Структурно представяне

Таблица на свързаност

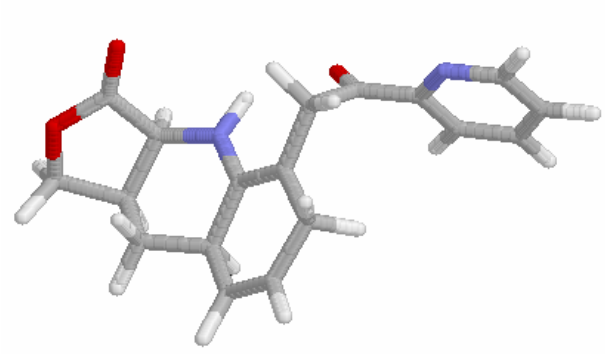
Линейни нотации

Файлови формати

Молекулни дескриптори

Топологични индекси

3D дескриптори

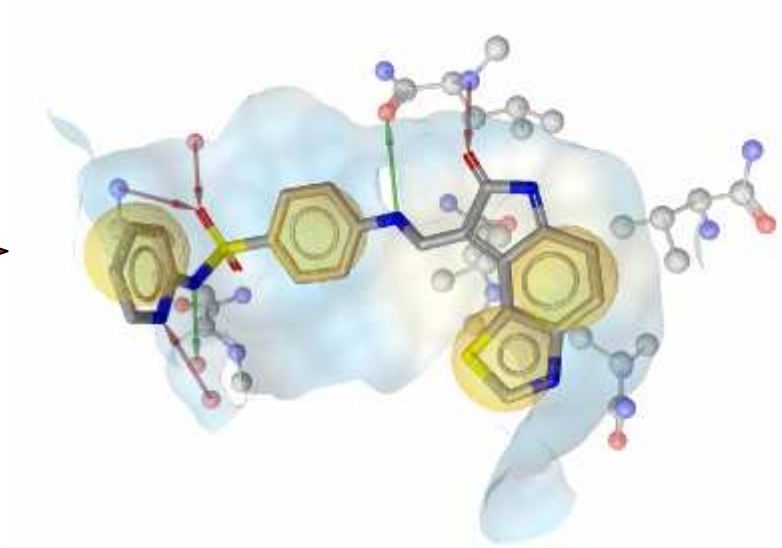


Химична информатика

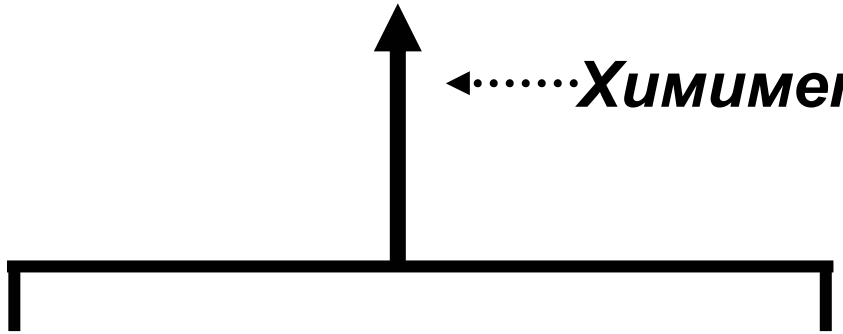
Химичната информатика е комбинирането на тези информационни ресурси, чрез които данните се трансформират в информация и информацията в знание с основна цел да се правят по-добри и по-бързи решения при идентификацията и оптимизацията на водещи структури (кандидати за лекарства)

?

F.K. Brown 1998



Химичната информатика е дисциплина, която прилага методите на информатиката при решаване на химични проблеми



Химиметрия

Математика,
Статистика,
Логика,

Теория на информацията

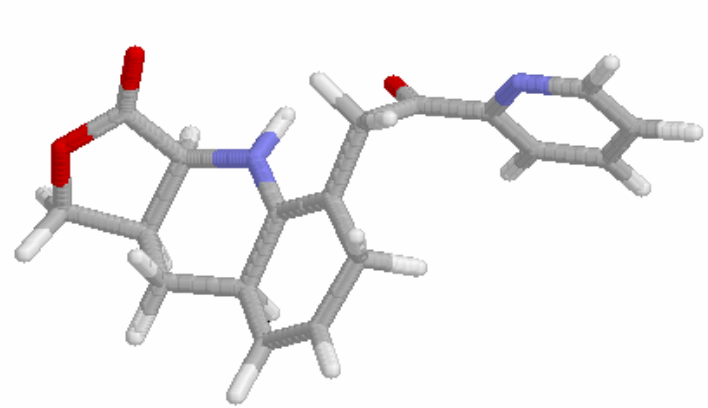
Информатиката

Химията изследва съединенията, техните свойства и трансформации

(1) е научна дисциплина, която се занимава със събиране, класифици-ране, съхранение, обработка и анализ на информация;

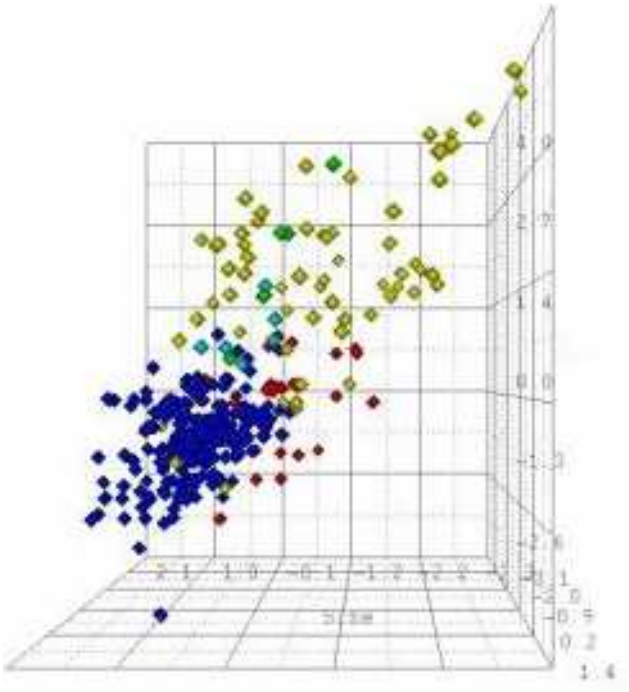
(2) се занимава с бази данни, онтология и създаване на софтуер

Химичната информатика винаги е свързана със структурите на химичните съединения.



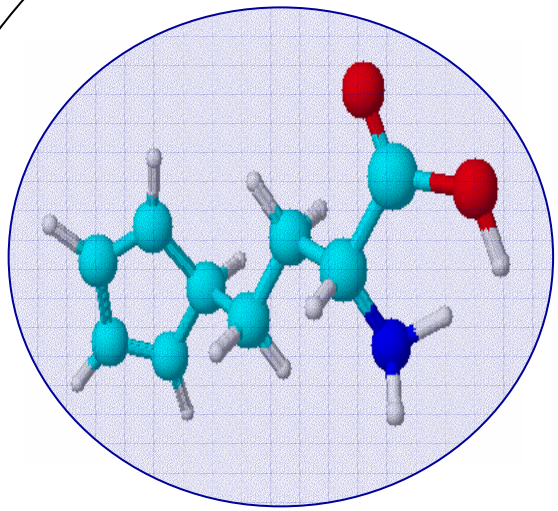
Химична информатика

Химичното пространство
съдържа поне 10^{60} молекули



Представяне на химично съединение

Структура



Свойства

-химични

-физични

-биологични

дескриптори

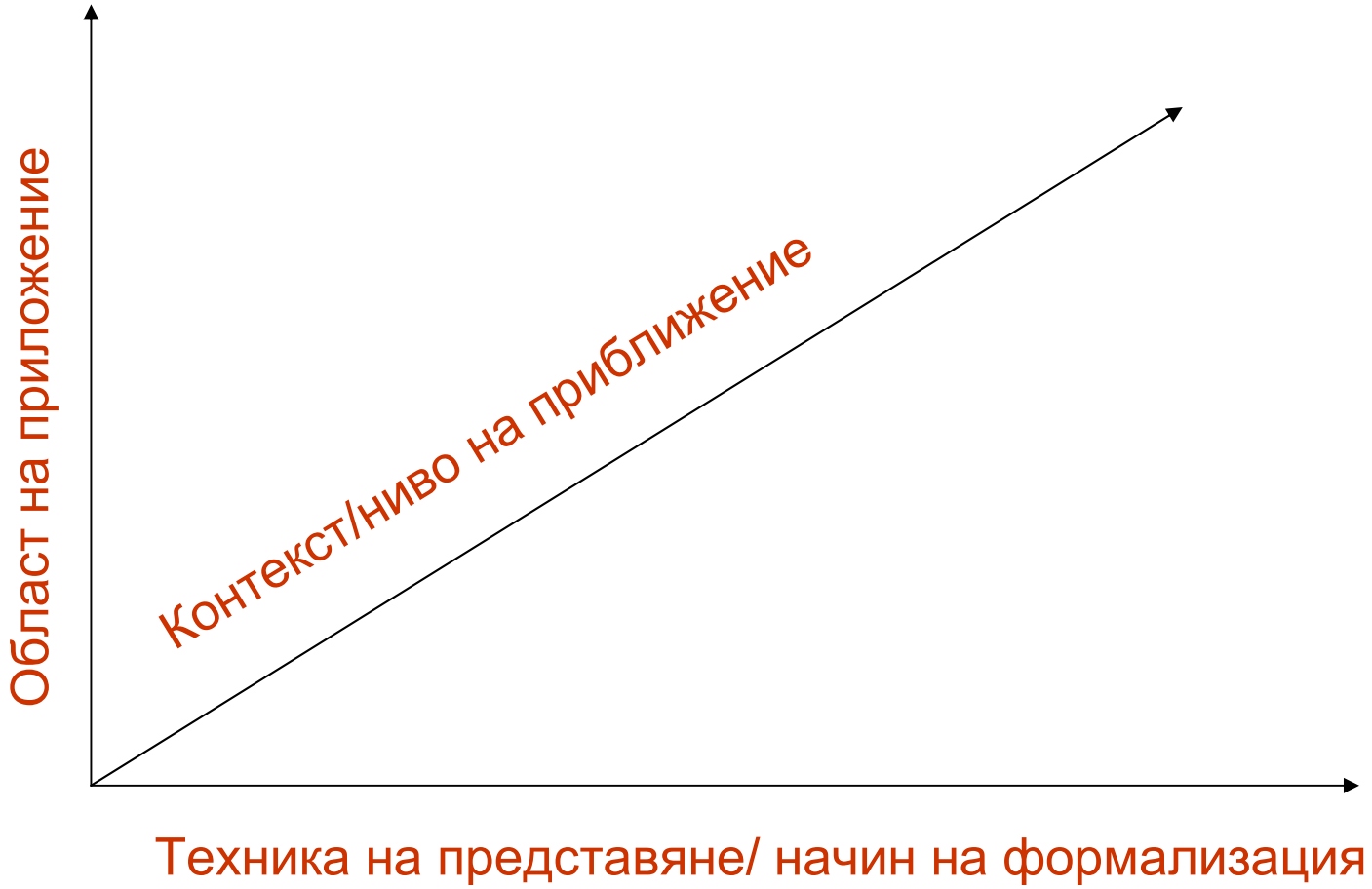
$(d_1, d_2, \ldots, d_n)$

Принцип в химичната информатика

Адекватността на представянето на химичните обекти е основна предпоставка за **ефективно моделиране**

