А. А. Рагойша

Информационные технологии в химии

2-й семестр

Избранные элементы хемоинформатики

Лекция 2

2D-структура

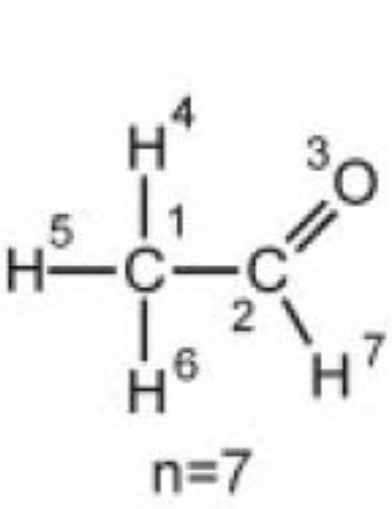
Двумерная форма хранения и обработки информации

Матричная форма представления молекулярного графа

Матрица смежности

Атомы нумеруются произвольно.

n атомов – матрица размерности *n* x *n*



	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

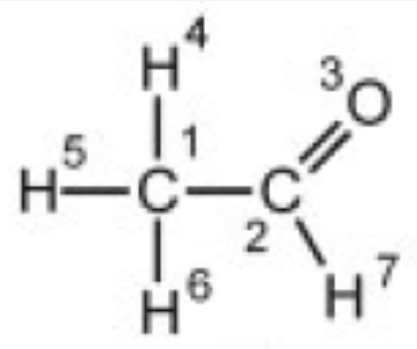
Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Элементы матрицы:

a_{ij} = 1, если между атомами ***i*** и ***j*** имеется химическая связь,

a_{ij} = 0 , если между атомами ***i*** и ***j*** нет химической связи.

Избыточная – неизбыточная матрица



	1	2	3	4	5	6	7
1	0	1	0	1	1	1	0
2	1	0	1	0	0	0	1
3	0	1	0	0	0	0	0
4	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	1	0	0	0	0	0

	1	2	3	4	5	6	7
1		1		1	1	1	
2	1		1				1
3		1					
4	1						
5	1						
6	1						
7		1					

Этапы упрощения матрицы:

удаление нулей,

устранение дублей,

удаление информации об атомах водорода.

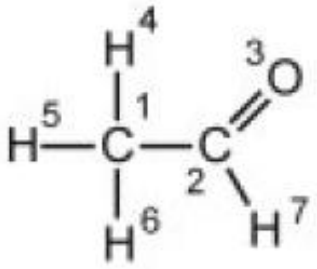
	1	2	3	4	5	6	7
1		1		1	1	1	
2			1				1
3							
4							
5							
6							
7							

	1	2	3
1		1	
2			1
3			

Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Матрица расстояний

(примеры иных типов матриц)



a)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1.400	2.190	1.022	1.023	1.022	2.106
C2	1.400	0	1.123	1.999	1.982	1.999	1.022
O3	2.190	1.123	0	2.349	2.708	2.995	1.859
H4	1.022	1.999	2.349	0	1.668	1.661	2.895
H5	1.023	1.982	2.708	1.668	0	1.668	2.562
H6	1.022	1.999	2.955	1.661	1.668	0	2.336
H7	2.106	1.022	1.859	2.895	2.566	2.336	0

b)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1	2	1	1	1	2
C2	1	0	1	2	2	2	1
O3	2	1	0	3	3	3	2
H4	1	2	3	0	2	2	3
H5	1	2	3	2	0	2	3
H6	1	2	3	2	2	0	3
H7	2	1	2	3	3	3	0

Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

а) геометрическое расстояние (ангстрем);

б) топологическое расстояние (число связей по кратчайшему пути)

Проблема разрастания объема базы данных

В матрице смежности:

Число элементов матрицы = $f(n^2)$

Нерационально для больших молекул.

Значительно лучше, если:

Число элементов = $f(n^1)$

Это достигается в форме
таблицы соединений (connection table).

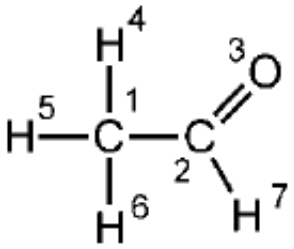
Таблица соединений -1

Таблица соединений – отображение состава вещества и связей между атомами в табличной форме.

Пример: этаналь.

Пронумеровать атомы в производном порядке.

Один из путей: Заполнить две таблицы.



Список атомов	
1	C
2	C
3	O
4	H
5	H
6	H
7	H

Список связей		
1-й атом	2-й атом	Порядок связи
1	2	1
2	3	2
2	7	1
1	4	1
1	5	1
1	6	1

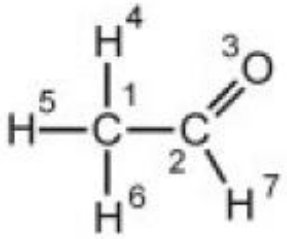
Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Таблица соединений - 2 (избыточная)

Пример: этаналь; второй путь.

Пронумеровать атомы в производном порядке.

Заполнить одну таблицу.



№	Атом	Сосед № 1	Порядок связи	Сосед № 2	Порядок связи	Сосед № 3	Порядок связи	Сосед № 4	Порядок связи
1	C	2	1	4	1	5	1	6	1
2	C	1	1	3	2	7	1		
3	O	2	2						
4	H	1	1						
5	H	1	1						
6	H	1	1						
7	H	2	1						

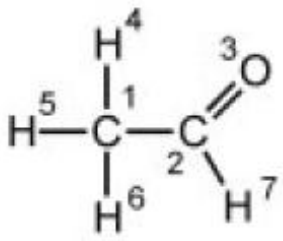
Информативность избыточна, т.к.
каждый атом упоминается дважды,
сведения о водороде стандартны.

Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Таблица соединений - 2 (неизбыточная)

Если убрать повторы, сжать, получаем:

В случае "обычных" органических соединений полезная информация при этом не теряется.



№	Атом	Сосед № 1	Порядок связи	Сосед № 2	Порядок связи
1	C	2	1		
2	C			3	2
3	O				

⇒

№	Атом	Сосед № 1	Порядок связи
1	C	2	1
2	C	3	2
3	O		

Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Информация о структуре:

- из измерительной аппаратуры,
- из молекулярных редакторов,
- из программ расчета.

Форматы разнообразны,
необходим стандарт обмена информацией.

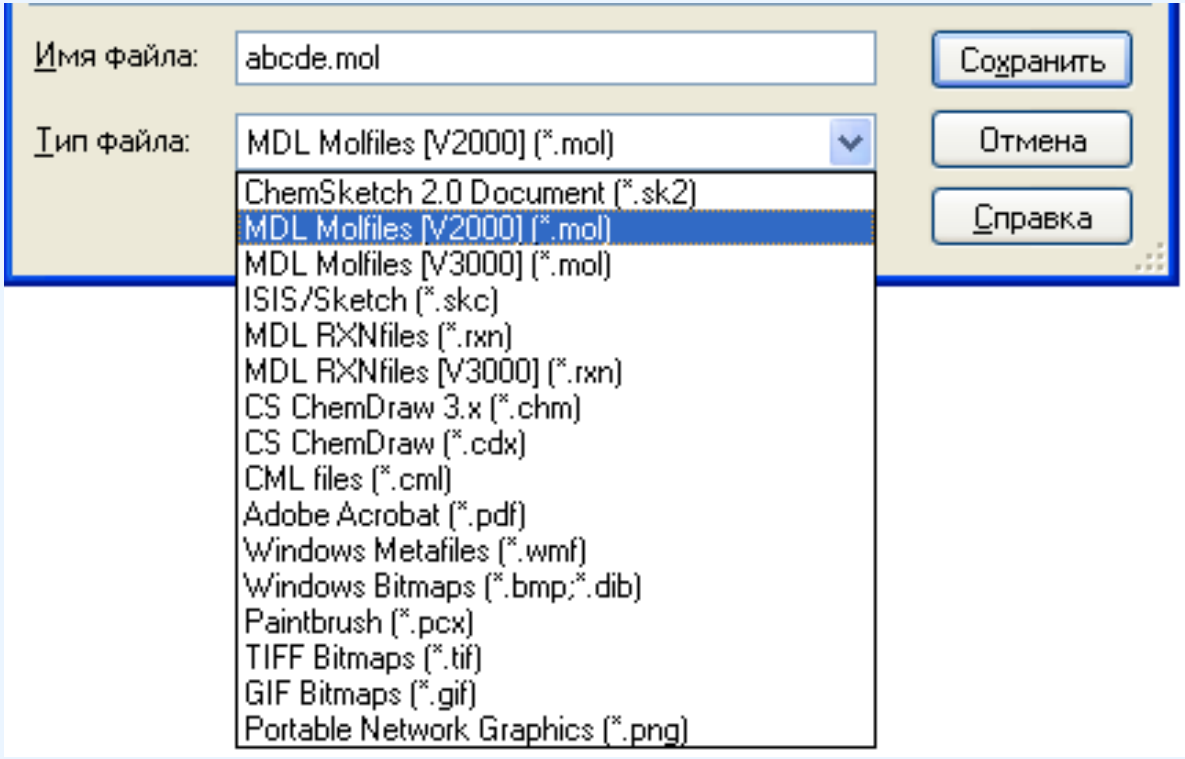
Де-факто:

MOL-файлы

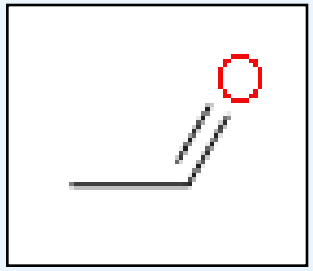
*abcde.**mol***

(есть варианты).

В основе MOL-файла – таблица соединений.



MOL-файл (2D, без атомов H)



3 атома

2 связи

список атомов

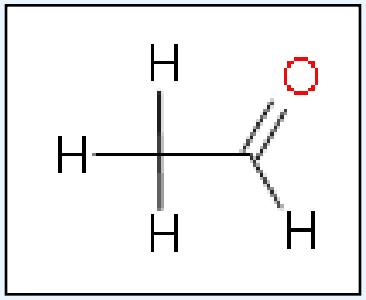
иные параметры

координаты
x, y, z

```
  SMXDraw01201020342D

  3  2  0  0  0  0  0  0  0  0999 V2000
    4.7059   -6.9865    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.5327   -6.9865    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.9461   -6.2705    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
M  END
```

список связей



MOL-файл (2D, с атомами H)

```
  SMMXDraw01201020342D

  7  6  0  0  0  0  0  0  0  0999 V2000
    4.9749   -5.2654    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.7945   -4.4244    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.8108   -5.9984    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    5.8016   -5.2654    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
    7.0418   -5.9814    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    7.0418   -4.5494    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    6.6284   -5.2654    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  4  1  1  0  0  0  0
  4  2  1  0  0  0  0
  4  3  1  0  0  0  0
  7  4  1  0  0  0  0
  7  5  1  0  0  0  0
  7  6  2  0  0  0  0
M  END
```

Кратко о структурной базе данных

Двумерная структура в запросе

Формирование структуры на поисковом бланке (апплет).

