideByMW (from A)
92. box area^2 square (from A)
93. box cross section extract from sample (from A)
94. box cross section/MW DivideByMW (from A)
95. box cross section^2 square (from A)
96. box volume extract from sample [angstrom^3] (from A)
97. box volume/MW DivideByMW (from A)
98. box volume^2 square (from A)
99. carboxylate count/MW DivideByMW (from A)
100. charge extract from sample [charge_au] (from A)
101. charge/MW DivideByMW (from A)
102. depth extract from sample (from A)
103. depth/MW DivideByMW (from A)
104. depth^2 square (from A)
105. double bond count double bonds (from A)
106. double bond count/MW DivideByMW (from A)
107. double bond count^2 square (from A)
108. guanidine count guanidine (from A)
109. guanidine count/MW DivideByMW (from A)
110. hydrogen donor partial surface area standard procedure (from A)
111. hydrogen donor partial surface area/MW DivideByMW (from A)
112. hydrogen donor partial surface area/total accessible surface area divide by total accessible surface area (from A)
113. hydrogen donor partial surface area/total accessible surface area/MW DivideByMW (from A)
114. hydrogen donor partial surface area/total accessible surface area^2 square (from A)
115. hydrogen donor partial surface area^2 square (from A)

116. hydrophobic dipole standard procedure (from A)
117. hydrophobic dipole/MW DivideByMW (from A)
118. hydrophobic dipole^2 square (from A)
119. hydrophobicity weighted area standard procedure (from A)
120. hydrophobicity weighted area/MW DivideByMW (from A)
121. hydrophobicity weighted area^2 square (from A)
122. hydrophobicity weighted negative area standard procedure (from A)
123. hydrophobicity weighted negative area/MW DivideByMW (from A)
124. hydrophobicity weighted negative area/total accessible surface area divide by total accessible surface area (from A)
125. hydrophobicity weighted negative area/total accessible surface area/MW DivideByMW (from A)
126. hydrophobicity weighted negative area/total accessible surface area^2 square (from A)
127. hydrophobicity weighted negative area^2 square (from A)
128. hydrophobicity weighted positive area standard procedure (from A)
129. hydrophobicity weighted positive area/MW DivideByMW (from A)
130. hydrophobicity weighted positive area/total accessible surface area divide by total accessible surface area (from A)
131. hydrophobicity weighted positive area/total accessible surface area/MW DivideByMW (from A)
132. hydrophobicity weighted positive area/total accessible surface area^2 square (from A)
133. hydrophobicity weighted positive area^2 square (from A)
134. hydroxyl count hydroxyl (from A)
135. hydroxyl count/MW DivideByMW (from A)
136. length extract from sample (from A)
137. length/MW DivideByMW (from A)
138. length/width DivideBywidth (from A)
139. length/width/MW DivideByMW (from A)
140. length/width^2 square (from A)
141. length>10 length>10 (from A)
142. length>10/MW DivideByMW (from A)
143. length>11 length>11 (from A)
144. length>11/MW DivideByMW (from A)
145. length>12 length>12 (from A)
146. length>12/MW DivideByMW (from A)
147. length>13 length>13 (from A)
148. length>13/MW DivideByMW (from A)
149. length>14 length>14 (from A)
150. length>14/MW DivideByMW (from A)
151. length>15 length>15 (from A)
152. length>15/MW DivideByMW (from A)
153. length>16 length>16 (from A)
154. length>16/MW DivideByMW (from A)
155. length>17 length>17 (from A)
156. length>17/MW DivideByMW (from A)
157. length>7 length>7 (from A)
158. length>7/MW DivideByMW (from A)
159. length>8 length>8 (from A)