Всеки тип представяне е ефективен за определен вид приложения

Характеристики/измерения на дадено представяне

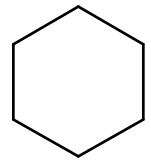


Нива на структурно представяне

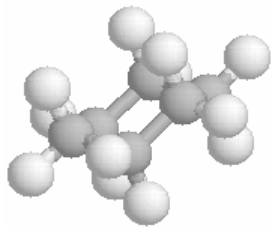
0D конституционни дескриптори,
1D брутна формула, номенклатури

cyclohexane
C6H12

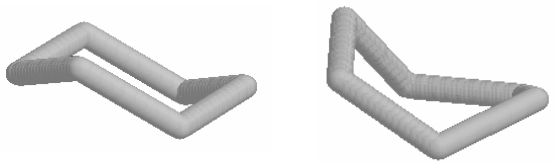
2D таблица на свързаност, топологични представяния, 2D координати, топологични дескриптори



3D 3D структура, молекулни повърхности, дескриптори базирани на 3D информация



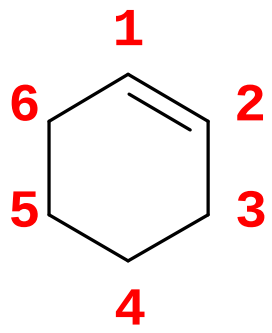
4D конформации



електронни дескриптори

ТС – първа основна форма

Състои се от два списъка: списък на атомите (ATOMLIST) и списък на връзките (BONDLIST)

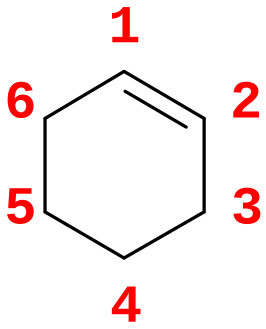


ATOMLIST		H
1	C	1
2	C	1
3	C	2
4	C	2
5	C	2
6	C	2

BONDLIST		
1	2	2
2	3	1
3	4	1
4	5	1
5	6	1
6	1	1

ТС – втора основна форма

Нарича се таблица на свързаност с излишъци (Redundant Connection Table). Състои се от един комбиниран списък, който описва атомите и техните първи топологични околности



RCT

1	C	2,2	6,1
2	C	1,2	3,1
3	C	2,1	4,1
4	C	3,1	5,1
5	C	4,1	6,1
6	C	1,1	5,1

Линейни нотации

Линейните нотации представят химичните структури като линейна последователност от букви и числа (*символен низ*).

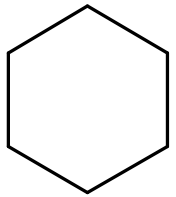
(1) Съкратено представяне

(2) Компютърно третиране

(3) Лесно генериране на допълнителна информация за структурата

Линейните нотации описват топологията на химичните структури.

Различни линейни нотации за циклохексана



ROSDAL	**1-2-3-4-5-6-1**
SMILES	**C1CCCCC1**
SLN	**C[1]H2CH2CH2CH2CH2CH2@1**
InChI	**1/C6H12/c1-2-4-6-5-3-1/h1-6H2**

Линейна нотация SMILES

C-C(N)=C …

Линейна нотация SMILES

Simplified **M**olecular **I**nput **L**ine **E**ntry **S**ystem

Система за опростено линейно въвеждане на молекули

Започната от Дейвид Уейнингер през 1986 / USEPA

Дизайнът е завършен в колежа Помона (Pomona Colledge)

Реализация на SMILES е направена в системата DayLight CIS

Литература и документация за SMILES:

(1) "SMILES 1. Introduction and Encoding Rules", Weininger, D., *J. Chem. Inf. Comput. Sci.*, **1988**, 28, 31.

(2) http://daylight.com

SMILES – основни концепции

Линейната нотация SMILES представя даден валентен модел на молекулите. *SMILES не е ефективна за химични обекти, които не се описват добре с валентния модел.*

SMILES не определя (не диктува) кой валентен модел се използва при представянето на структурите. *Валентният модел е начин за описание на местоположението на протоните, неутроните и електроните.*

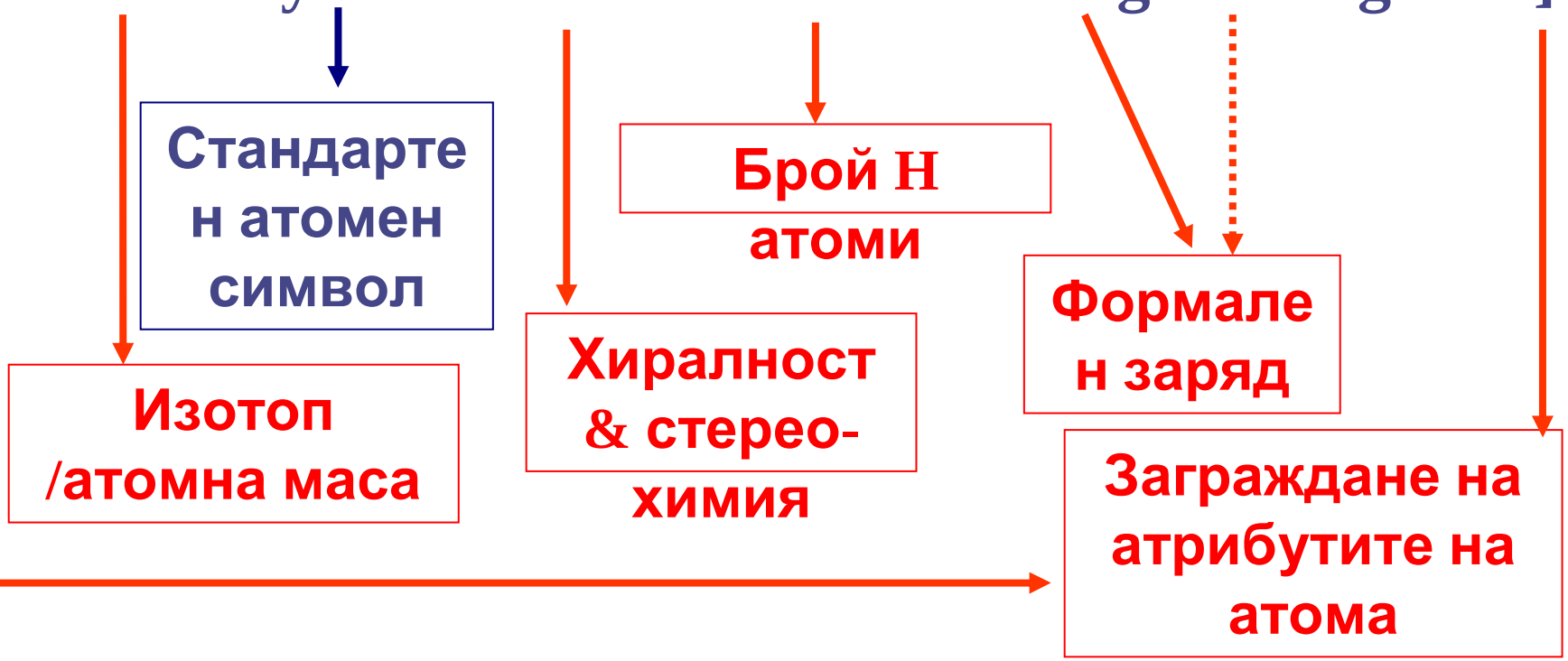
Пример: CN(=O)=O C[N+](=O)[O-]

SMILES не е дефинирана в рамките на определен софтуер. *SMILES представлява химичен модел на молекули, а не компютърен модел на структурни данни.*

SMILES е универсална номенклатура / индустриален и научен стандарт.

SMILES - спецификация на атоми

'[' <mass> symbol <chiral> <hcount> <sign<charge>> ']'



Атомният символ е задължителен. Ако е двубуквен втората буква трябва да е малка.

Например: **Br** а не **BR**

SMILES - спецификация на атоми

Атрибутите, които не са указани се асоциират със стойности по подразбиране.

Маса - не указана
Заряд – 0
Брой H атоми – 0

Хиралност - не указана

---

sp2 хибридизирани атоми може да се отбележат с малки букви

* - указва атом с неспецифиран тип. *Асоциира се с атомен номер 0.*

Основните “органични елементи” може да се указват без скоби [], ако нямат зададени атрибути и броят на водородните атоми съответства на най-ниската нормална валентност.

B(3), C(4), N(3,5), O(2), P(3,5), S(2,4,6), F(1), Cl(1), Br(1), I(1)

Елементи, които могат да се употребяват без скоби [ ]

1a	2a	3b	4b	5b	6b	7b	8	8	8	1b	2b	3a	4a	5a	6a	7a	0
1 **H**																	2 **He**
3 **Li**	4 **Be**											5 **B**	6 **C**	7 **N**	8 **O**	9 **F**	10 **Ne**
11 **Na**	12 **Mg**											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 **Ar**
19 **K**	20 **Ca**	21 **Sc**	22 **Ti**	23 **V**	24 **Cr**	25 **Mn**	26 **Fe**	27 **Co**	28 **Ni**	29 **Cu**	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 **Kr**
37 **Rb**	38 **Sr**	39 **Y**	40 **Zr**	41 **Nb**	42 **Mo**	43 **Tc**	44 **Ru**	45 **Rh**	46 **Pd**	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 **Xe**
55 **Cs**	56 **Ba**	57 **La**	72 **Hf**	73 **Ta**	74 **W**	75 **Re**	76 **Os**	77 **Ir**	78 **Pt**	79 **Au**	80 **Hg**	81 **Tl**	82 **Pb**	83 **Bi**	84 **Po**	85 **At**	86 **Rn**
87 **Fr**	88 **Ra**	89 **Ac**	104 **Rf**	105 **Ha**													
		58 **Ce**	59 **Pr**	60 **Nd**	61 **Pm**	62 **Sm**	63 **Eu**	64 **Gd**	65 **Tb**	66 **Dy**	67 **Ho**	68 **Er**	69 **Tm**	70 **Yb**	71 **Lu**		
		90 **Th**	91 **Pa**	92 **U**	93 **Np**	94 **Pu**	95 **Am**	96 **Cm**	97 **Bk**	98 **Cf**	99 **Es**	100 **Fm**	101 **Md**	102 **No**	103 **Lr**		

Примери за SMILES атоми

2DP	SMILES	Име	Обяснения
S	[S]	sulfur	По подразбиране: маса 0, заряд 0
Au	[Au]	Gold	По подразбиране: брой H атоми 0
CH_4	C	methane	Нормалната валентност на C е 4
PH_3	P	phosphine	Най-ниската нормална валентност на P е 3
H_2S	S	hydrogen sulfide	Най-ниската нормална валентност на S е 2
HCl	Cl	hydrocloric acid	Най-ниската нормална валентност на Cl е 1
OH^-	[OH-]	hydroxide anion	Заряд -1; “-” е еквивалентно на -1 т.е [OH-1]
Fe^{+2}	[Fe++]	Iron(II) cation	Заряд +2; “++” е еквивалентно на +2 т.е [Fe+2]
^{235}U	[235U]	Uranium-235	Водещото число е атомната маса
***+2**	[*+2]	Не е молекула	Атом от неизвестен тип със заряд +2

SMILES - спецификация на връзки

Единична връзка	<нищо> или -
Двойна връзка	=
Тройна връзка	#
Ароматна връзка	:

Единична насочена “нагоре”	/
Единична насочена “надолу”	\

Затварянето на цилъл е алтернативен метод за указване на връзки.