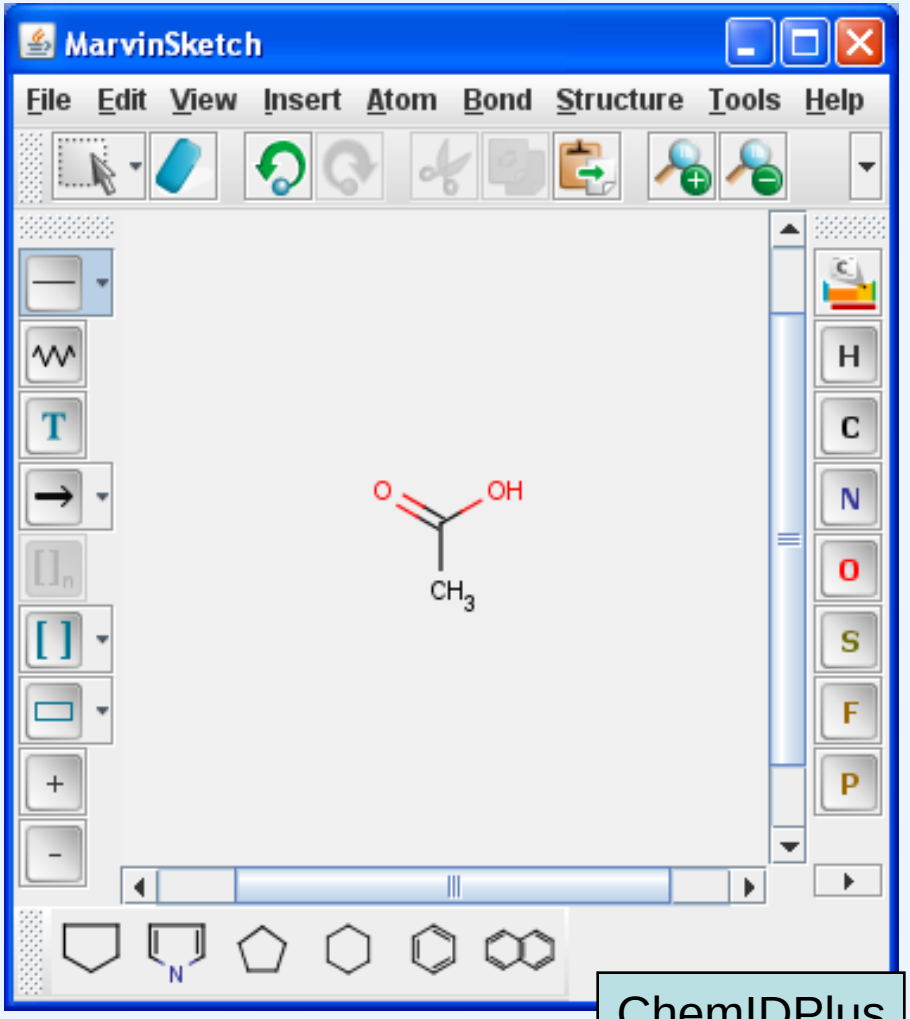
Загрузка файла пользователя (обычно MOL).

Search for Species Data by Structure or Substructure

There are three structure search options available:

1. Use applet to draw a structure.
   This option requires a Java capable browser.
2. Submit a mol (MDL) file containing the structure. This option requires a browser which can upload files.
3. Specify structure properties and subgrou
   form.

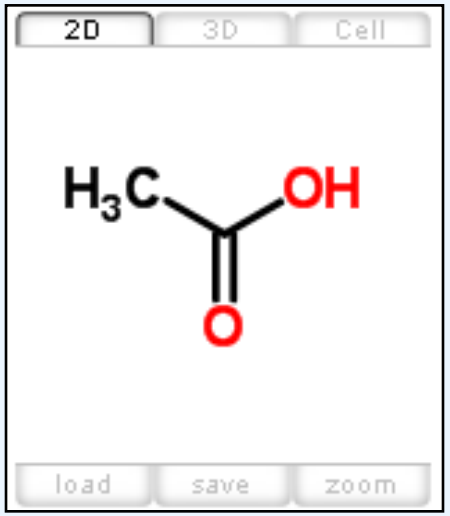
NIST



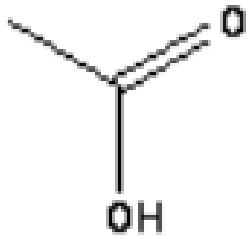
ChemIDPlus

Структура в результатах поиска

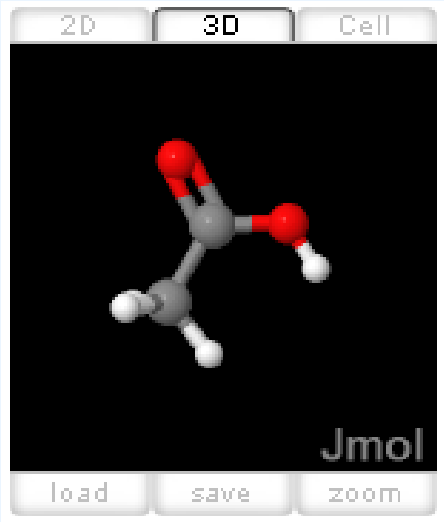
- Изображение на странице поиска
- Ссылка на файл



- **Chemical structure:**

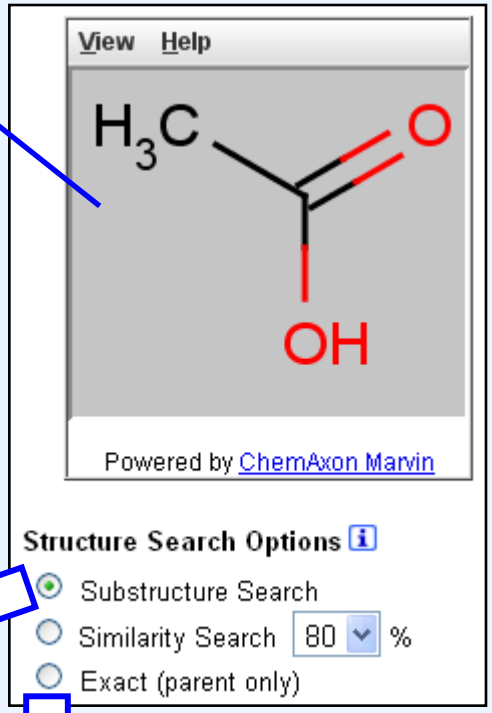


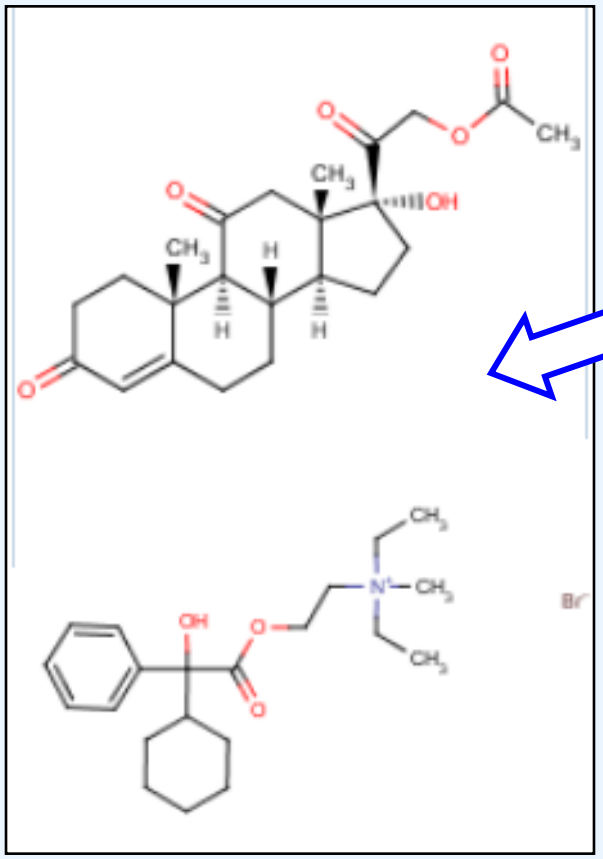
This structure is also available as a 2d Mol file or as a computed 3d Mol file.



Структура, подструктура (субструктура)

Запрос





Поиск идентичной структуры и поиск структур, имеющих заданный остов,- профессионалы такие задачи считают тривиальными.

Молекулярное (химическое) подобие

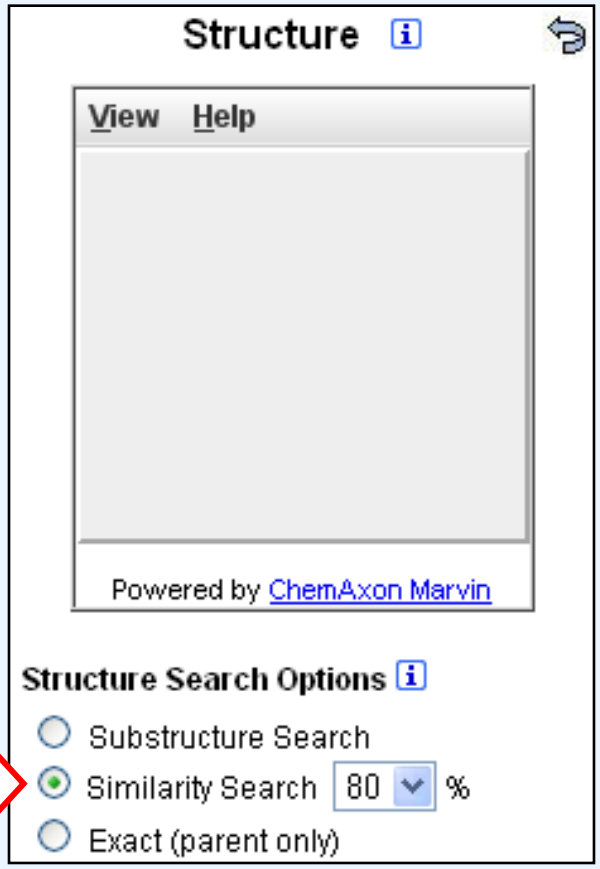
Молекулярное (химическое) подобие Similarity

Молекулярное подобие - это близость, сходство, подобие структур химических соединений.

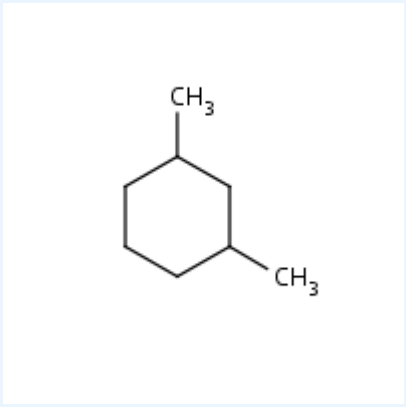
Предполагается, что **подобные соединения обладают близкими химическими свойствами**, в т. ч. подобной биологической активностью.

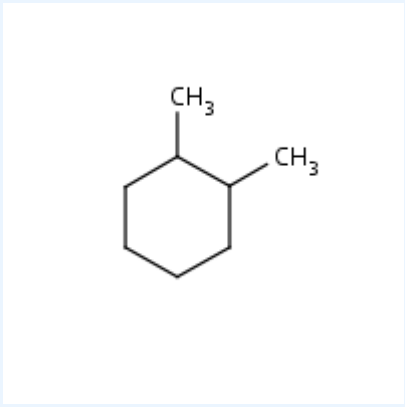
(Предположение не всегда верно).