160. length>8/MW DivideByMW (from A)
161. length>9 length>9 (from A)
162. length>9/MW DivideByMW (from A)
163. length^2 square (from A)
164. ln(Carbon count) ln (from A)
165. ln(Csp^2 bonded to 1 C) ln (from A)
166. ln(Csp^3 bonded to 2 C) ln (from A)
167. ln(H-bond acceptor count) ln (from A)
168. ln(H-bond donor count) ln (from A)
169. ln(Hydrogen count) ln (from A)
170. ln(Nitrogen count) ln (from A)
171. ln(Oxygen count) ln (from A)
172. ln(all bond count) ln (from A)
173. ln(all count) ln (from A)
174. ln(box area) ln (from A)
175. ln(box cross section) ln (from A)
176. ln(box volume) ln (from A)
177. ln(depth) ln (from A)
178. ln(double bond count) ln (from A)
179. ln(hydrogen donor partial surface area) ln (from A)
180. ln(hydrogen donor partial surface area/total accessible surface area) ln (from A)
181. ln(hydrophobic dipole) ln (from A)
182. ln(hydrophobicity weighted negative area) ln (from A)
183. ln(hydrophobicity weighted negative area/total accessible surface area) ln (from A)
184. ln(hydrophobicity weighted positive area) ln (from A)
185. ln(hydrophobicity weighted positive area/total accessible surface area) ln (from A)
186. ln(length) ln (from A)
187. ln(length/width) ln (from A)
188. ln(molecular weight) ln (from A)
189. ln(nonpolar area) ln (from A)
190. ln(rotatable bond count nonterminal) ln (from A)
191. ln(rotatable bond count) ln (from A)
192. ln(single bond count) ln (from A)
193. ln(total accessible surface area) ln (from A)
194. ln(total positive charges) ln (from A)
195. ln(width) ln (from A)
196. ln(width/depth) ln (from A)
197. log P atom typing scheme (from A)
198. log P/MW DivideByMW (from A)
199. log P/box area DivideBybox area (from A)
200. log P/box cross section DivideBybox cross section (from A)
201. log P/box cross section^2 square (from A)
202. log P/box volume DivideBybox volume (from A)
203. log P/width/depth DivideBywidth/depth (from A)
204. log P/width/depth/MW DivideByMW (from A)
205. log P/width/depth^2 square (from A)
206. log P^2 square (from A)

207. methyl count methyl (from A)
208. methyl count/MW DivideByMW (from A)
209. methyl count^2 square (from A)
210. methylene count methylene (from A)
211. methylene count/MW DivideByMW (from A)
212. methylene count^2 square (from A)
213. molecular weight extract from sample [mass_au] (from A)
214. molecular weight>240 molecular weight>240 (from A)
215. molecular weight>240/MW DivideByMW (from A)
216. molecular weight>260 molecular weight>260 (from A)
217. molecular weight>260/MW DivideByMW (from A)
218. molecular weight>280 molecular weight>280 (from A)
219. molecular weight>280/MW DivideByMW (from A)
220. molecular weight>300 molecular weight>300 (from A)
221. molecular weight>300/MW DivideByMW (from A)
222. molecular weight>320 molecular weight>320 (from A)
223. molecular weight>320/MW DivideByMW (from A)
224. molecular weight>340 molecular weight>340 (from A)
225. molecular weight>340/MW DivideByMW (from A)
226. molecular weight>360 molecular weight>360 (from A)
227. molecular weight>360/MW DivideByMW (from A)
228. molecular weight^2 square (from A)
229. nonpolar area standard procedure (from A)
230. nonpolar area/MW DivideByMW (from A)
231. nonpolar area/total accessible surface area divide by accessible area (from A)
232. nonpolar area/total accessible surface area/MW DivideByMW (from A)
233. nonpolar area^2 square (from A)
234. phenol count phenol (from A)
235. phenol count/MW DivideByMW (from A)
236. ring count 5 member five-membered rings (from A)
237. ring count 5 member/MW DivideByMW (from A)
238. ring count 5 member^2 square (from A)
239. ring count 6 member six-membered rings (from A)
240. ring count 6 member/MW DivideByMW (from A)
241. ring count 6 member^2 square (from A)
242. ring count all all rings (from A)
243. ring count all aromatic all aromatic rings (from A)
244. ring count all aromatic/MW DivideByMW (from A)
245. ring count all aromatic^2 square (from A)
246. ring count all nonaromatic all nonaromatic rings (from A)
247. ring count all nonaromatic/MW DivideByMW (from A)
248. ring count all nonaromatic^2 square (from A)
249. ring count all/MW DivideByMW (from A)
250. ring count all^2 square (from A)
251. ring count aromatic 5 five-membered aromatic rings (from A)
252. ring count aromatic 5/MW DivideByMW (from A)
253. ring count aromatic 5^2 square (from A)
254. ring count aromatic 6 six-membered aromatic rings (from A)
255. ring count aromatic 6/MW DivideByMW (from A)
256. ring count aromatic 6^2 square (from A)