Примери за SMILES връзки

2D	SMILES	Име	Обяснения
	CC C-C [CH3]- [CH3]	ethane	По подразбиране два съседни атома се приема че са свързани с единична връзка. Указването на единична връзка не е необходимо
	C=O O=C	formaldehyde	Редът на атомите е без значение
	C#N N#C	hydrogen cyanide	

Спецификация на SMILES разклонения

Атом (Разклонение 1) Разклонение 2
Атом (Разклонение 1) (Разклонение 2) Разклонение 3
Атом (Разклонение 1) (Разклонение 2) (Разклонение 3) Разклонение 4

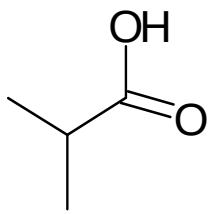
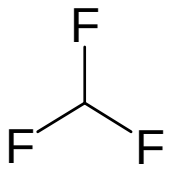
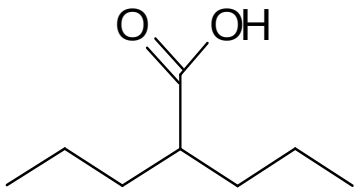
Разклоненията към даден атом се поставят в скоби

Всяко разклонение се постява в отделна скоба с изключение на последното. *Последното разклонение се смята за продължение на веригата*

Разклоненията могат да се влагат едно в друго

Разклонението в скобата може да започне с тип връзка, ако връзката към разклонението не е единична

Примери за SMILES разклонения

2D	SMILES	Име	Обяснения
	CC(C)C(=O)O или CC(C)C(O)=O	isobutyric acid	Символ за връзка може да се употреби вътре в скобите като начало на разклонението.
	FC(F)F или C(F)(F)F	fluoroform	Атомите са ранопоставени при избиране на начален атом.
	CCCC(C(=O)O)CCC	4-heptanoic acid	Разклоненията могат да се влагат едно в друго

Спецификация на циклични съединения

SMILES цикъл се затваря посредством две еднакви числа (индекси) указани към атомите, които затварят цикъла.

<Атом 1><Число> ....<Атом2><Число>
<Атом 1><Число> ....<Атом2>=<Число>
<Атом 1><Число> ....<Атом2>:<Число>

Числата към атомите означават, че между Атом 1 и Атом 2 има връзка.

Пред числата може да се сложи символ за връзка, когато затварянето не е с единична връзка. *Символът за връзка обикновено се слага само на втория атом.*

Ако числото е двуцифрено, пред него трябва да се сложи %.

Например: C%12

Спецификация на циклични съединения

Възможно е едно число да се употреби повтарно, след като е било употребено за затваряне на връзка между два атома. Езикът SMILES позволява със същото число след това да се затвори връзка между други два атома (друг цикъл) .

НЕ СЕ ПРЕПОРЪЧВА ПОВТОРНАТА УПОТРЕБА НА ЧИСЛА ЗА ЗАТВЯРЯНЕ НА ЦИКЛИ

Възможно е в даден атом едновременно да се затварят два цикъла.
Атомът съответно ще има два индекса

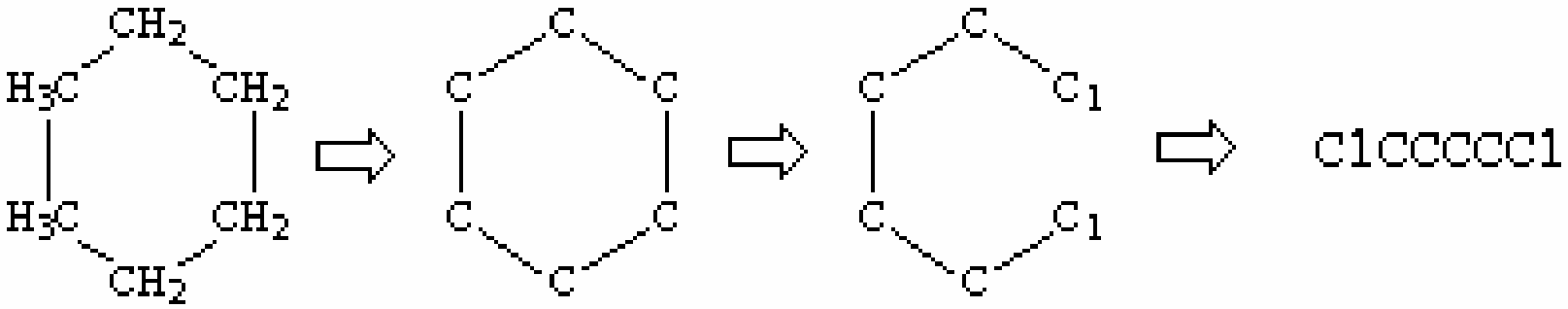
Например: C12

Да се прави разлика с C%12, което означава, че атомът е с един двуцифрен индекс а не два индекса 1 и 2

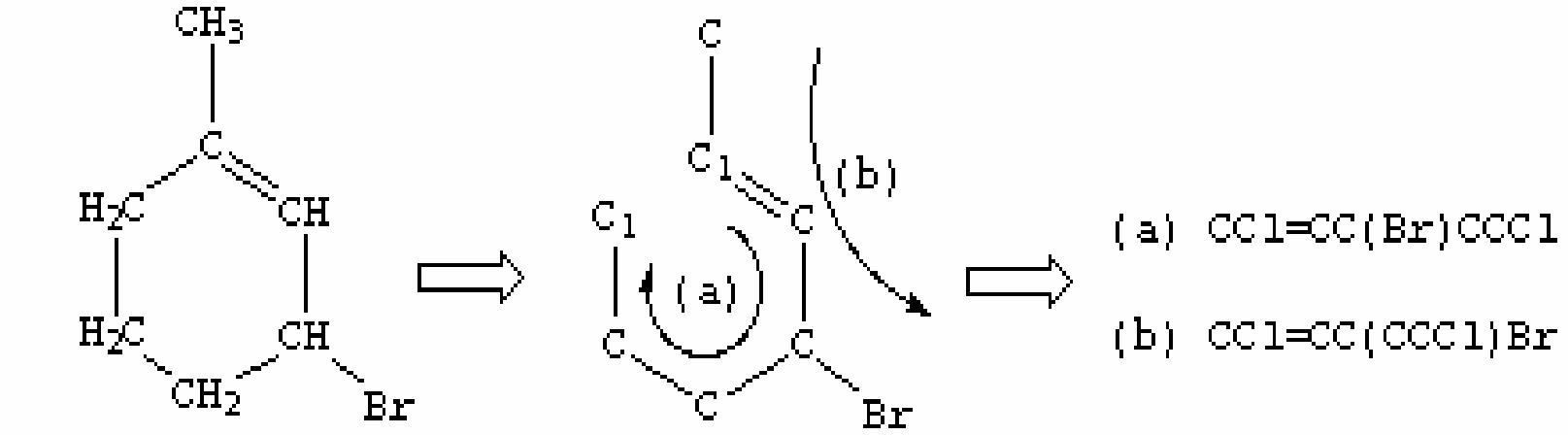
Практическо правило за определяне на SMILES на циклично съединение

Някои връзки в съединението се премахват така, че да се получи **ациклично** съединение със **свързана** структура.

Липсващите връзки в ацикличното съединение ЗАДЪЛЖИТЕЛНО се добавят посредством индекси (числа).



Обикновено има няколко различни, но еднакво валидни начина за да се опише една структура.



Без значение е в кое разклонение ще се поставят атомите, които затварят цикъл посредством индекси.

Граф-теоретични факти относно цикличните структури

Всеки един граф може от цикличен да стане ацикличен, като се премахнат определен брой връзки.

Минималният брой връзки, които трябва да се премахнат за да се получни ацикличен граф се нарича цикломатично число – **μ.**

Теорема:

За всеки един граф цикломатичното число **μ** е по-малко или равно на броя на циклите в графа.

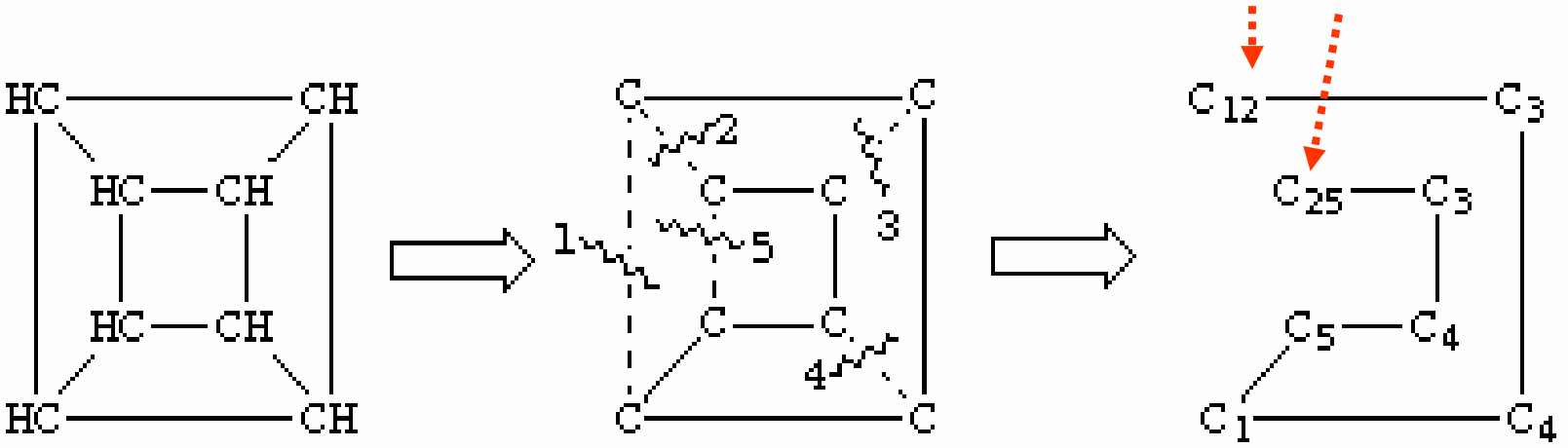
Следствие:

Ако от всеки цикъл се премахне по една връзка, със сигурност ще се получи ациклична структура.