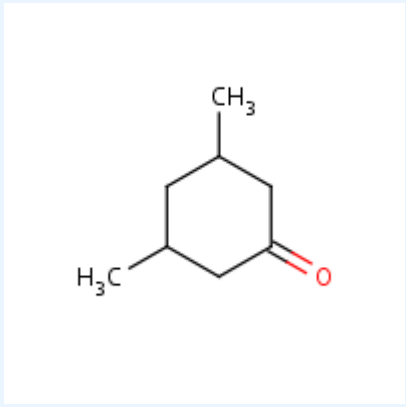
поиск подобных структур

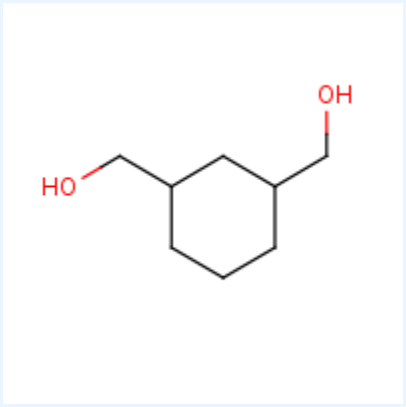


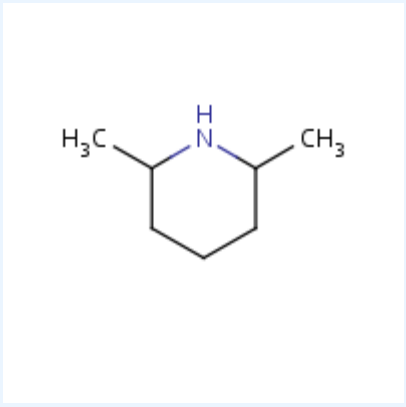
Пример: подобные структуры

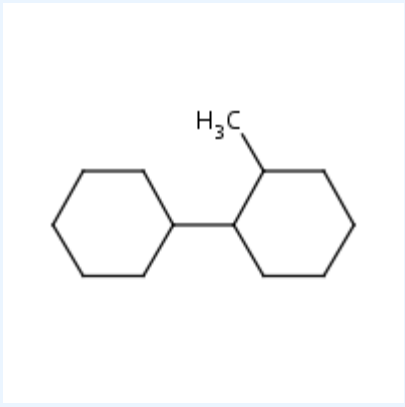


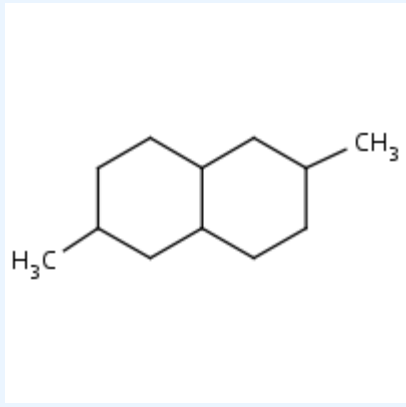


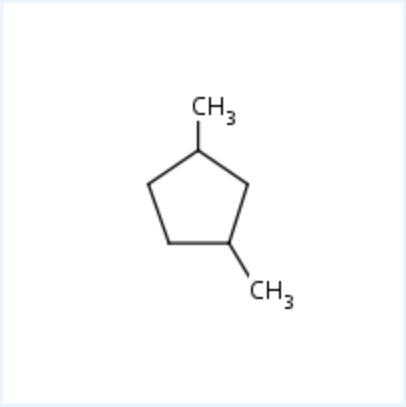












Как количественно охарактеризовать степень подобия?

Строка битов (Bitstring)

Строка битов – отображение структуры последовательностью бинарных чисел 0 и 1.

Формируют словарь признаков (например, разные группы атомов).

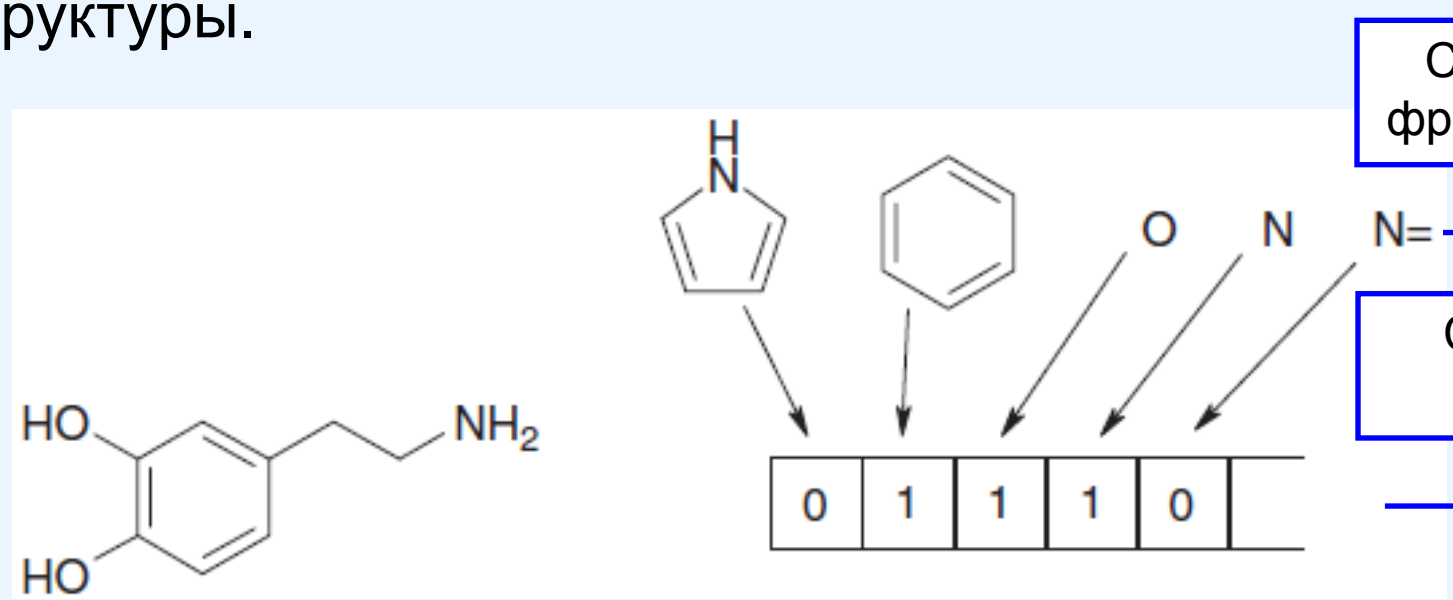
В строке битов на ***i***-й позиции отмечают ***i***-й признак:
1 – признак имеется,
0 – признак отсутствует.

Для каждой структуры генерируют строку битов, например:

011100110001000...

Структурный код (Structural key)

Структурный код – строка битов, в которой отражено наличие или отсутствие определенного фрагмента структуры.

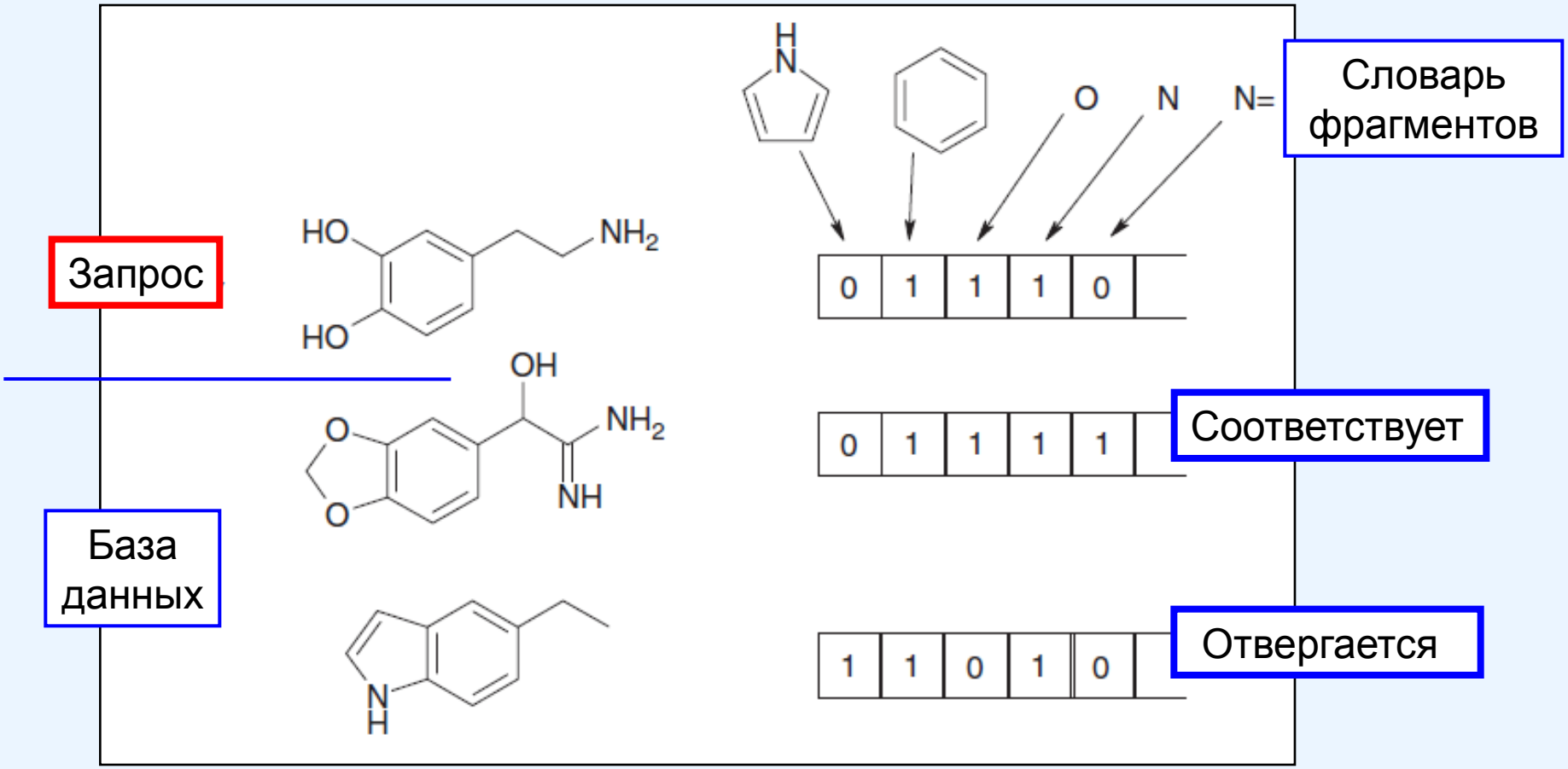


Полученная строка битов – отпечаток пальцев (fingerprint) данной *двумерной* структуры.

A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

Отпечатки пальцев в структурном поиске

Первая стадия структурного поиска: скрининг, т.е. отсеивание заведомо ненужной информации



A.R.Leach, V.J.Gillet. An Introduction to Chemoinformatics. 2007.

Количественная оценка подобия
Коэффициент Танимото

Сравнивают отпечатки пальцев двух структур.

Напоминание: отпечаток пальцев – строка битов.