257. ring count nonaromatic 5 five-membered nonaromatic rings (from A)
258. ring count nonaromatic 5/MW DivideByMW (from A)
259. ring count nonaromatic 5^2 square (from A)
260. ring size largest size of largest ring (from A)
261. ring size largest/MW DivideByMW (from A)
262. ring size largest^2 square (from A)
263. ring size smallest size of smallest ring (from A)
264. ring size smallest/MW DivideByMW (from A)
265. ring size smallest^2 square (from A)
266. rotatable bond count nonterminal nonterminal single bonds not in rings (from A)
267. rotatable bond count nonterminal/MW DivideByMW (from A)
268. rotatable bond count nonterminal^2 square (from A)
269. rotatable bond count single bonds not in rings (from A)
270. rotatable bond count/MW DivideByMW (from A)
271. rotatable bond count^2 square (from A)
272. sec-amine count sec-amine (from A)
273. sec-amine count/MW DivideByMW (from A)
274. sec-amine count^2 square (from A)
275. single bond count single bonds (from A)
276. single bond count/MW DivideByMW (from A)
277. single bond count^2 square (from A)
278. sqrt(Carbon count) square root (from A)
279. sqrt(Csp^2 bonded to 1 C) square root (from A)
280. sqrt(Csp^3 bonded to 2 C) square root (from A)
281. sqrt(H-bond acceptor count) square root (from A)
282. sqrt(H-bond donor count) square root (from A)
283. sqrt(Hydrogen count) square root (from A)
284. sqrt(Nitrogen count) square root (from A)
285. sqrt(Oxygen count) square root (from A)
286. sqrt(all bond count) square root (from A)
287. sqrt(all count) square root (from A)
288. sqrt(box area) square root (from A)
289. sqrt(box cross section) square root (from A)
290. sqrt(box volume) square root (from A)
291. sqrt(depth) square root (from A)
292. sqrt(double bond count) square root (from A)
293. sqrt(hydrogen donor partial surface area) square root (from A)
294. sqrt(hydrogen donor partial surface area/total accessible surface area) square root (from A)
295. sqrt(hydrophobic dipole) square root (from A)
296. sqrt(hydrophobicity weighted negative area) square root (from A)
297. sqrt(hydrophobicity weighted negative area/total accessible surface area) square root (from A)
298. sqrt(hydrophobicity weighted positive area) square root (from A)
299. sqrt(hydrophobicity weighted positive area/total accessible surface area) square root (from A)
300. sqrt(length) square root (from A)
301. sqrt(length/width) square root (from A)
302. sqrt(molecular weight) square root (from A)

303. sqrt(nonpolar area) square root (from A)
304. sqrt(rotatable bond count nonterminal) square root (from A)
305. sqrt(rotatable bond count) square root (from A)
306. sqrt(single bond count) square root (from A)
307. sqrt(total accessible surface area) square root (from A)
308. sqrt(total positive charges) square root (from A)
309. sqrt(width) square root (from A)
310. sqrt(width/depth) square root (from A)
311. sulfide count sulfide (from A)
312. sulfide count/MW DivideByMW (from A)
313. sulfide count^2 square (from A)
314. tertiary-amine count tert-amine (from A)
315. tertiary-amine count/MW DivideByMW (from A)
316. total accessible surface area standard procedure [angstrom^2] (from A)
317. total accessible surface area/MW DivideByMW (from A)
318. total accessible surface area^2 square (from A)
319. total negative charges/MW DivideByMW (from A)
320. total positive charges extract from sample (from A)
321. total positive charges/MW DivideByMW (from A)
322. total positive charges^2 square (from A)
323. width extract from sample (from A)
324. width/MW DivideByMW (from A)
325. width/depth DivideBydepth (from A)
326. width/depth/MW DivideByMW (from A)
327. width/depth^2 square (from A)
328. width>10 width>10 (from A)
329. width>10/MW DivideByMW (from A)
330. width>7 width>7 (from A)
331. width>7/MW DivideByMW (from A)
332. width>8 width>8 (from A)
333. width>8/MW DivideByMW (from A)
334. width>9 width>9 (from A)
335. width>9/MW DivideByMW (from A)
336. width^2 square (from A)

Descriptors with no variance have been omitted from the list above.