Не винаги това е задължително

Пример със структурата на кубена

има атоми с два индекса



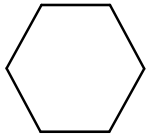
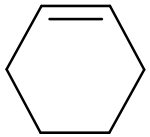
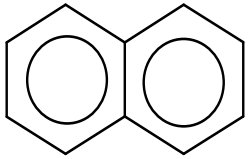
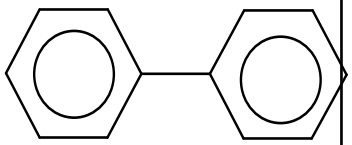
CCCCCCCC → **C12C3C4C1C5C4C3C25**

ацикличен SMILES *цикличен SMILES*

Броят на циклите е **6** (това са най-малките прости цикли – множеството им се означава със SSSR)

Цикломатичното число **μ** = 5

Примери: SMILES кодове на циклични съединения

2D	SMILES	Име	Обяснения
	C1CCCCC1 C(CC1)CCC1	cyclohexane	Ако не е указана, връзката, която се затваря цикъл е единична
	C1=CCCCC1 C=1CCCCC1 C1CCCCC=1 C=1CCCCC=1	Cyclohexene	Връзката, която затваря цикъла може да се укаже дори и при двата индекса стига да не си противоречат. C=1CCCCC-1 – не е добре
	c12c(cccc1)cccc2 или c1cc2ccccc2cc1	naphtalene	Атомите могат да имат повече от един индекс (затваряне на цикъл)
	c1ccccc1c2ccccc2 или c1ccccc1c1ccccc1	biphenyl	Възможно е индексите за затваряне на цикъл да се използват повторно. НЕ СЕ ПРЕПОРЪЧВА !

Прекъснати SMILES структури

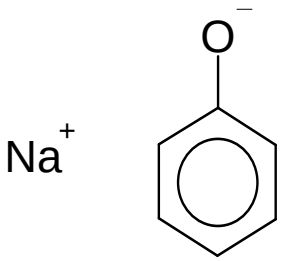
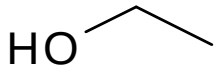
Прекъсване на структурата се задава посредством символа “.”(точка). *Според валентния модел точката съответства на връзка с формален порядък нула.*

В практиката, черз един SMILES стринг може да се представят няколко структури едновременно (АНСАМБЪЛ) като се разделят с точки.

Например: C.CC.CCC.CCCC.CCCCC

Прекъснатите SMILES кодове може да се използват за описание на йони, лиганди и други структурни фрагменти. *Редът, в които те се указват не е от значение.*

Примери: SMILES кодове с прекъсване

2D	SMILES	Име	Обяснения
	[Na+].[Cl-]	Sodim cloride	Точката типично означава прекъсване
	[Na+].[O-]c1ccccc1 или c1cc([O-].[Na+])ccc1	Sodium phenoxide	Точката означава “нулева връзка / без връзка”.
	C1.O2.C12 или CCO	Ethanol	Странно и изопачено използване на точки и индекси – HO е ВЯРНО. Не винаги наличието на точки означава прекъсната структура, тъй като връзки меже да се указват чрез индекси за затваряне на цикли

Спецификация на изотопи

Изотопи се указват посредством префикс към атомния символ. Като префикс се използва желаната атомна маса.

[<маса> Елементен символ …]

Атомната маса може да се укаже само вътре в квадратните скоби.

Примери: SMILES кодове – указване на изотопи

2D	SMILES	Име	Обяснения
CH_4	C	methane	Масата на въглерода не е указана
C	[C]	Elemenal carbon	Масата на въглерода не е указана
^{12}C	[12C]	Elemenal carbon -12	Указана е маса 12
^{13}C	[13C]	Elemenal carbon -13	Внимание: възможно е да се укаже маса, която няма смисъл
CH_4	[13CH4]	C-13 methane	В този случай свързаните въглеродни атоми трябва да се укажат в квадратните скоби.

Спецификация на конфигурация около двойна връзка

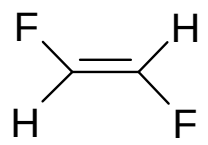
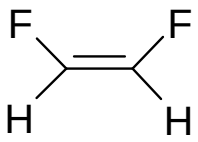
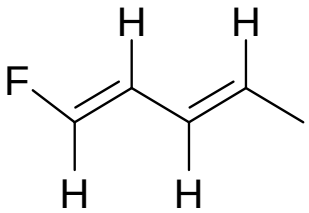
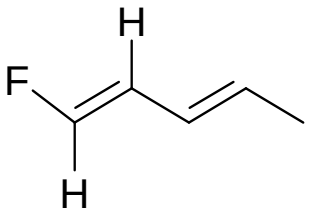
Конфигурацията около двойна връзка се указва посредством символите “/” и “\”.

Тези символи имат смъсъл само когато са указани едновременно при двата атома около двойна връзка

<Atom 1> / C = C /<Atom 2> – trans конфигурация
<Atom 1> \ C = C \ <Atom 2> – trans конфигурация

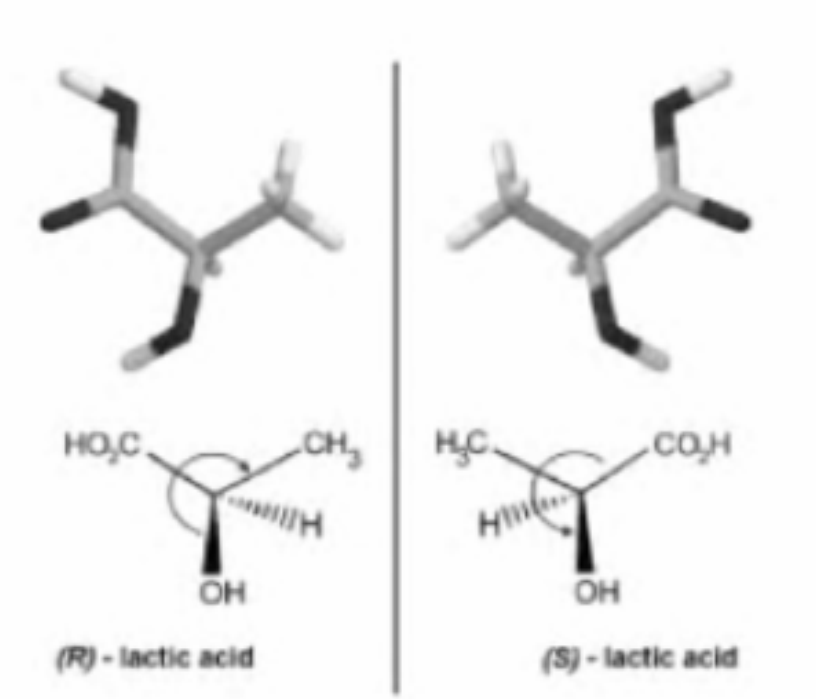
<Atom 1> / C = C \<Atom 2> – cis конфигурация
<Atom 1> \ C = C / <Atom 2> – cis конфигурация

Примери: SMILES конфигурация на двойна връзка

2D	SMILES	Име	Обяснения
	F/C=C/F или F\C=C\F	Trans-diflouroethene	‘F’атомите са разположени на ‘противоположни’ страни спрямо двойната връзка
	F/C=C\F или F\C=C/F	Cis-diflouroethene	‘F’атомите са разположени от ‘една и съща’ страна спрямо двойната връзка
	F/C=C/C=C/ C	trans, trans-1-fluoro-penta-1,3-diene	Ориентациите на двойните връзки са напълно специфицирани
	F/C=C/C=CC	trans, unspec-1-fluoro-penta-1,3-diene	Ориентациите на двойните връзки са частично специфицирани

Представяне на стерео информация

Директно 3D представяне чрез декартови координати



2.5D представяне (2D + хиралност)

Представяне на стерео информация

Абсолютно представяне – CIP номенклатура за тетраедрична хиралност - *правилото на Кан, Инголд, Прелог*

Заместителите (лигандите) на хиралния център се подреждат по старшинство L1 > L2 > L3 > L4

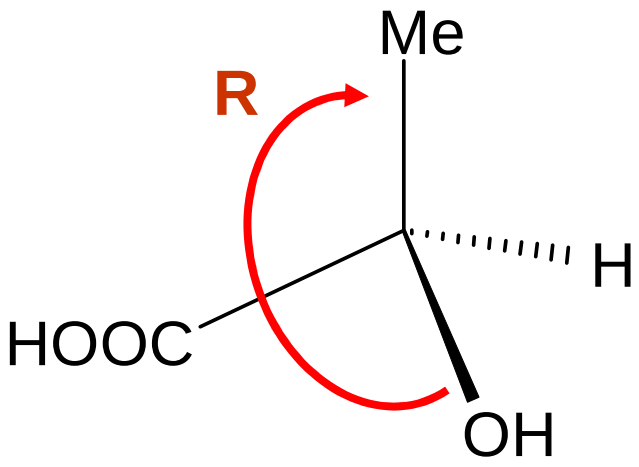
Най-младшия заместител се разполага най-далече от наблюдателя

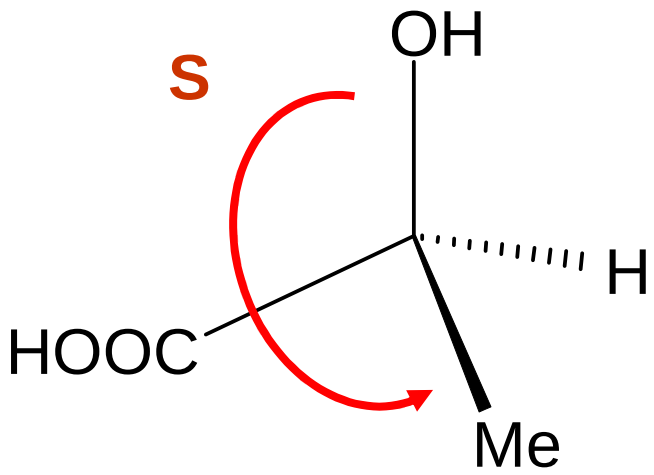
Ако L1, L2, L3 са подредени по часовниковата стрелка, то конфигурацията е R

Представяне на стерео информация

Подреждане на заместителите по старшинство

OH > CO2H > Me > H





Представяне на стерео информация

Относително представяне за тетраедрична хиралност

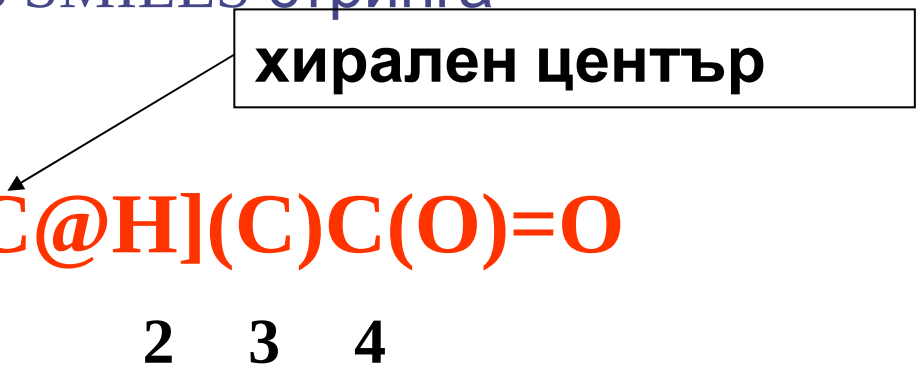
Заместителите на хиралния център се подреждат според вътрешно номериране, което не съответства на старшинството на лигандите L1, L2, L3, L4

Начина за описание на симетрията според вътрешната номерация е относителен и зависи от номерацията.