1	0	1	1	1	0

Для двух структур A и B:

S_{AB} – коэффициент Танимото,

a – количество "единиц" у A,

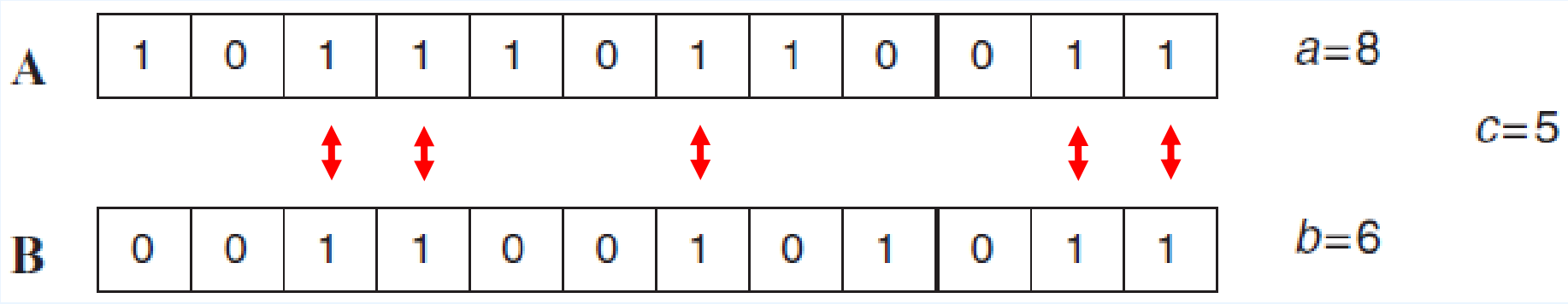
b – количество "единиц" у B,

c – количество "единиц", общих для A и B.

$$S_{AB} = \frac{c}{a + b - c}$$

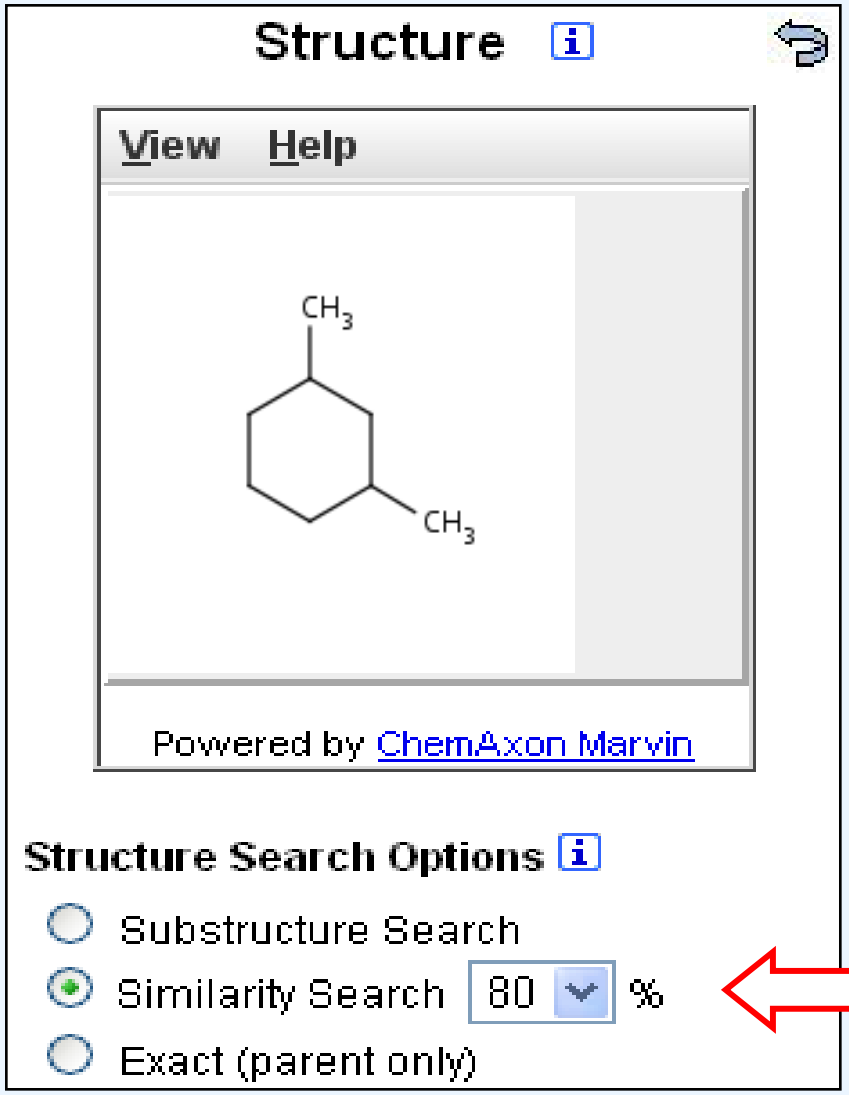
$$0 < S_{AB} < 1$$

Пример расчета коэффициента Танимото



$$S_{AB} = \frac{5}{8+6-5} = 0.56$$

Структуры A и B подобны на 56 %



Увеличить
или уменьшить
список
результатов
поиска

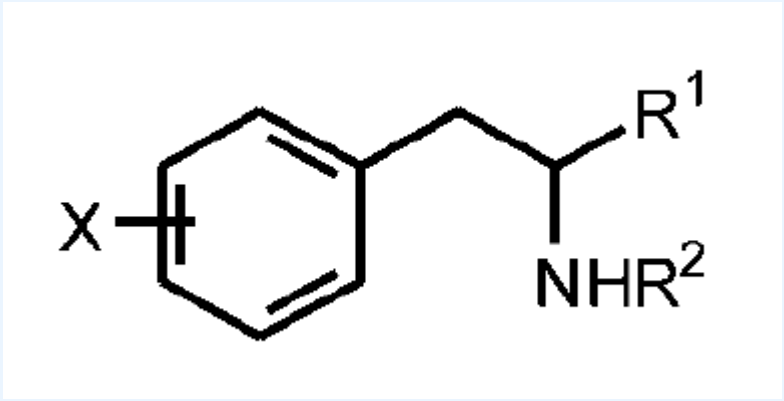
Структуры Маркуша

Структура Маркуша
(в патентных базах данных)

Структура Маркуша – способ отображения серии соединений с помощью общего для них ядра и варьируемых частей.

Варьируемые части записываются отдельно от графической формулы.

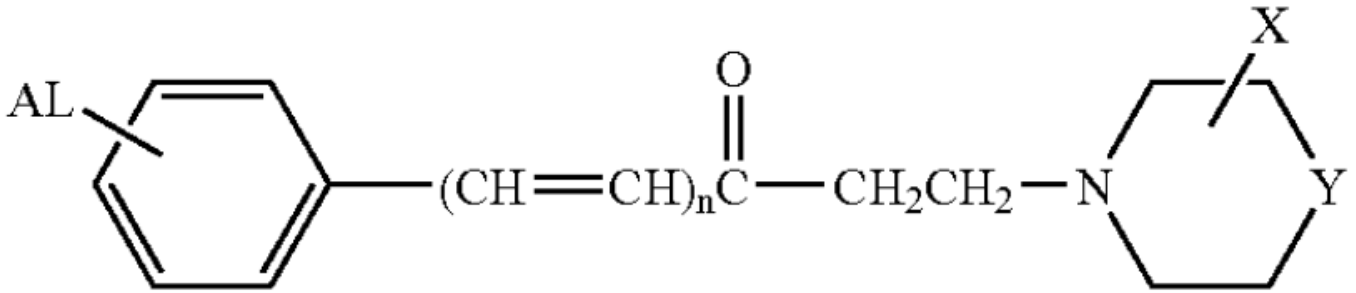
Одна структура Маркуша может отображать собой множество соединений разных классов.



R^1 = H, OH, COOH, галоген

R^2 = H, CH_3

X = H, $(CH_2)_nCH_3$



characterized in that AL is selected from the group consisting of hydrogen, hydroxy, halogen (F, Cl, Br, I), CF_3, CN, NO_2, NR_1R_2 (R_1, R_2=C_{1-6} alkyl), C_{1-6} alkyl, C_{1-6} alkoxy, methylenedioxy, 3,4-di-C_{1-6} alkoxy, 3,4,5-tri-C_{1-6} alkoxy, 3-methoxy-4-hydroxy, 3,4-methylenedioxy-5-methoxy, 3-hydroxy-4-methoxy;

n=0, 1, 2;

Y is selected from the group consisting of C, N, O;

X is selected from the group consisting of hydrogen, C_{1-6} alkyl, COOR (R=hydrogen, C_{1-6} alkyl, $C(CH_3)_3$, substituted or unsubstituted aryl, CO-Ph, CH_2Ph, CH_2CH_2OH, $CONR_1R_2$ (R_1, R_2=C_{1-6} alkyl).

Структуры Маркуша
в патентных базах данных:

Одной формуле могут соответствовать триллионы структур ---

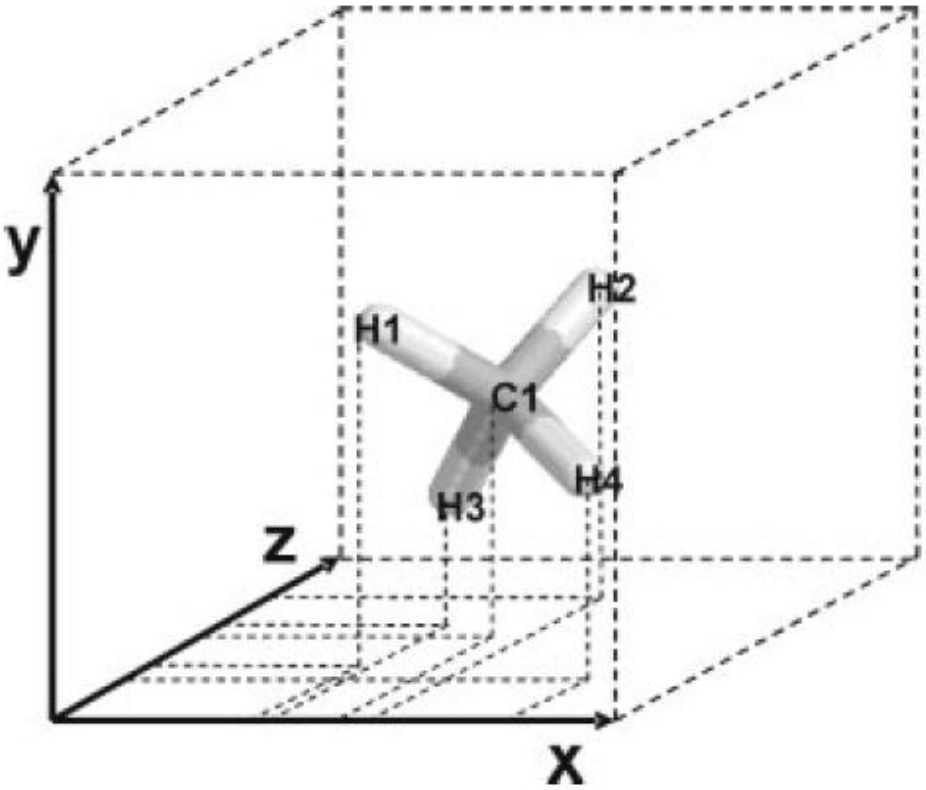
на много порядков больше,
чем число известных веществ.

- Как проверить свойства каждого из запатентованных веществ?
- Поиск по формуле Маркуша — только в коммерческих базах данных.