SCIGRESS-extracted report for QSAR model of tripeptide DPP IV inhibitors

Part 1: Summary

a. Property predicted

The property used to develop the QSAR is named *log10ic50 (from A)* and its values were obtained from a training set of 33 chemical samples.

b. Best QSAR equation

Using the *Custom QSAR* option, the following regression equation, which is the best of 184 descriptors selected with Enhanced Replacement Method, gave the highest correlation coefficient (r^2=0.8293).

log10ic50 = 0.5126**Csp^3 bonded to 3 C* - 0.1210**methyl count^2* + 260.4559**ln(molecular weight)* + 3.47930e+04**1.0/molecular weight* - 17.1577**sqrt(molecular weight)* - 1301.7825

c. Quality of the best QSAR equation

The cross-validated correlation coefficient (cvr^2= 0.7099) suggests that the stability of the equation on addition of similar training data is likely to be reasonable as it is above 0.70. A more detailed analysis is provided in Part 2.

The average error for the training set is 0.2088 and the standard deviation is 0.2958.

The F-ratio is 26.2274. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05 (1 in 20), there is at least one significant descriptor in the model, i.e. this is a valid and stable equation. A probability above 0.05 indicates that the equation might be a chance correlation and not stable.

Based on the partial-F value of each descriptor, there is a near 93% probability that all descriptors are significant.

The training data and QSAR predictions were checked for the following and warnings were noted and are discussed in more detail in Part 2:

1. There are enough observed data values per descriptor.
2. The data is not evenly distributed in thirds, with bin ratios: 5:26:2.
3. The training set had these notes:

   WRA:
   Nonbonded atoms H22 and H57 are too close.
   WRE:
   Nonbonded atoms H22 and H56 are too close.
   WRF:
   Nonbonded atoms H22 and H59 are too close.
   WRH:

Nonbonded atoms H23 and O26 are too close.
WRP:
Only sample in set with an ester count.
WRQ:
Nonbonded atoms H22 and H58 are too close.
WWW:
The depth of 11.377 is more than 3 sigma from the average of 8.174.

LPL:

The error in the prediction is more than three times the standard deviation.

An independent set of chemical samples should be used to test this equation.

d. Applicable prediction range and chemical space

This equation should be used to estimate only data values that fall within the training range from 0.869 to 4.602. Predictions that fall outside this range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that have properties in these ranges:

1. *Csp^3 bonded to 3 C* from 0.000 to 2.000.
2. *methyl count^2* from 0.000 to 16.000.
3. *ln(molecular weight)* from 5.494 to 6.357.
4. *1.0/molecular weight* from 0.002 to 0.004.
5. *sqrt(molecular weight)* from 15.597 to 24.013.

Predictions for chemical samples with properties that fall outside the training set property range should be treated with caution as there is no way to know if the correlation holds outside the training set range.

The QSAR should be used to predict values only for chemical samples that are chemically similar to the training set or share a common mode of action. The 33 samples in the training set included the following elements (min:max:# samples): H(17:38:33) C(10:33:33) N(3:10:33) O(4:6:33) S(0:1:3). The 33 samples in the training set included the following functional groups: *ring*, *ring 5 member*, *sec-amine*, *ring aromatic*, *ring 6 member*, *H-bond acceptor*, *hydroxyl*, *guanidine*, *amide*, *ring aromatic 6*, *phenol*, *H-bond donor*, *methyl*, *sulfide*, *ring aromatic 5*, *amine*, *tertiary-amine*.

e. Mechanistic interpretation

The descriptors and their relative importance are listed below:

Descriptor	Relative importance
Csp^3 bonded to 3 C	0.0053
methyl count^2	-0.0091
ln(molecular weight)	1.0000
1.0/molecular weight	0.3657
sqrt(molecular weight)	-0.6413

Part 2: Detailed Analysis

Data distribution

There are 6.600 data values per descriptor in the QSAR model.