Не може директно да се направи съответствие между относително предстявяне и R/S

Представяне на стерео информация

Хиралността в SMILES се описва на базата на подредбата на съседите на хиралния център според реда, в който са в SMILES стринга



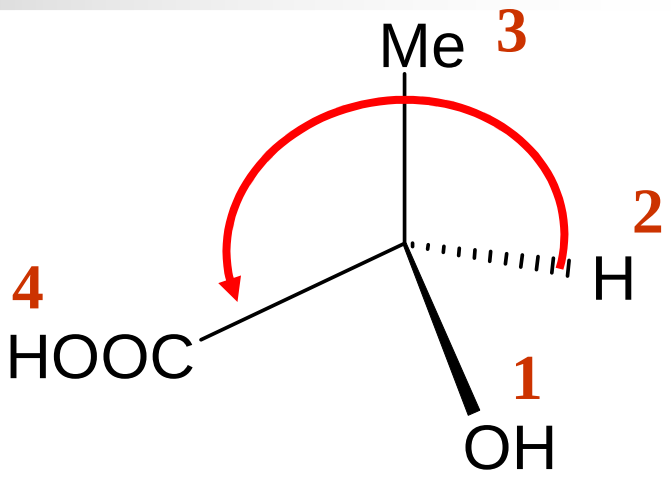
@ - означава, че ако гледаме от съсед 1 към хиралния център другите съседни атоми 2, 3 и 4 (H, C и C(O)=O) са подредени обратно на часовниковата стрелка

@@ - означава, че атоми 2, 3 и 4 са подредени по посока на часовниковата стрелка

Представяне на стерео информация

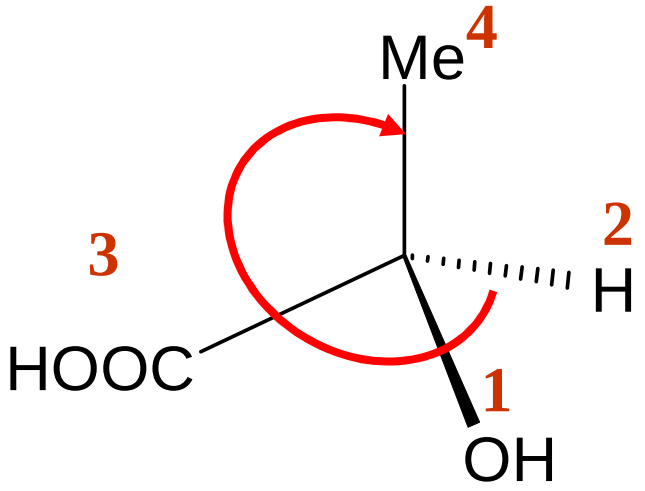
O[C@H](C)C(O)=O

1 2 3 4



O[C@@H](C(O)=O)C

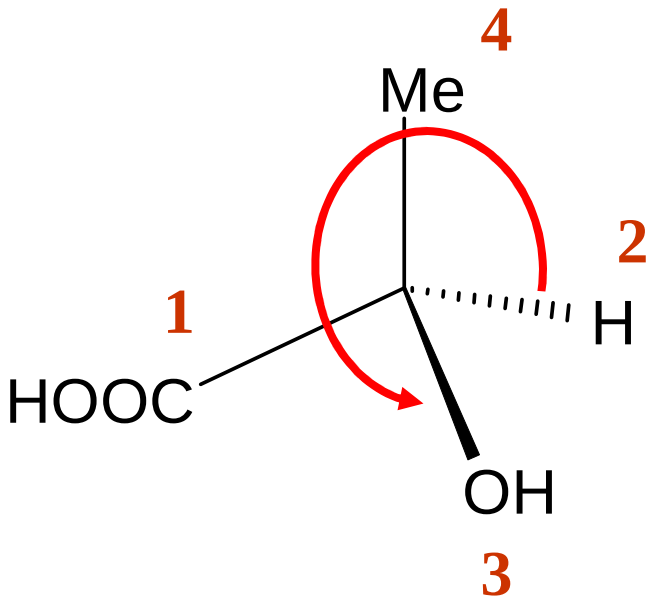
1 2 3 4

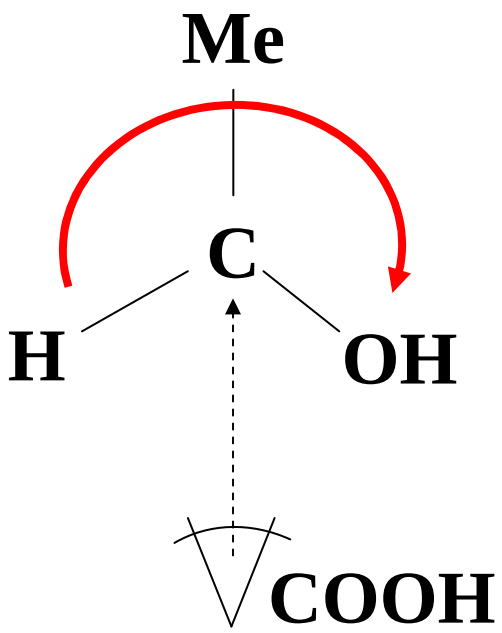


Представяне на стерео информация

C((O)=O)[C@@H](C)O

1 2 3 4





Спецификация на стерео-информация / хиралност

SMILES прилага обобщен метод за локално описание на хиралност

Частта <chiral> от описанието на атомните атрибути […<chiral> ...] се представя във вида:

@ <клас> <подредба>

Възможно е класът / подредбата да не се укажат (т.е. използа се само символът ‘@’) - тогава се работи със стойности по подразбиране. *Стоиностите по подразбиране се определят от хибридизацията или валентността на атома*

Хиралността в SMILES се описва на базата на подредбата на съседите на даден атом според реда, в който са в SMILES стринга

N **C** (O) (C) Cl
1 2 3 4

C (N) (O) (C) Cl
1 2 3 4

N [**C** H] (C) Cl
1 2 3 4

Задаване на тетраедрична хиралност

Задава се за атоми с 4 съседа. Атомът, при който е указан атрибутът <chiral> е хиралният център.
Това е най-популярния случай на задаване на локална стерео информация.

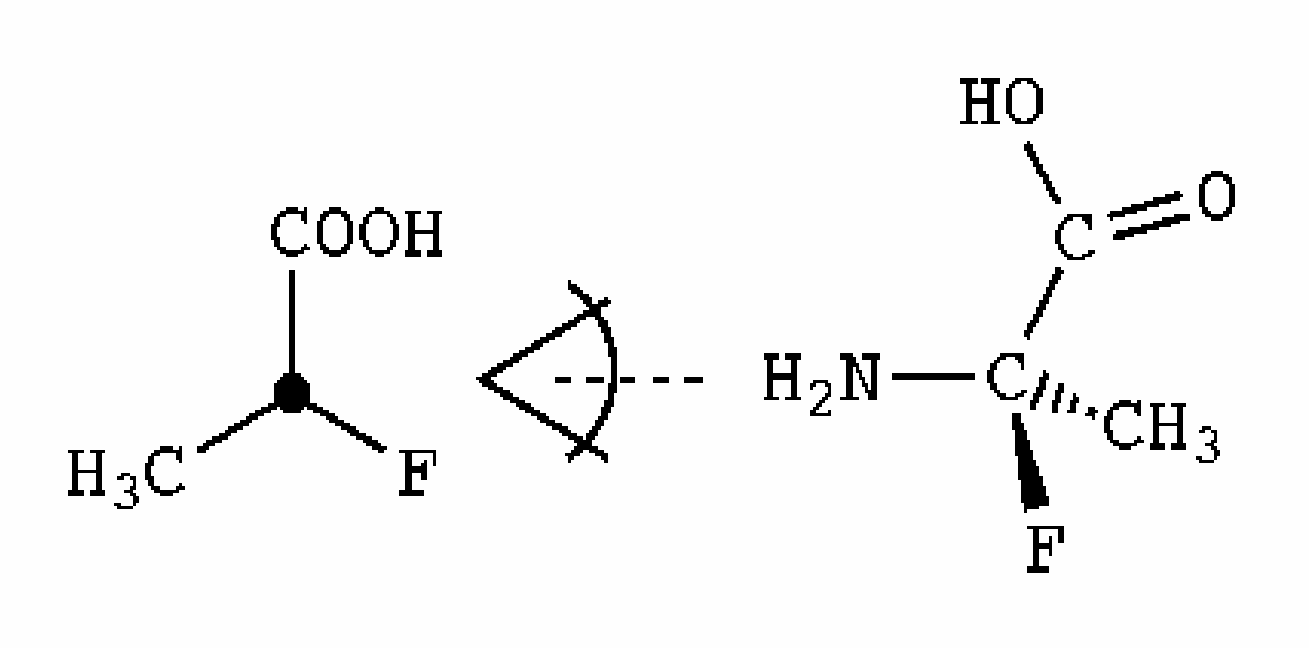
НЕ трябва да има симетрия около хиралния център, в противен случай няма смисъл указването на хиралност.

Има два ентантиомера, които се специфицират в <u>съкратен вид</u> съответно с означенията @ и @@

@ - означава, че ако гледаме от атом 1 към хиралния център другите атоми 2, 3 и 4 са подредени обратно на часовниковата стрелка **N[C@](O)(C)Cl**

@@ - атоми 2, 3 и 4 са подредени по посока на часовниковата стрелка **N[C@@](O)(C)Cl**

N **C** (O) C Cl
1 2 3 4



N[C@](C)(F)C(=O)O

Задаване на тетраедрична хиралност

Ако атомът, при който се дефинира хиралност има H атоми по подзрабиране, те трябва да се опишат изрично вътре в квадратните скоби след символа за хиралния клас:

N[C@H]CCl

Ако хиралният център участва в затваряне на цикли посредством индекси, индексът се брои за атом.