Трехмерные структуры
3D-структуры

Двумерная форма
хранения информации
о молекулярной структуре

Метан
Декартовы координаты

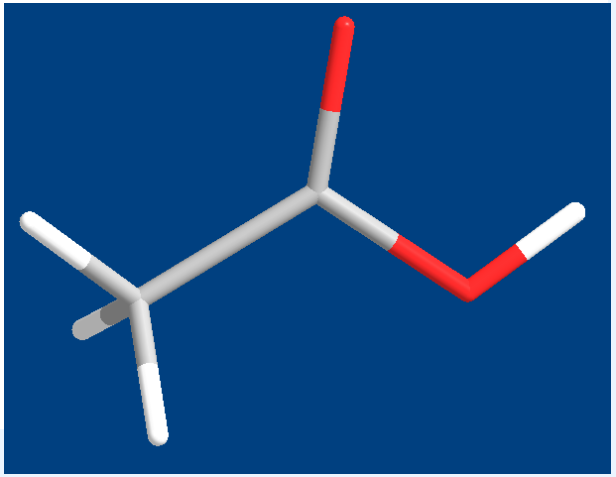


	x	y	z
C1	-0.0127	1.0858	0.0080
H1	0.0021	-0.0041	0.0020
H2	1.0099	1.4631	0.0003
H3	-0.5399	1.4469	-0.8751
H4	-0.5229	1.4373	0.9048

Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

MOL-файл (3D)

```
Acetic acid, ID: C64197
  NIST     04042217093D 1   1.00000     0.00000
NIST Chemistry WebBook
  8  7  0     0  0               1 V2000
    0.7649    0.9627    1.0051 C   0  0  0  0  0  0           0  0  0
    2.0422    1.7381    0.9144 C   0  0  0  0  0  0           0  0  0
    3.1225    1.0260    0.5122 O   0  0  0  0  0  0           0  0  0
    2.2424    2.9166    1.1481 O   0  0  0  0  0  0           0  0  0
    0.0000    1.5240    1.5565 H   0  0  0  0  0  0           0  0  0
    0.3742    0.7552    0.0000 H   0  0  0  0  0  0           0  0  0
    0.9140    0.0000    1.5112 H   0  0  0  0  0  0           0  0  0
    3.8882    1.5907    0.4746 H   0  0  0  0  0  0           0  0  0
  1  2  1  0     0  0
  1  5  1  0     0  0
  1  6  1  0     0  0
  1  7  1  0     0  0
  2  3  1  0     0  0
  2  4  2  0     0  0
  3  8  1  0     0  0
M  END
```

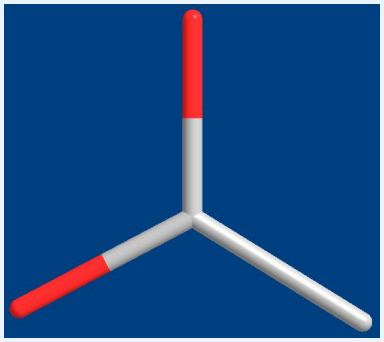


уксусная кислота

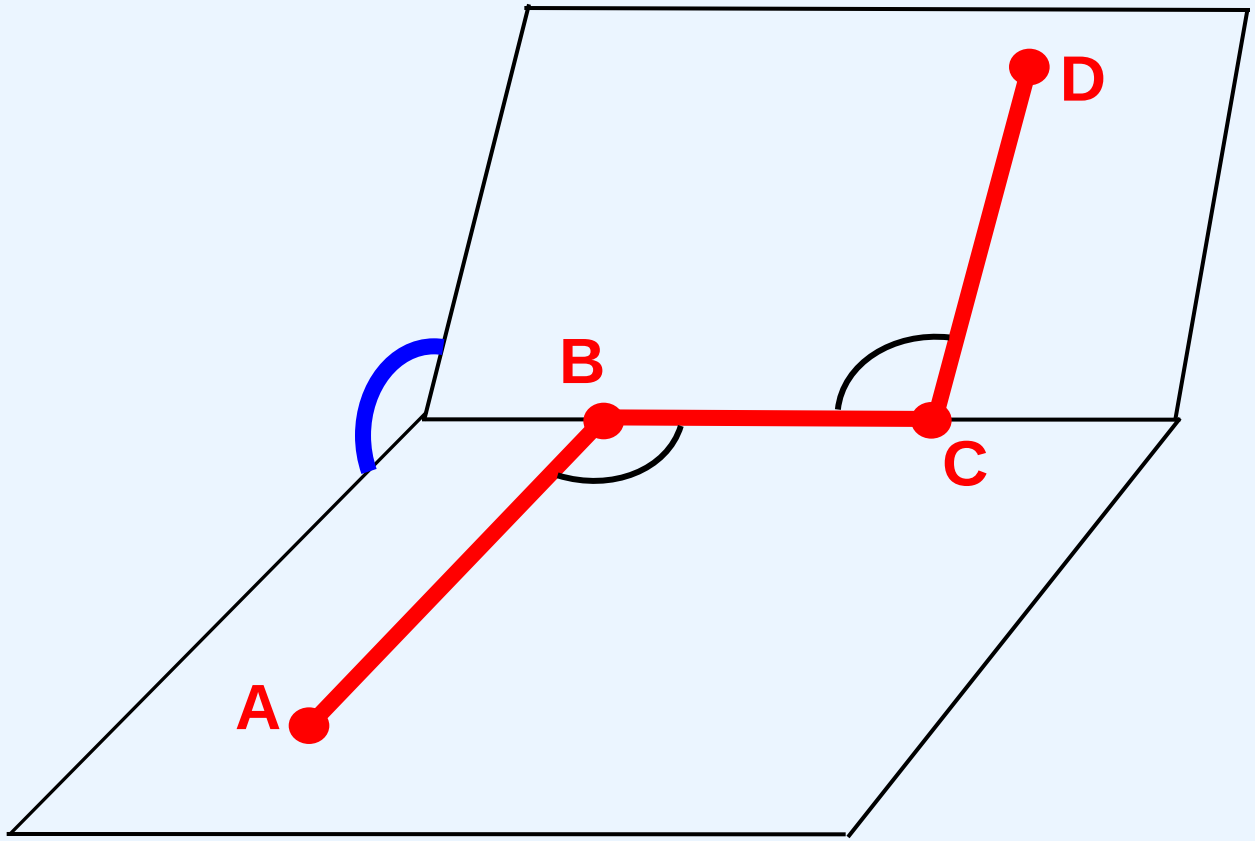
PDB- файл (3D)

```
HEADER    PROTEIN
COMPND    149048
AUTHOR    GENERATED BY BABEL 1.6
ATOM      1  O   UNK     1       0.007   1.510   0.000  1.00  0.00
ATOM      2  O   UNK     1      -1.317  -0.703   0.000  1.00  0.00
ATOM      3  C   UNK     1       0.007  -0.003   0.000  1.00  0.00
ATOM      4  C   UNK     1       1.307  -0.807   0.000  1.00  0.00
CONECT    1    3
CONECT    2    3
CONECT    3    1    2    4
CONECT    4    3
MASTER        0    0    0    0    0    0    0    0    4    0    4    0
END
```

Уксусная кислота.
Атомы водорода не отображены



Внутренние координаты

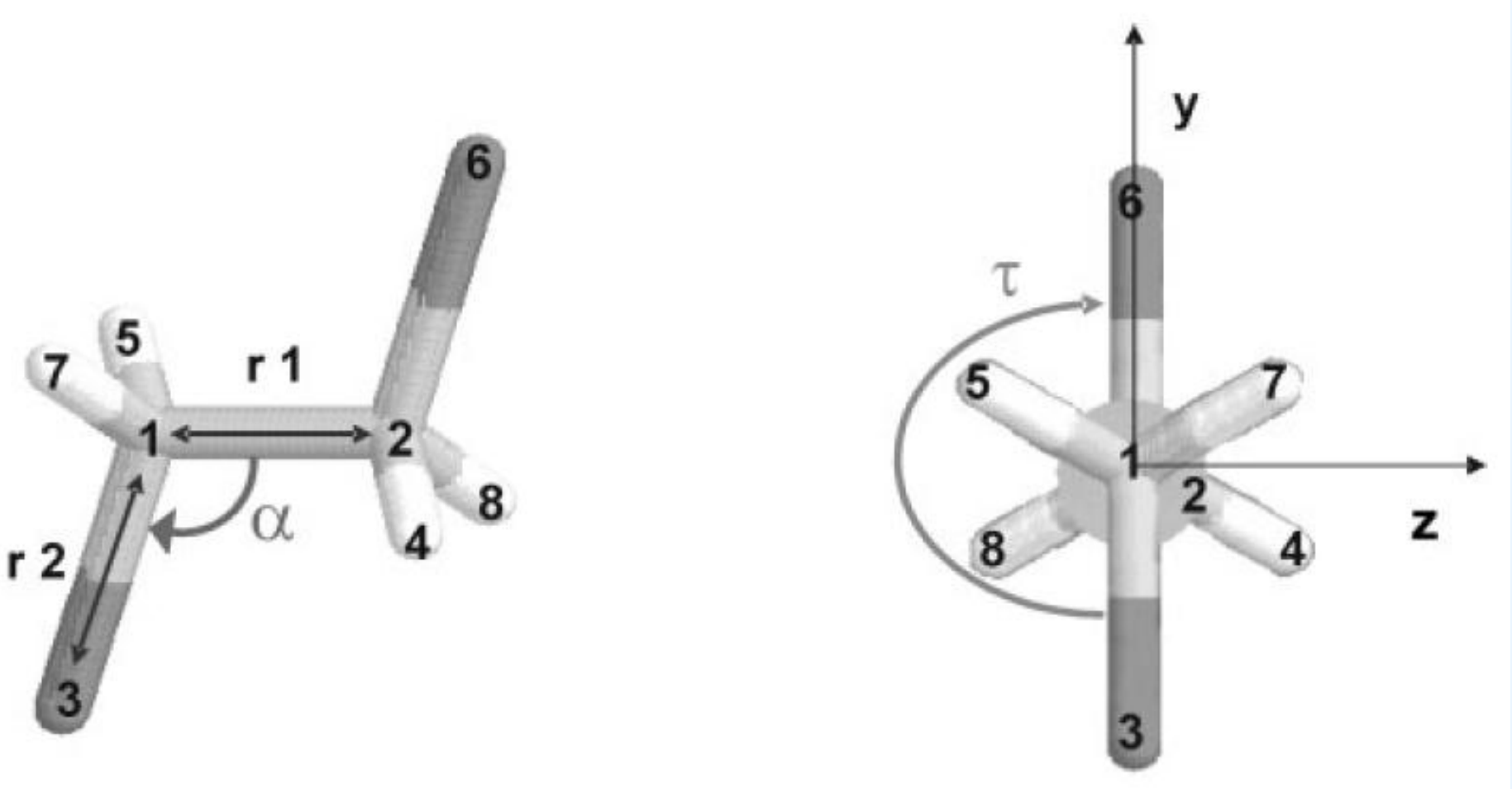


Длина связи

Угол между связями

Двугранный угол

1,2-дихлорэтан: внутренние координаты



Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

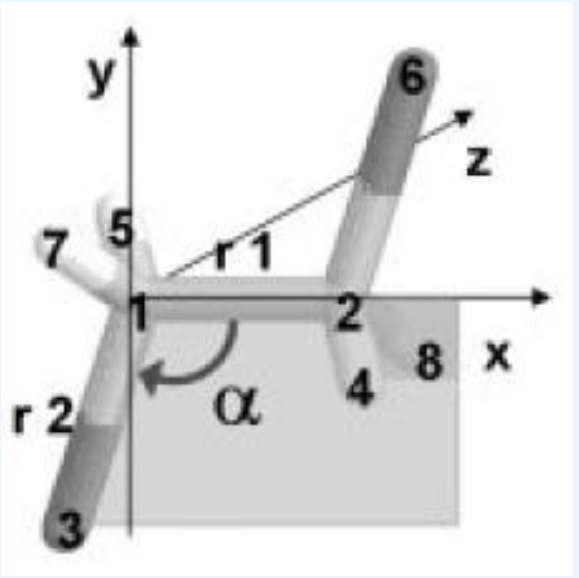
1,2-дихлорэтан: Z-матрица

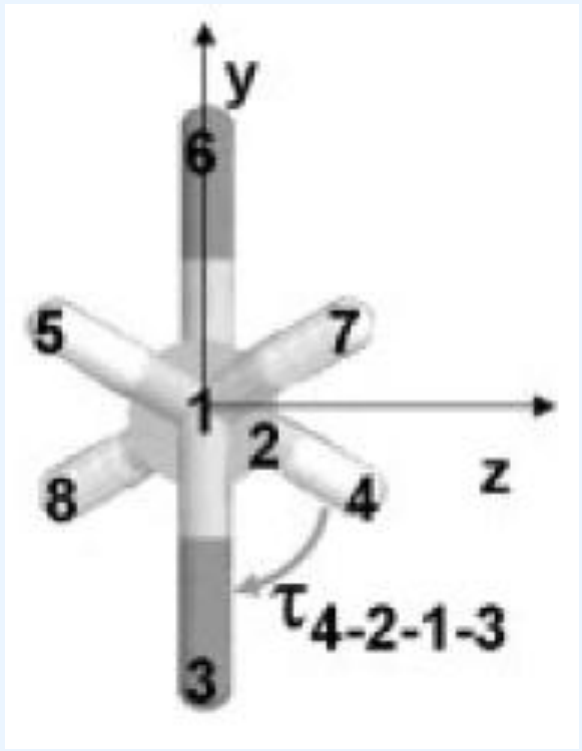
длина связи

угол между связями

двугранный угол

```
C1
C2    1.5   1
C13   1.7   1   109   2
H4    1.1   2   109   1   -60   3
H5    1.1   1   109   2   180   4
C16   1.7   2   109   1    60   5
```