The QSAR equation was derived using a training set of 33 chemical samples with a three-sigma range for log10ic50 (from A) from 0.543 to 4.772. The average was 2.658 and the standard deviation was 0.705 with a minimum data value of 0.869 and a maximum of 4.602. The data skewness measure is 0.378. The data skewness is between -2.0 and 2.0 which indicates that the data is not skewed. Partitioning the data into equal thirds from lowest to highest data values gives three bins with these counts: 5:26:2.

Chemical samples

Chemical samples in the training set had molecular weights from 243.26 to 576.645 and these elements and counts:

Element	Lowest	Highest
Hydrogen	17	38
Carbon	10	33
Nitrogen	3	10
Oxygen	4	6
Sulfur	0	1

The chemical samples had these charges (and counts): +0(17) +1(13) +2(3).

The chemical samples had these groups and counts:

Group	Lowest	Highest
ring	1	6
ring 5 member	1	3
sec-amine	0	4
ring aromatic	0	6
ring 6 member	0	3
H-bond acceptor	6	11
hydroxyl	0	1
guanidine	0	2
amide	1	2
ring aromatic 6	0	3
phenol	0	2
H-bond donor	3	16
methyl	0	4

sulfide	0	1
ring aromatic 5	0	3
amine	0	4
tertiary-amine	0	1

Chemical samples were preconditioned by electronic structure for QSAR.

The training set samples had these notes:

WRA:
Nonbonded atoms H22 and H57 are too close.

WRE:
Nonbonded atoms H22 and H56 are too close.

WRF:
Nonbonded atoms H22 and H59 are too close.

WRH:
Nonbonded atoms H23 and O26 are too close.

WRP:
Only sample in set with an ester count.

WRQ:
Nonbonded atoms H22 and H58 are too close.

WWW:
The depth of 11.377 is more than 3 sigma from the average of 8.174.

Analysis of QSAR equation

The following equation predicts log10ic50 (from A):

log10ic50 = 0.5126*Csp^3 bonded to 3 C - 0.1210*methyl count^2 + 260.4559*ln(molecular weight) + 3.47930e+04*1.0/molecular weight - 17.1577*sqrt(molecular weight) - 1301.7825

The average error for the training set is 0.2088. The standard deviation of the error is 0.2958.

r^2 is 0.8293; the degrees-of-freedom adjusted r^2 is 0.7976; and the leave-one-out cross-validated r^2 is 0.7099. The standard deviation in the error predicted by leave-one-out cross-validation is 0.3855. The F-ratio is 26.2274. The probability that a greater F-ratio can be obtained by chance alone is 0.0000. Since the probability is less than 0.05, there is at least one significant descriptor in the model.

Use the normalized coefficients in the following analysis section to interpret the relative importance of each descriptor.

82.9% of the variability in log10ic50 (from A) is explained by this equation.

The relative weight of each normalized contribution is:

Descriptor	Coefficient	Normalized coefficient	Descriptor standard deviation	Partial-F	Probability of greater F-ratio
Csp^3 bonded to 3 C	0.5126	0.0053	0.6629	3.4655	0.0736

methyl count^2	-0.1210	-0.0091	4.8500	11.0810	0.0025
ln(molecular weight)	260.4559	1.0000	0.2478	20.7152	0.0001
1.0/molecular weight	34793.0455	0.3657	0.0007	24.0262	0.0000
sqrt(molecular weight)	-17.1577	-0.6413	2.4121	19.3041	0.0002
Constant	-1301.7825				

The equation is the best of 184 descriptors selected with Enhanced Replacement Method. The correlation between any pair of descriptors that appear in the equation is less than 0.9500

A plot of predicted values against the original data may be found in the 18. Best QSAR plot. Other equations with r^2 from 0.8293 to 0.8293 may be found in the 18. Other QSARs table.

The predicted values had these notes:

LPL:

The error in the prediction is more than three times the standard deviation.