1 2 4

N1CCCO[C@H]1CC

3

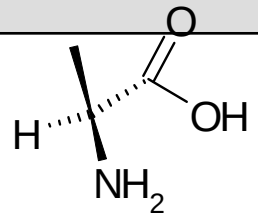
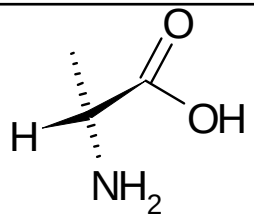
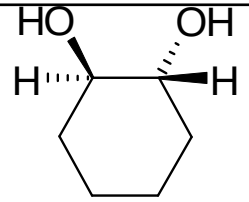
Тетраедрична хиралност може да се специфицира в пълна форма:

@TH1 (@ - кратка форма)

@TH2 (@@ - кратка форма)

Възможни са и кратките форми: **@1 @2**

Примери: Конфигурация на тетраедрична хиралност

2D	SMILES	Име	Обяснения
	N[C@@H](C)C(=O)O	L-alanine	Погледнати от N: **H**, **метиловата група** и **карбоксилната група** са разположени по посока на часовниковата стрелка
	N[C@H](C)C(=O)O	D-alanine	Погледнати от N: **H**, **метиловата група** и **карбоксилната група** са разположени обратно на часовниковата стрелка
	O[C@@H]1CCCC[C@H]1O	trans-resorcinol	За първия хирален атом: погледнати от OH групата : **H**, **COH** и **C** са разположени по посока на часовниковата стрелка

Обща спецификация на хиралност

@ <клас> <подредба>

Хиралният клас се задава с двубуквен код. *Подредбата се задава с положително число*

Класове хиралност:

TH – тетраедрична (@TH1, @TH2) - *по подразбиране за атоми с 4 съседа)*

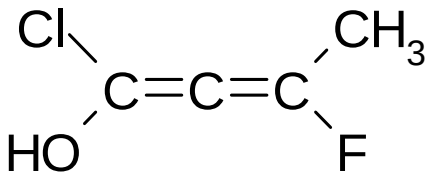
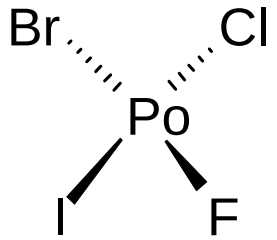
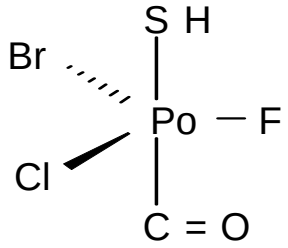
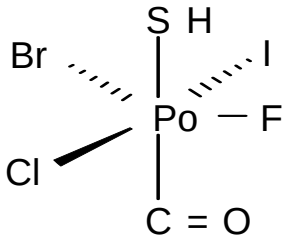
AL – аленил (@AL1, @AL2) - *по подразбиране за атоми с 2 съседа, например =C=*

SP – квадратно –планарна (@SP1…@SP3)

TB – тригонална-бипирамидална (@TB1…@TB20) - *по подразбиране за атоми с 5 съседа*

OH – октаедрична (@OH1…@OH30) - *по подразбиране за атоми с 6 съседа*

Примери: Обща хиралност

Клас	2D/пример	SMILES
AL Allene-like		OC(Cl)=[C@]=C(C)F или OC(Cl)=[C@AL1]=C(C)F
SP Square-planar		F[Po@SP1](Cl)(Br)I
TB Trigonal-bipyramidal		O=C[As@](F)(Cl)(Br)S
OH Octahedral		O=C[Co@](F)(Cl)(Br)(I)S

Проблеми при структурното представяне

ароматност

тавтомери

неорганични и координационни съединения

макромолекули и полимери

Маркуш структури

канонизация (уникалност)

автоматично разпознаване на цикли

автоматично конвертиране от номенклатура в ТС и обратно

SMARTS

Субструктурното търсене е една от най-важните практически задачи в химичната информатика

SMARTS е език, чрез който може да описват субструктури (субструктурни шаблони). Те се използват за да се оформят заявки за търсене в структурни бази данни

SMARTS е директно продължение на езика SMILES

Всеки SMILES е и валиден SMARTS (с много редки изключения), **но значението му се променя**.

SMARTS

!!! При SMARTS не се прави насищане с водородни атоми по подразбиране.

Стрингът “CC” има различно значение при SMILES и SMARTS

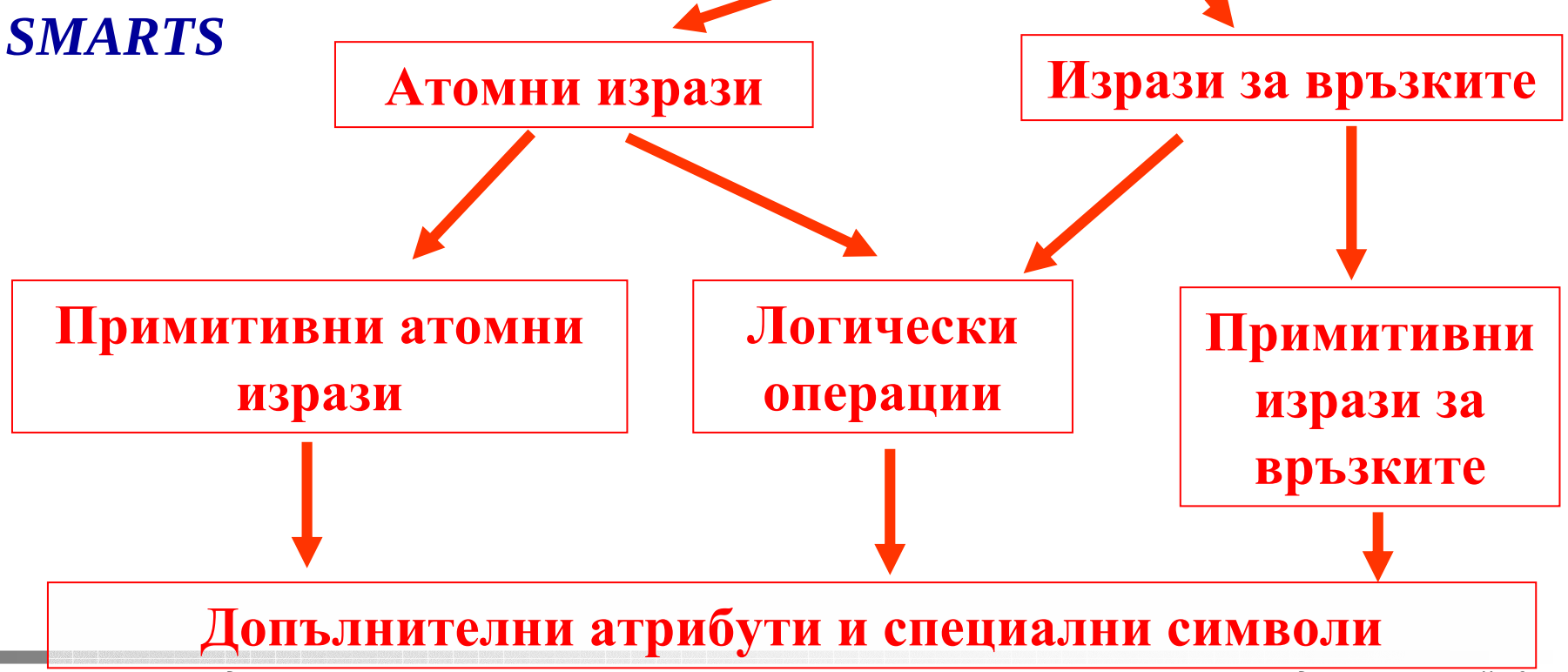
SMILES	SMARTS
CC описва моекулата на етана. CC е еквивалентен на [CH3][CH3]	CC описва двуатомен фрагмент. CC е еквивалентен на [C][C]

SMARTS

основните компоненти на графа са разширени за да поддържат логически операции и специфични атрибути

SMILES представя структурния граф **S[A,B]**

SMARTS



SMARTS – атомни примитиви

символ	значение	стойност по подразбиране
*	произволен атом (шаблон)	
a	произволен ароматен атом (шаблон)	
A	произволен алифатен атом (шаблон)	
D<n>	топологична степен (degree); <n> указва броя на експлицитно указаните връзки	1
H<n>	общ брой на водородните атоми - <n>	1
h<n>	брой на имплицитните водородни атоми - <n>	1
R<n>	атомът участва точно в <n> цикъла	атом в цикъл
r<n>	атомът участва в цикъл с размер <n>	атом в цикъл

SMARTS – атомни примитиви

символ	значение	стойност по подразбиране
v<n>	валентност <n>	1
X<n>	брой на всички връзки за този атом <n>	1
- <n>	отрицателен заряд	-1
+ <n>	положителен заряд	+1
#<n>	тип на атома указан чрез номер на елемента	
@	хиралност – обратно на часовниковата стрелка	
@@	хиралност – по часовниковата стрелка	
<n>	атомна маса (изотоп)	

SMARTS – примитиви за връзки

символ	значение
-	единична връзка (алифатна)
/	единична връзка - посока “нагоре”
\	единична връзка - посока “надолу”
/?	единична връзка - посока “нагоре или неуказана”
\?	единична връзка - посока “надолу или неуказана”
=	двойна връзка
#	тройна връзка
:	ароматна връзка
~	произволна връзка (шаблон)
@	произволна връзка в цикъл

SMARTS – логически операции

символ	израз	значение
!	!e1	отрицание (“не e1”)
&	e1&e2	логическо “И” (висок приоритет)
,	e1,e2	логическо “ИЛИ”
;	e1;e2	логическо “И” (нисък приоритет)
	e1e2	при липса на логическа операция се подразбира логическо “И” с висок приоритет

SMARTS – примерни атомни изрази

SMARTS	значение
[CH2]	алифатен въглерод с 2 водорода
[!C;R]	атом, който не е алифатен въглерод и е в цикъл
[!C;!R0]	същото като горното (“!R0” означава “не е в 0 цикъла”)
[n;H1]	ароматен азот с 1 водород
[n&H1]	същото като горното
[nH1]	същото като горното
[c,n&H1]	ароматен въглерод или ароматен азот с 1 водород
[c,n;H1]	атом, който е ароматен въглерод или ароматен азот и този атом има точно 1 водород

SMARTS – примерни атомни изрази

SMARTS	значение
[X3&H0]	атом с три връзки и никакви водороди
[35*]	произволен атом с маса 35
[35Cl]	Хлорен атом с маса 35
[F,Cl, Br, I]	атом, който е някой от първите 4 халогени
[#6]	въглероден атом (алифатен или ароматен)
[C,c]	алифатен или ароматен въглерод (като горното)

SMARTS – рекурсивни атомни изрази

Нотацията SMARTS позволява да се задават по сложни топологични конфигурации около даден атом. За целта се използва така наречения рекурсивен SMARTS.

[... $(валиден SMARTS израз)...]

В логическия израз за даден атом може да се ‘вмъкнат’ цели фрагменти, които са описани чрез друг валиден SMARTS израз.

SMARTS – примерни рекурсивни атомни изрази

SMARTS	значение
[$(CC), $(CO)]	въглероден атом, който участва в една от групите CC или CO
[C;!$(CCO)]	въглероден атом, който не участва във фрагмент CCO
[C;$(C*O)]	въглероден атом, който е на разстояние 2 от кислороден атом т.е. участва във фрагмент C*O

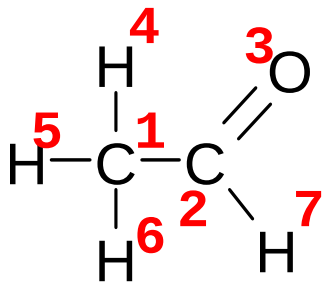
Двумерно представяне на структури (2D)

Двумерното представяне се състои от

(1) Таблица на свързаност

(2) Списък с координатите на атомите

Двумерното представяне може да се разглежда като разширена таблица на свързаност, при което е разширен атомният списък



ATOMLIST		X	Y
1	C	0	0
2	C	1	0
3	O	2	1
4	H	0	1
5	H	-1	0
6	H	0	-1
7	H	2	-1

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

3D представяне

Използват се два основни подхода за описание на геометрията (3D конфигурацията) на химични обекти.