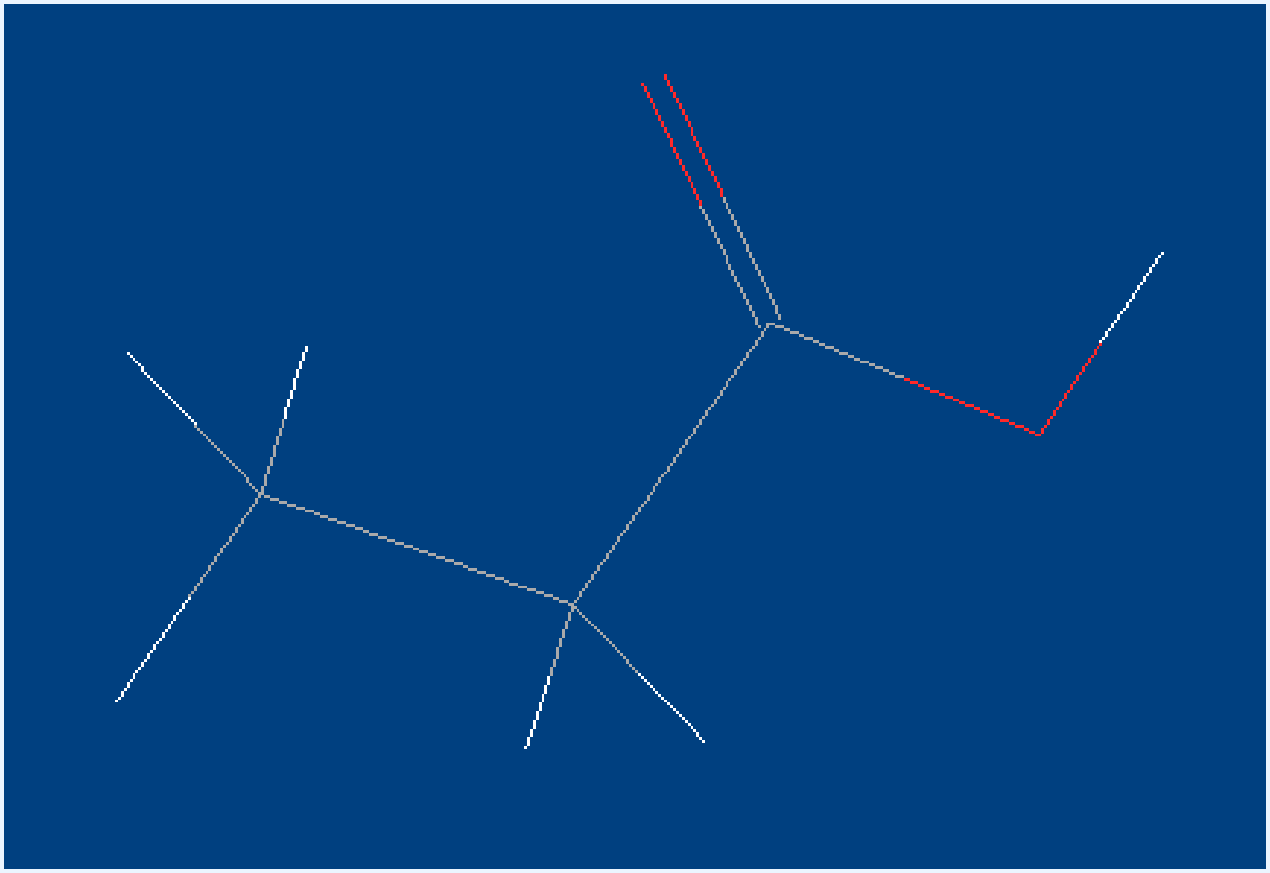
Chemoinformatics: A Textbook. Ed. J.Gasteiger, T.Engel. 2003

Трехмерные структуры
3D-структуры

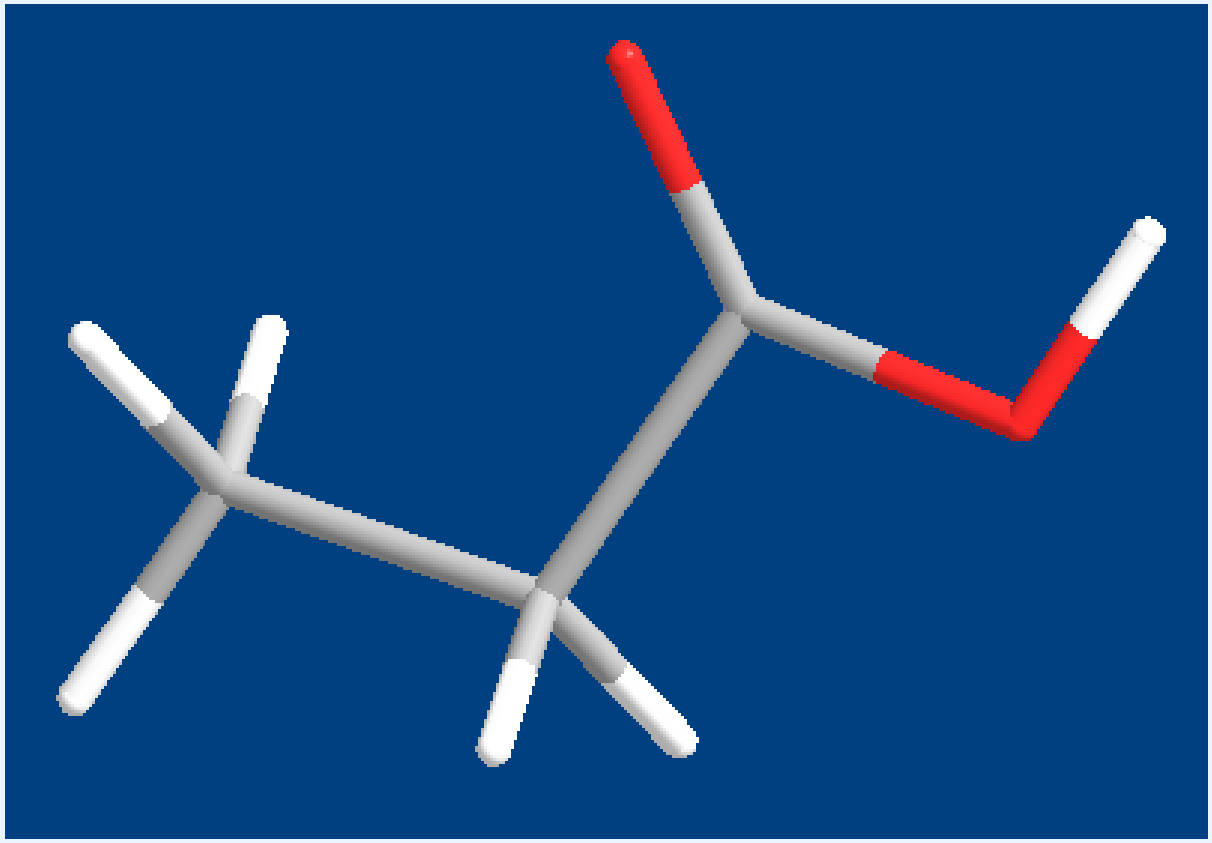
Модели визуального отображения структуры

Проволочная модель (Wire Frame)