This equation should be used only to predict values of similar chemicals that have molecular weights, elements, groups and charges within the lowest to highest ranges of the training set.

SCIGRESS-extracted report of the list of 184 descriptors for QSAR modeling of tripeptide DPP IV inhibitors

Descriptors

1. 1.0/Carbon count (from D)
2. 1.0/Csp^2 bonded to 1 C (from D)
3. 1.0/Csp^3 bonded to 2 C (from D)
4. 1.0/H-bond acceptor count (from D)
5. 1.0/H-bond donor count (from D)
6. 1.0/Nitrogen count (from D)
7. 1.0/Oxygen count (from D)
8. 1.0/amide count (from D)
9. 1.0/double bond count (from D)
10. 1.0/energy dielectric (from D)
11. 1.0/heat of formation (from D)
12. 1.0/molecular weight (from D)
13. 1.0/polarizability (from D)
14. 1.0/ring count 5 member (from D)
15. 1.0/ring count all (from D)
16. 1.0/ring size largest (from D)
17. 1.0/rotatable bond count (from D)
18. 1.0/rotatable bond count nonterminal (from D)
19. 1.0/total negative charges (from D)
20. 1.0/total positive charges (from D)
21. Carbon count (from D)
22. Carbon count/MW (from D)
23. Carbon count^2 (from D)
24. Csp^2 bonded to 1 C (from D)
25. Csp^2 bonded to 1 C/MW (from D)

26. Csp^2 bonded to 1 C^2 (from D)
27. Csp^2 bonded to 2 C (from D)
28. Csp^2 bonded to 2 C/MW (from D)
29. Csp^2 bonded to 2 C^2 (from D)
30. Csp^2 bonded to 3 C (from D)
31. Csp^2 bonded to 3 C/MW (from D)
32. Csp^2 bonded to 3 C^2 (from D)
33. Csp^3 bonded to 1 C (from D)
34. Csp^3 bonded to 1 C/MW (from D)
35. Csp^3 bonded to 1 C^2 (from D)
36. Csp^3 bonded to 2 C (from D)
37. Csp^3 bonded to 2 C/MW (from D)
38. Csp^3 bonded to 2 C^2 (from D)
39. Csp^3 bonded to 3 C (from D)
40. Csp^3 bonded to 3 C/MW (from D)
41. Csp^3 bonded to 3 C^2 (from D)
42. H-bond acceptor count (from D)
43. H-bond acceptor count/MW (from D)
44. H-bond acceptor count^2 (from D)
45. H-bond donor count (from D)
46. H-bond donor count/MW (from D)
47. H-bond donor count^2 (from D)
48. Nitrogen count (from D)
49. Nitrogen count/MW (from D)
50. Nitrogen count^2 (from D)
51. Oxygen count (from D)
52. Oxygen count/MW (from D)
53. Oxygen count^2 (from D)
54. Sulfur count (from D)
55. Sulfur count/MW (from D)
56. abs(charge) weighted area (from D)
57. all bond count (from D)
58. all bond count/MW (from D)
59. all bond count^2 (from D)
60. amide count (from D)
61. amide count/MW (from D)
62. amide count^2 (from D)
63. amine count (from D)
64. amine count/MW (from D)
65. amine count^2 (from D)
66. atomic charge weighted negative area (from D)
67. atomic charge weighted positive area (from D)
68. atomic charge weighted positive area-atomic charge weighted negative area (from D)
69. charge [charge_au] (from D)
70. charge weighted area (from D)
71. charge weighted polar area (from D)
72. charge/MW (from D)
73. charge^2 (from D)
74. double bond count (from D)
75. double bond count/MW (from D)