Използват се съответно два основни типа координатни системи:

(1) Декартови координати

(2) Вътрешни координати

Подходи за получаване на 3D информация

Експериментални методи:

- ренгенова кристалография
- микровълнова спектроскопия
- електронна дифракция
- ЯМР

In silico / компютърно моделиране

- квантова химия
- молекулна механика
- 3D генератори

Файлови формати

MOL (SDF)

MOL2

XYZ

PDB

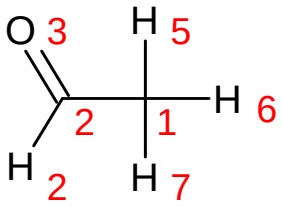
CTX

HIN

CML

…

Структура на .mol (V2000)



1 2 3	ACD/Labs02121017162D	Заглавен блок	
4	7 6 0 0 0 0 0 0 0 0 1 V2000	Начален ред	Таблица на свързаност
5 6 7 8 9 10 11	13.8288 -11.3611 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 15.1588 -11.3611 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 13.1638 -10.2093 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 13.1638 -12.5129 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 15.1588 -12.6911 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 16.4888 -11.3611 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0 15.1588 -10.0311 0.0000 H 0 0 0 0 0 0 0 0 0 0 0 0	Блок на атомите	
12 13 14 15 16 17	1 2 1 0 0 0 0 1 3 2 0 0 0 0 1 4 1 0 0 0 0 5 2 1 0 0 0 0 6 2 1 0 0 0 0 7 2 1 0 0 0 0	Блок на връзките	
18	Log P (octanol-water): Value : 0.35	Блок на свойства	
19	M END		

Заглавен блок на Molfile (v2000)

Описание на параметрите		име на потре - бителя	име на програма- та, чрез която е създаден файла	дата /час на създаване на файла		пространствена структура	мащаб	енергия	вътрешен ре- гистрационен номер
Стойности			ACD/Labs	021210	1716	2D			

Заглавен ред на таблицата на свързаност

Описание на параметрите	Брой атоми	Брой връзки	Брой на атомни списъци	Излязъл от употреба	Знак за хирал- ност	Свойства, не- поддържани в Molfile					Брой до- пълнителни свойства	Версия на таблицата на свързаност
Стойности	18	18	0	0	0	0	0	0	0	0	1	V2000

Блок на атомите

Описание на параметрите	Декартови координати			Атомен символ	Масова разлика	Заряд	Друга информация*
	x	y	z				
Стойности	13.1638	-10.2093	0.0000	O	0	0	0 0 0 …
	13.1638	-12.5129	0.0000	H	0	0	0 0 0 …
	15.1588	-12.6911	0.0000	H	0	0	0 0 0 …

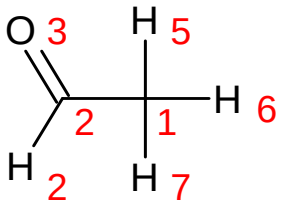
*Заряд, атомна маса, пространствена конфигурация, валентност и др.

Блок на връзките

Описание на параметрите	Първи атом	Втори атом	Порядък на връзката	Конфигурация на връзката	Друга информация*
Стойности	1	2	1	0	0 0 0 …
	1	3	2	0	0 0 0 …
	1	4	1	0	0 0 0 …

*Пространствена конфигурация, реакционен център и др.

Структура на .mol (V3000)



1 2 3 4	ACD/Labs02121017172D 0 0 0 0 0 0 0 0 0 0 0999 V3000	Заглавен блок	
5	M V30 BEGIN CTAB	Начален ред	Таблица на свързаност
6	M V30 COUNTS 7 6 0 0 0	Описание	
7 8 9 10 11 12 13 14 15	M V30 BEGIN ATOM M V30 1 C 13.8288 -11.3611 0 0 M V30 2 C 15.1588 -11.3611 0 0 M V30 3 O 13.1638 -10.2093 0 0 M V30 4 H 13.1638 -12.5129 0 0 M V30 5 H 15.1588 -12.6911 0 0 M V30 6 H 16.4888 -11.3611 0 0 M V30 7 H 15.1588 -10.0311 0 0 M V30 END ATOM	Блок на атомите	
16 17 18 19 20 21 22 23	M V30 BEGIN BOND M V30 1 1 1 2 M V30 2 2 1 3 M V30 3 1 1 4 M V30 4 1 5 2 M V30 5 1 6 2 M V30 6 1 7 2 M V30 END BOND	Блок на връзките	
24	M V30 END CTAB		
25	M END		

Блок на атомите

Описание на параметрите		Номер на атома	Атомен символ	Декартови координати			Atom-atom mapping	Друга информация*
				x	y	z		
Стойности	M V30	3	O	13.1638	-10.2093	0.0000	0	0…
	M V30	4	H	13.1638	-12.5129	0.0000	0	0…
	M V30	5	H	15.1588	-12.6911	0.0000	0	0…

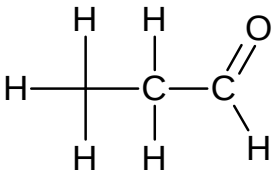
*Заряд, атомна маса, пространствена конфигурация, валентност и др.

Блок на връзките

Описание на параметрите		Номер на връзката	Тип на връзката	Първи атом	Втори атом	Друга информация*
Стойности	M V30	1	1	1	2	…
	M V30	2	2	1	3	…
	M V30	3	1	1	4	…

*Пространствена конфигурация, реакционен център и др.

Структура на Tripos .mol2




```
# Name:         no166
# Creation time: Mon Jan 25 11:31:13 1999
```

Коментарни редове

```
@<TRIPOS>MOLECULE
no166
  10    9    1    1    1
SMALL
AMPAC_CHARGES
```

Обща информация за молекулата

```
@<TRIPOS>ATOM
     1 H1        -0.5164   -0.5501    2.7255 H       1 <1>      -0.002
     2 C2         0.0234   -1.2585    3.3732 C.3     1 <1>       0.029
     3 C3         0.9380   -2.1755    2.5202 C.3     1 <1>      -0.075
     4 C          0.1553   -2.9713    1.5084 C.2     1 <1>       0.246
     5 O          0.7381   -3.7506    0.7729 O.2     1 <1>      -0.297
     6 H6         0.6261   -0.6840    4.0952 H       1 <1>       0.003
     7 H7        -0.7102   -1.8627    3.9300 H       1 <1>      -0.002
     8 H8         1.4758   -2.8786    3.1765 H       1 <1>       0.037
     9 H9         1.6811   -1.5656    1.9825 H       1 <1>       0.037
    10 H10       -0.9257   -2.8708    1.4126 H       1 <1>       0.025
```

Информация за атомите в молекулата

[номер] [име на атома] [X] [Y] [Z] [тип на атома] [номер на подструктура] [име на подструктура][заряд]

```
@<TRIPOS>BOND
    1    2    1 1
    2    3    2 1
    3    4    5 2
    4    4    3 1
    5    2    6 1
    6    2    7 1
    7    3    8 1
    8    3    9 1
    9    4   10 1
```

Информация за връзките в молекулата

[номер] [1-ви атом] [2-ри атом] [тип (кратност)]

Файлов формат - XYZ

<брой атоми>
линия за коментар
атомен_символ$_1$ x_1 y_1 z_1

атомен_символ$_2$ x_2 y_2 z_2

…

атомен_символ$_n$ x_n y_n z_n

Файлов формат – XYZ (TINKER)

N – броят атоми

...

...

<Номер> <Елемент> <X> <Y> <Z> <Атомен Клас> <първа околност>

...

...

Файлов формат - PDB

```
HEADER    MOLECULE
COMPND
AUTHOR    GENERATED BY PCMODEL V7.0
ATOM      1  C    UNK   1      0.267  -0.390  -0.764  1.00  0.00
ATOM      2  C    UNK   1     -0.241   0.102   0.594  1.00  0.00
ATOM      3 Cl    UNK   1     -0.270   1.889   0.654  1.00  0.00
ATOM      4 Cl    UNK   1      0.297  -2.177  -0.824  1.00  0.00
ATOM      5  H    UNK   1      1.310  -0.050  -0.962  1.00  0.00
ATOM      6  H    UNK   1     -0.395  -0.057  -1.597  1.00  0.00
ATOM      7  H    UNK   1     -1.284  -0.238   0.792  1.00  0.00
ATOM      8  H    UNK   1      0.421  -0.231   1.427  1.00  0.00
CONECT    1    2    4    5    6
CONECT    2    1    3    7    8
CONECT    3    2
CONECT    4    1
CONECT    5    1
CONECT    6    1
CONECT    7    2
CONECT    8    2
MASTER        0    0    0    0    0    0    0    0    8    0    8    0
END
```

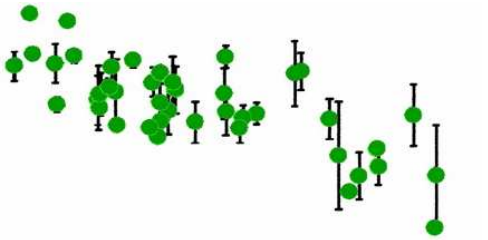
Молекулни дескриптори

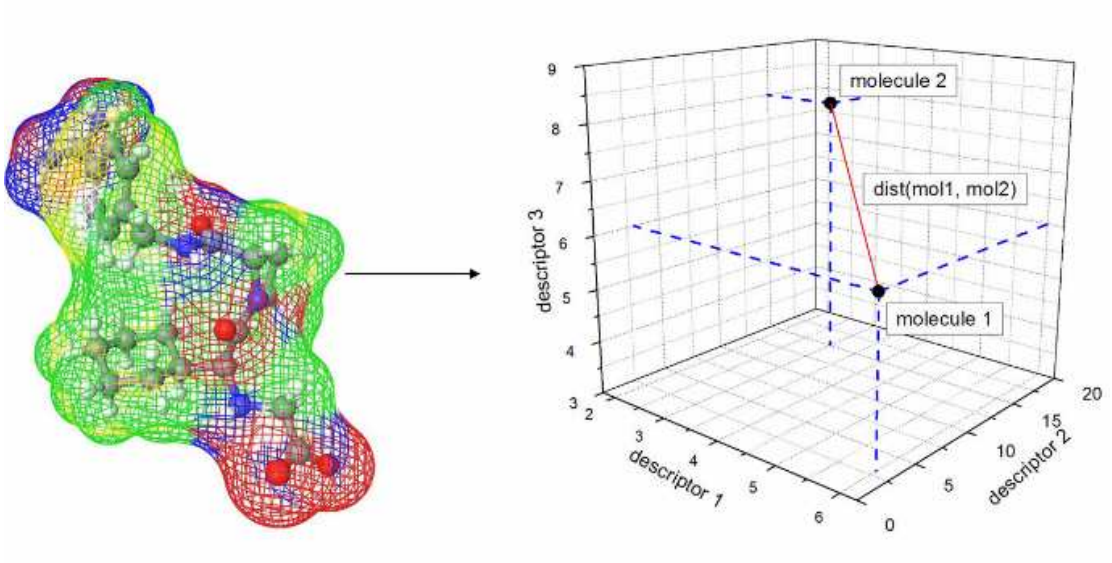
Конституционни дескриптори

Топологични дескриптори

Геометрични дескриптори

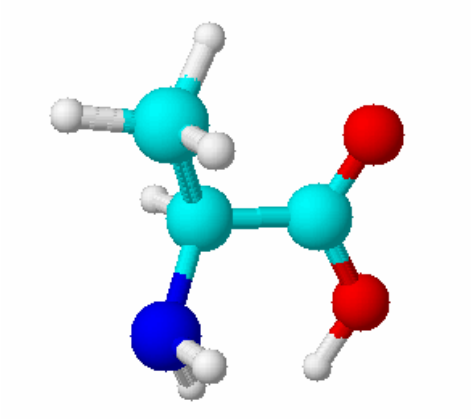
Електростатични дескриптори





Молекулни дескриптори

Молекулните дескриптори описват (*кодират*) информация за структурата и свойствата на химичните съединения.