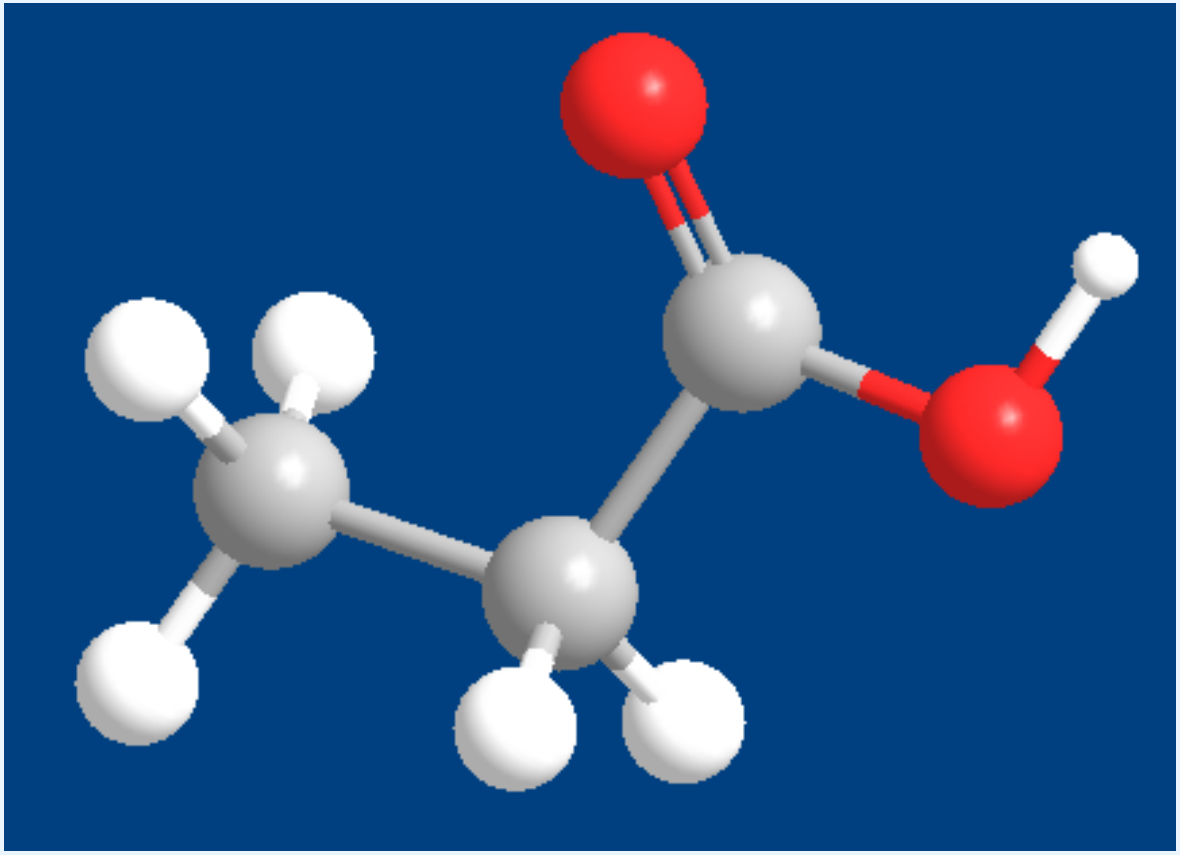
CH_3CH_2COOH



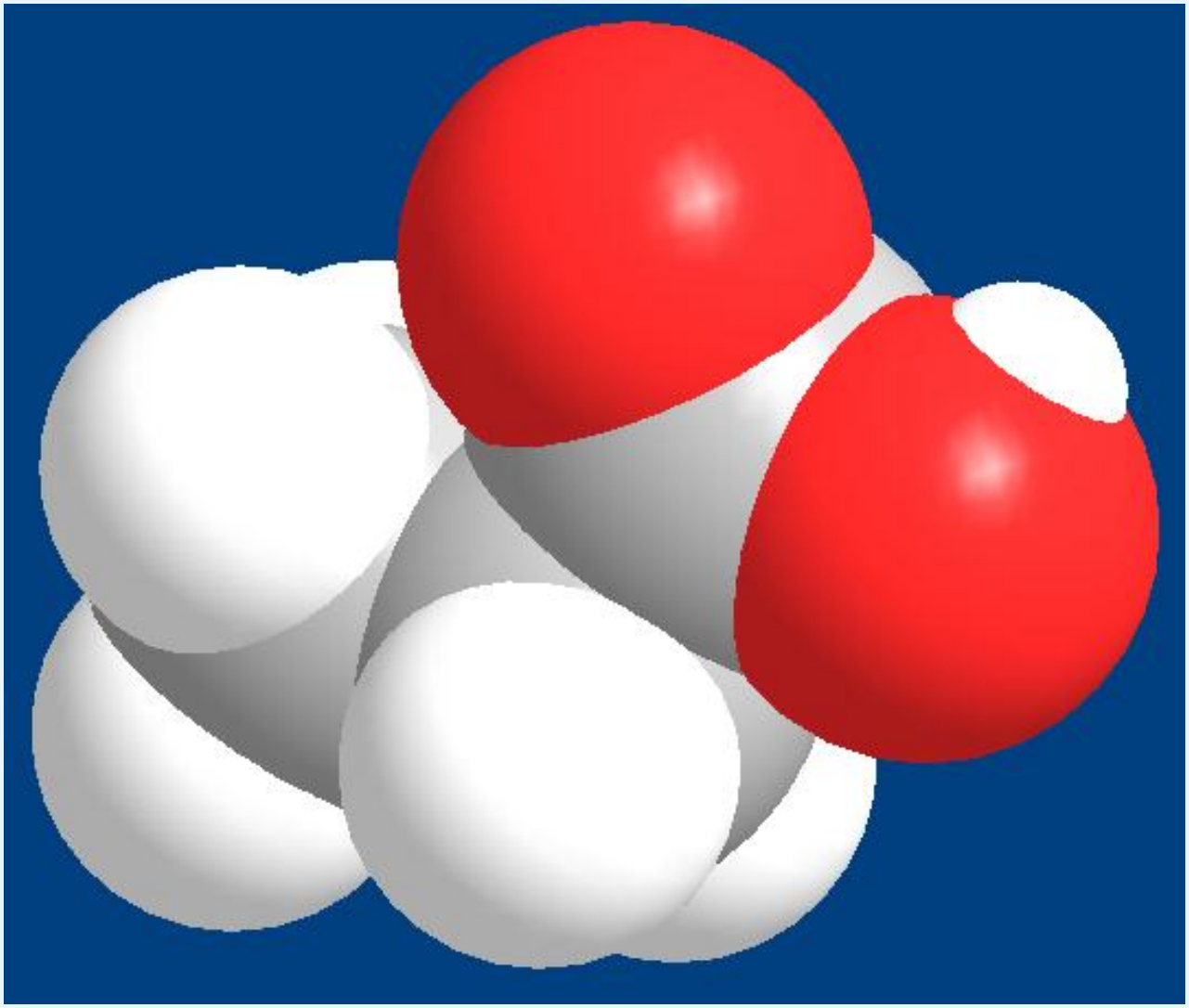
Стержневая модель (Sticks)



Шаростержневая модель (Balls and Sticks)