76. double bond count^2 (from D)
77. electrophilic weighted area (from D)
78. energy dielectric [kcal/mol] (from D)
79. energy dielectric/MW (from D)
80. energy dielectric^2 (from D)
81. guanidine count (from D)
82. guanidine count/MW (from D)
83. guanidine count^2 (from D)
84. heat of formation [kcal/mol] (from D)
85. heat of formation/MW (from D)
86. heat of formation^2 (from D)
87. hydroxyl count (from D)
88. hydroxyl count/MW (from D)
89. ln(Carbon count) (from D)
90. ln(Csp^2 bonded to 1 C) (from D)
91. ln(Csp^3 bonded to 2 C) (from D)
92. ln(H-bond acceptor count) (from D)
93. ln(H-bond donor count) (from D)
94. ln(Nitrogen count) (from D)
95. ln(Oxygen count) (from D)
96. ln(amide count) (from D)
97. ln(double bond count) (from D)
98. ln(molecular weight) (from D)
99. ln(polarizability) (from D)
100. ln(ring count 5 member) (from D)
101. ln(ring count all) (from D)
102. ln(ring size largest) (from D)
103. ln(rotatable bond count nonterminal) (from D)
104. ln(rotatable bond count) (from D)
105. ln(total positive charges) (from D)
106. methyl count (from D)
107. methyl count/MW (from D)
108. methyl count^2 (from D)
109. molecular weight [mass_au] (from D)
110. molecular weight^2 (from D)
111. nucleophilic weighted area (from D)
112. phenol count (from D)
113. phenol count/MW (from D)
114. phenol count^2 (from D)
115. polarizability [angstrom^3] (from D)
116. polarizability/MW (from D)
117. polarizability^2 (from D)
118. radical weighted area (from D)
119. ring count 5 member (from D)
120. ring count 5 member/MW (from D)
121. ring count 5 member^2 (from D)
122. ring count 6 member (from D)
123. ring count 6 member/MW (from D)
124. ring count 6 member^2 (from D)
125. ring count all (from D)

126. ring count all aromatic (from D)
127. ring count all aromatic/MW (from D)
128. ring count all aromatic^2 (from D)
129. ring count all nonaromatic (from D)
130. ring count all nonaromatic/MW (from D)
131. ring count all nonaromatic^2 (from D)
132. ring count all/MW (from D)
133. ring count all^2 (from D)
134. ring count aromatic 5 (from D)
135. ring count aromatic 5/MW (from D)
136. ring count aromatic 5^2 (from D)
137. ring count aromatic 6 (from D)
138. ring count aromatic 6/MW (from D)
139. ring count aromatic 6^2 (from D)
140. ring count nonaromatic 5 (from D)
141. ring count nonaromatic 5/MW (from D)
142. ring count nonaromatic 5^2 (from D)
143. ring size largest (from D)
144. ring size largest/MW (from D)
145. ring size largest^2 (from D)
146. rotatable bond count (from D)
147. rotatable bond count nonterminal (from D)
148. rotatable bond count nonterminal/MW (from D)
149. rotatable bond count nonterminal^2 (from D)
150. rotatable bond count/MW (from D)
151. rotatable bond count^2 (from D)
152. sec-amine count (from D)
153. sec-amine count/MW (from D)
154. sec-amine count^2 (from D)
155. solvent accessible surf area [angstrom^2] (from D)
156. sqrt(Carbon count) (from D)
157. sqrt(Csp^2 bonded to 1 C) (from D)
158. sqrt(Csp^3 bonded to 2 C) (from D)
159. sqrt(H-bond acceptor count) (from D)
160. sqrt(H-bond donor count) (from D)
161. sqrt(Nitrogen count) (from D)
162. sqrt(Oxygen count) (from D)
163. sqrt(all bond count) (from D)
164. sqrt(amide count) (from D)
165. sqrt(double bond count) (from D)
166. sqrt(molecular weight) (from D)
167. sqrt(polarizability) (from D)
168. sqrt(ring count 5 member) (from D)
169. sqrt(ring count all) (from D)
170. sqrt(ring size largest) (from D)
171. sqrt(rotatable bond count nonterminal) (from D)
172. sqrt(rotatable bond count) (from D)
173. sqrt(total positive charges) (from D)
174. sulfide count (from D)
175. sulfide count/MW (from D)

176. tertiary-amine count (from D)
177. tertiary-amine count/MW (from D)
178. total accessible surface area [angstrom^2] (from D)
179. total negative charges (from D)
180. total negative charges/MW (from D)
181. total negative charges^2 (from D)
182. total positive charges (from D)
183. total positive charges/MW (from D)
184. total positive charges^2 (from D)

Descriptors with no variance have been omitted from the list above.