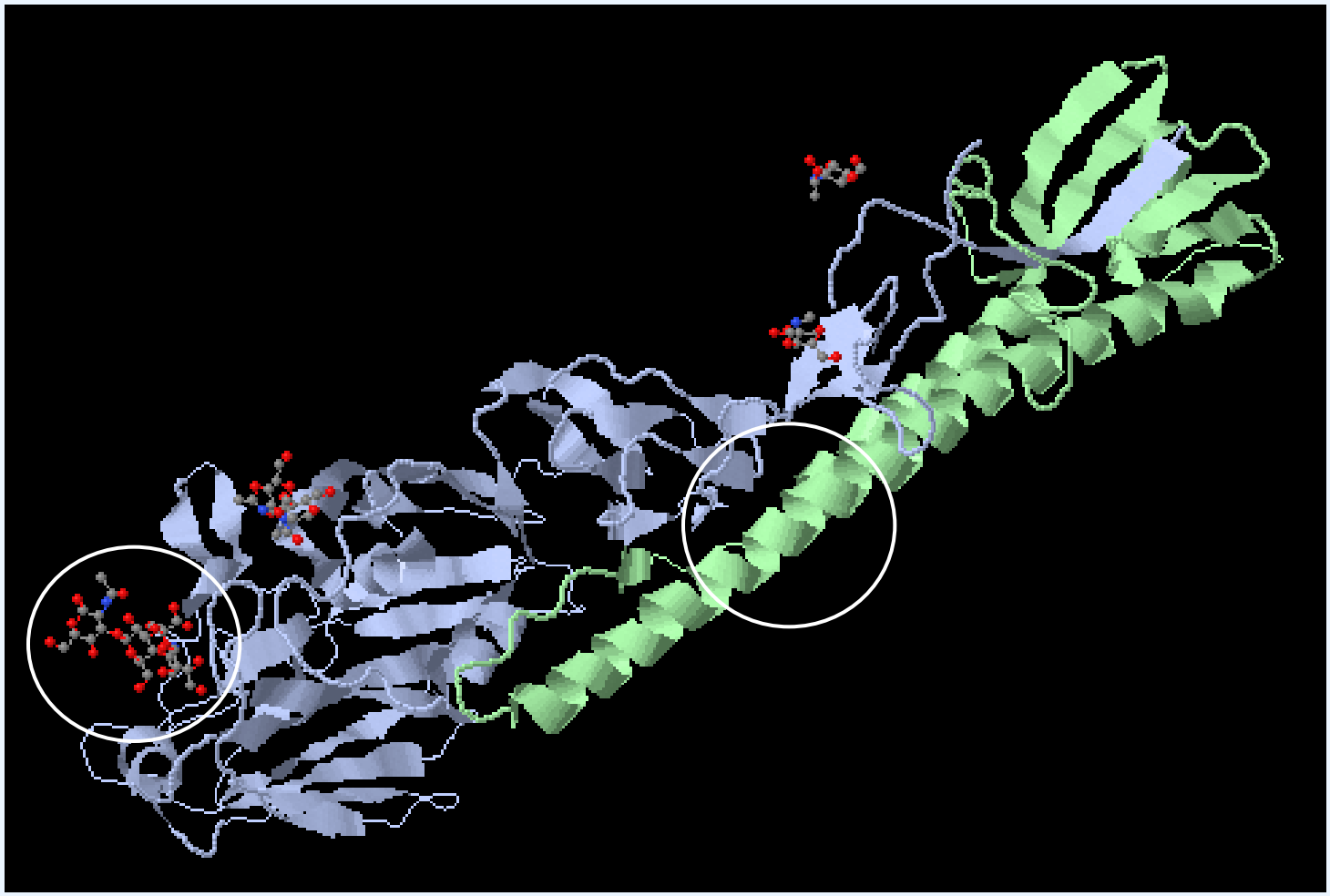


Объемная модель (Space Filling)
Ван-дер-Ваальсова поверхность



CPK-
модель

Ленточная - биомакромолекулы

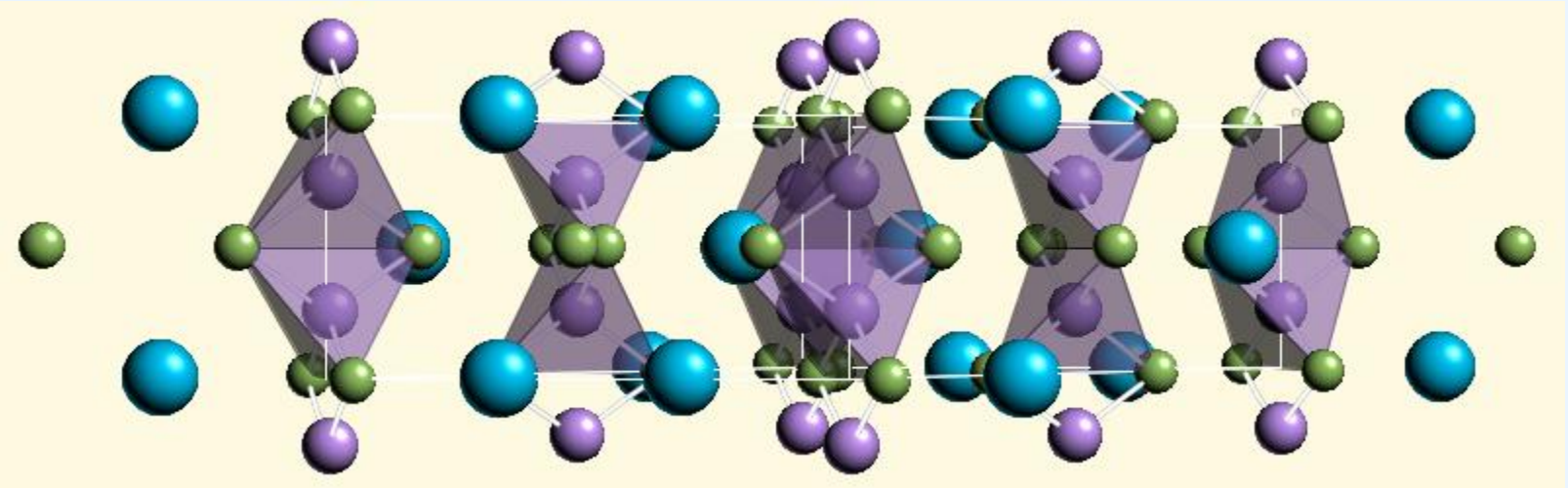


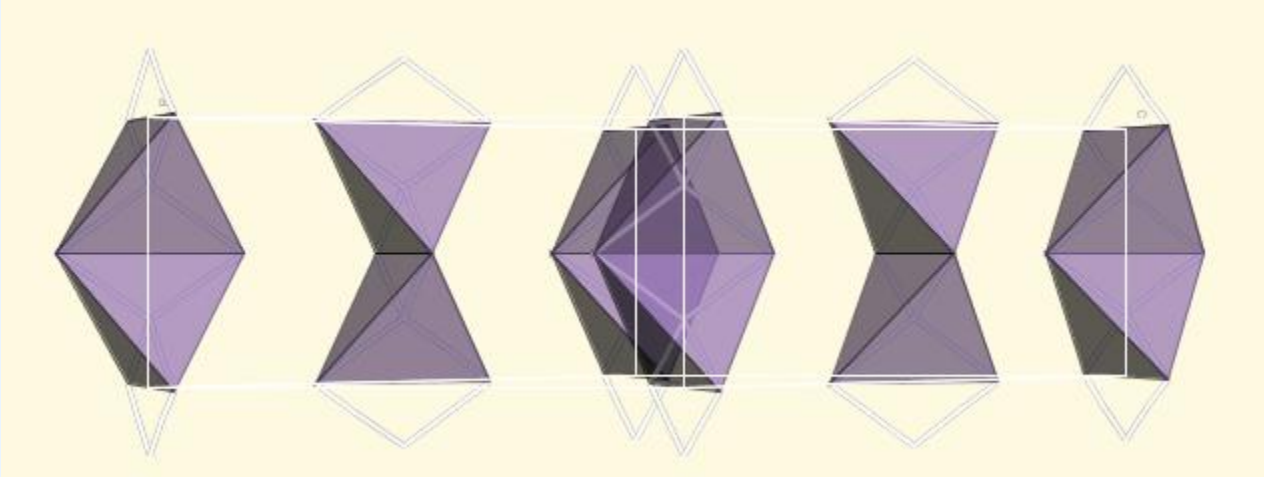
Ribbon

вирус птичьего гриппа

www.pdb.org

Многогранники

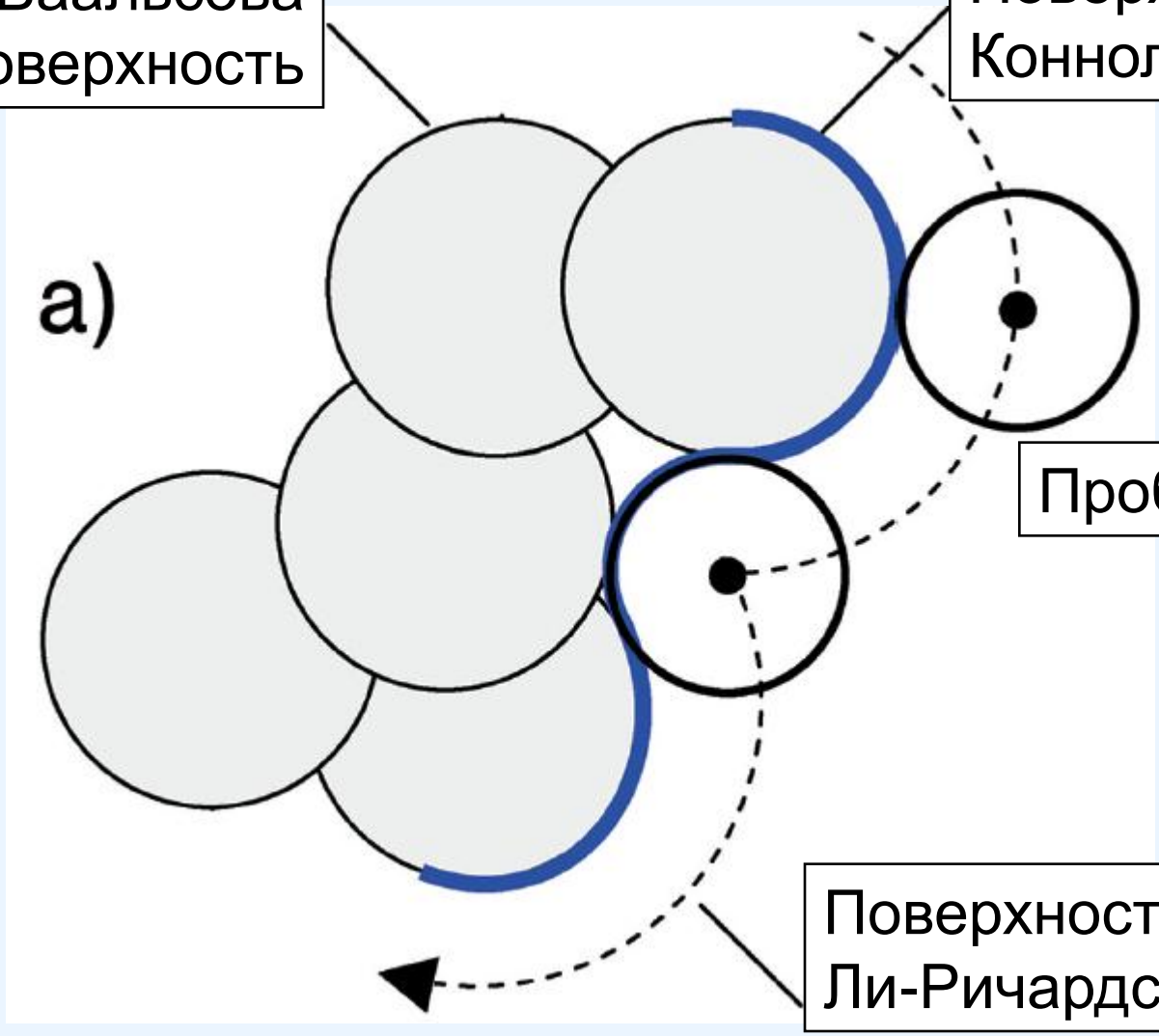




$TlFeS_2$

webmineral.com

Молекулярная поверхность



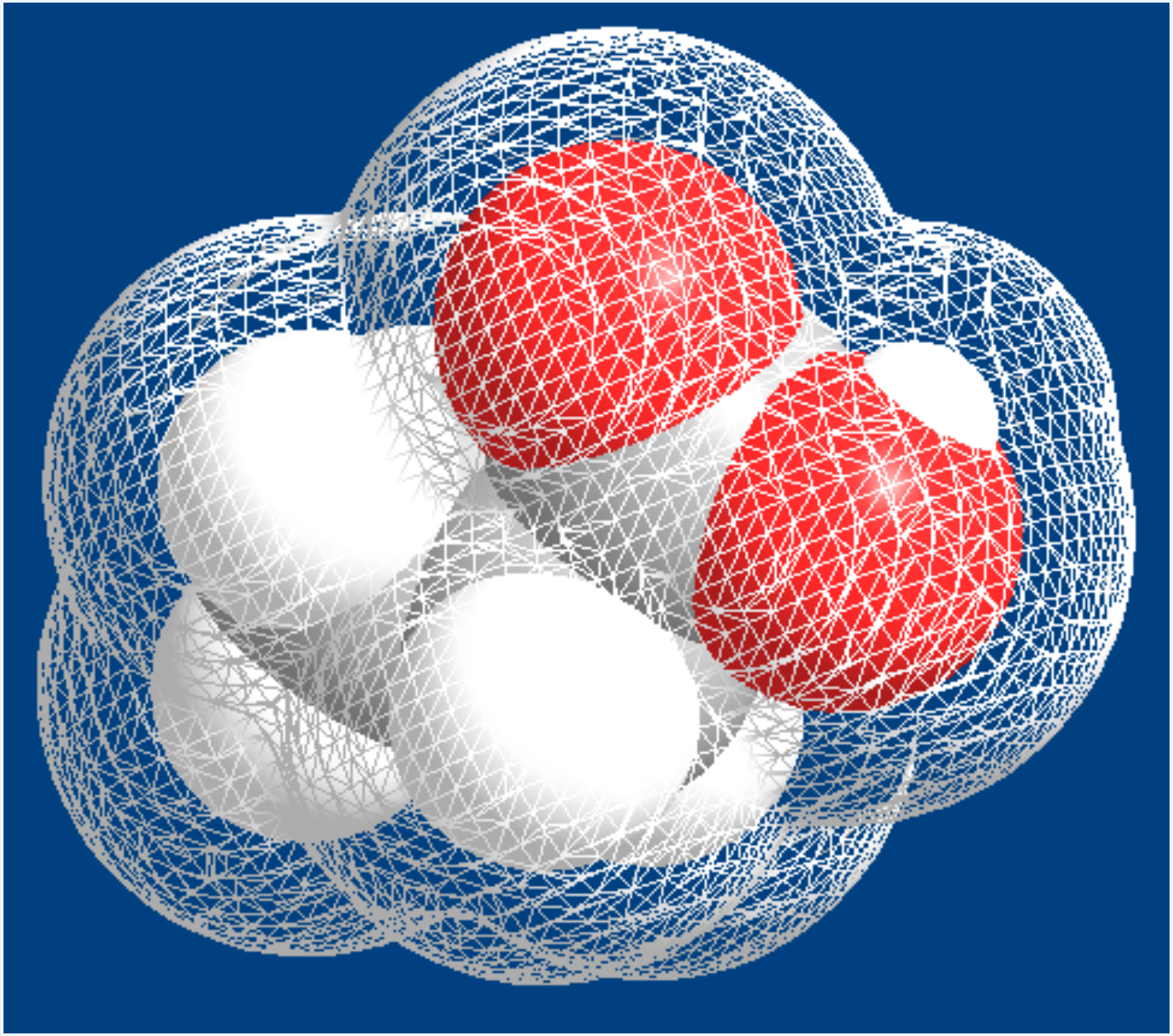
Поверхность, доступная растворителю
SAS – Solvent-accesible surface

Поверхность **Ли-Ричардса** –
доступная растворителю
(в некоторых источниках поверхности Ван-дер-Ваальса, Конноли тоже считают доступными растворителю).

Размер пробной сферы имеет значение.
r(H_2O) = 1,4 A
r(C_6H_6) = 2,6 A

Поверхность Конноли ограничивает
объем, недоступный растворителю.

Поверхность Ли-Ричардса

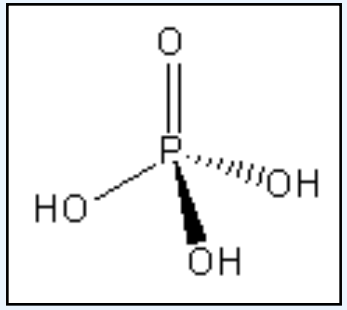


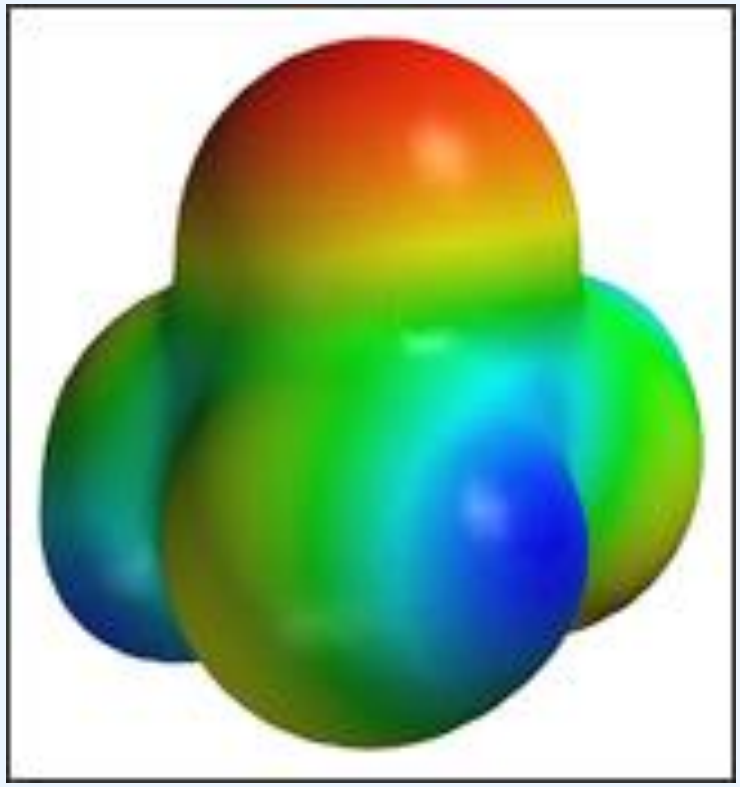
Solvent Accessible

Отображение электростатического потенциала

Цветовая гамма условна.
Цветовая гамма не стандартизирована.

Обычно:
синий цвет – положительный заряд,
красный – отрицательный.





http://chemwiki.ucdavis.edu/Theoretical_Chemistry/Chemical_Bonding/Electrostatic_Potential_maps