$d_1, d_2, \ldots, d_k$

Молекулни дескриптори

Чрез молекулните декриптори се получава химиметричен образ на обекта.

G[V,E], 3D свойства → $d_1, d_2, \ldots, d_k$

Химичната информация се трансформира от ниво **химичен граф** на ниво **белези (дескриптори)** на обекта.

За вектора с молекулните дескриптори може да се приложат множество математически и химиметрични методи.

Класификация според вида на данните

Булеви: 0 или 1 (false, true)

Биполярни: -1 или +1

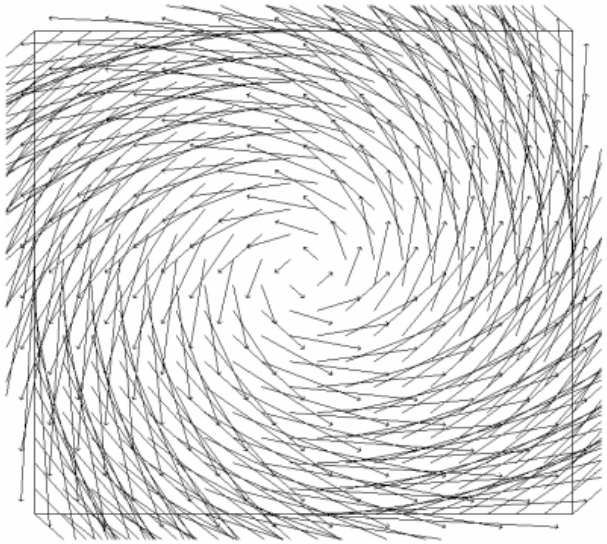
Скаларни: цели или реални числа

Вектори: $(d_1, d_2, \ldots, d_k)$

Матрици:

$$\begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{pmatrix}$$

Полета (скаларни, векторни)



Класификация според химичната информация

Конституционни

Топологични

Геометрични

Електростатични

Квантово-химични

МО / базирани на молекулни орбитали

Термодинамични

Биохимични

“Молекулни свойства”

“QSAR/QSPR дескриптори”

Класификация според размерността на структурното представяне

Вълнова ф-я или EP повърхнина

4D - конформации

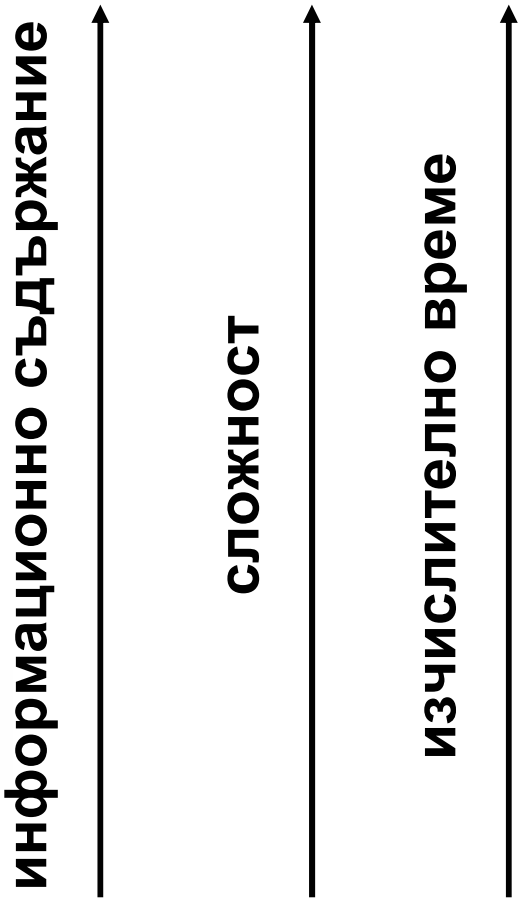
3D -тримерна информация

2.5D - хибридни представяния

2D - топологични

1D - конституционни

0D - конституционни



Изисквания за молекулните дескриптори

Възможност за ефективно изчисляване от наличните структурни представяния

Корелация с молекулни свойства

Дискриминантна способност - *различните класове химични обекти трябва да имат статистически различими стойности за дескриптора*

Адекватно използване за получаване на модели

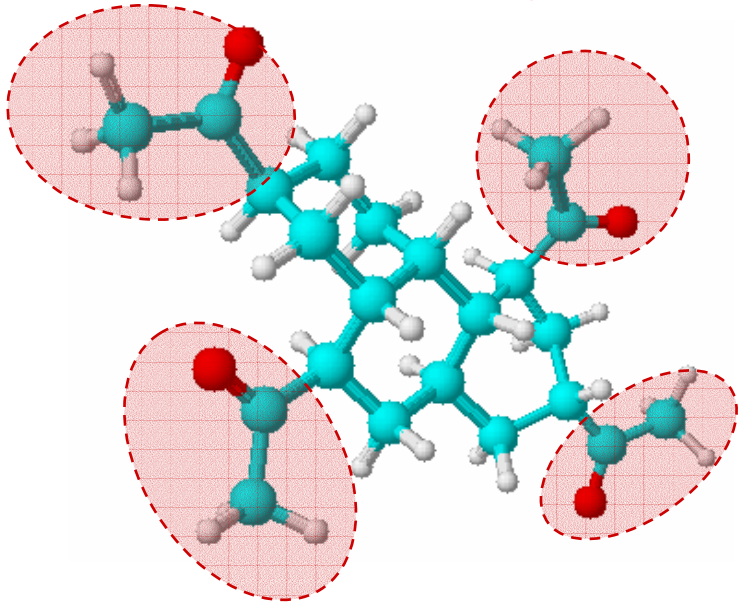
Конституционни дескриптори

0D дескриптори - не зависят от топологията.
Използват се и физико-химични свойства.

1D дескриптори - брой малки/функционални групи;
частично зависят от топологията.

Типично обозначение: N_X – честота на поява на група **X**

$N_{Acetyl} = 4$



0D дескриптори

структурен граф S[A,B] (G[V,E])

N_A - брой атоми в молекулата $N_A = |A|$

N_B - брой връзки $N_B = |B|$

MW - молекулно тегло

AMW – средно молекулно тегло $AMW = \frac{MW}{N_A}$

N_C, N_H, N_{Cl},… - брой въглеродни, водородни, хлорни, ... атоми в молекулата

1D дескриптори

Брой на функционални групи: N_{OH}, N_{CH3},N_{COOH}, …

Брой на B групите N_{A-B}

Брой на G групите N_{A-B-C}

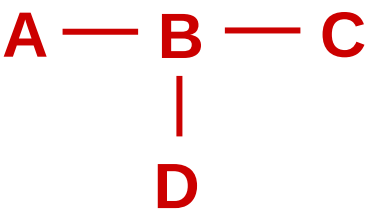
Брой на D групите $N_{A-B(-C)-D}$

1D дескриптори

B група A — B

G група A — B — C

D група



Топологични дескриптори

Топологичните дескриптори се изчисляват от топологичното представяне на структурата (граф или таблица на свързаност).

Топологичните дескриптори могат да бъдат:

матрични представяния

индекси

характеристични полиноми

Топологични индекси

Топологичният индекс е инвариант на структурния граф - не зависи от конкретното представяне или номериране на графа.

$$TI = f(G)$$

Възможно е две различни структури да имат една и съща стойност на топологичния индекс.

Топологичният индекс се характеризира със своята **степен на израждане** – *размера на най-малката структура, за която стойността на топологичния индекс се дублира със стойността за друга структура.*

Матрица на съседство

$$A(G) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

a_{ij} = 1 ако атомите **i** и **j** са съседни (свързани с връзка)

a_{ij} = 0 ако атомите **i** и **j** не са съседни

Матрица на разстоянията

$$D(G) = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

$\mathbf{d_{ij}}$ най-късото разстояние между атоми **i** и **j**

Реципрочна матрица на разстоянията

$$RD(G) = \begin{pmatrix} 0 & \frac{1}{d_{12}} & \dots & \frac{1}{d_{1n}} \\ \frac{1}{d_{21}} & 0 & \dots & \frac{1}{d_{2n}} \\ \dots & \dots & \dots & \dots \\ \frac{1}{d_{n1}} & \frac{1}{d_{n2}} & \dots & 0 \end{pmatrix}$$

d_{ij} са елементите на матрицата на разстоянията **D(G)**

Де-тур матрица

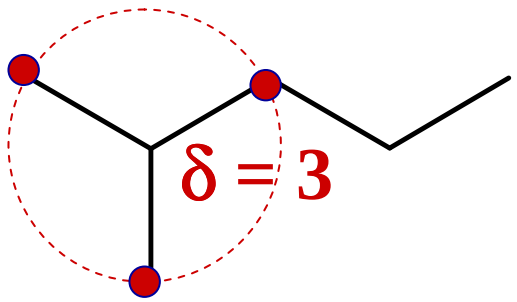
$$\Delta_{ij} = \begin{cases} \max(l(p_{ij})) & i \neq j \\ 0 & i = j \end{cases}$$

$l(p_{ij})$ е дължината на пътя **p_{ij}** между атоми **i** и **j**

За ациклични структури де-тур матрицата съвпада с матрицата на разстоянията

Топологична степен на атом

δ - броят на съседите на даден атом от структурата т.е. размера на неговата първа топологична околност; *броят на връзките, към този атом независимо от тяхната валентност.*



Топологична степен на атом

Вектор с топологичните степени

$$(\delta_1, \delta_2, ..., \delta_n)$$

Матрица на топологичните степени

$$DEG(G) = \begin{pmatrix} \delta_1 & 0 & ... & 0 \\ 0 & \delta_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \delta_n \end{pmatrix}$$

Матрица на Лаплас

L(G) - разликата от матрицата на степените и матрицата на съседство

L(G) = DEG(G)-A(G)

$$L_{ij} = \begin{cases} \delta_i & i = j \\ -1 & i \text{ е свързан с } j \\ 0 & i \text{ не е свързан с } j \end{cases}$$

Матрица на Лаплас

$$L(G)=\begin{pmatrix} \delta_1 & -a_{12} & \dots & -a_{1n} \\ -a_{21} & \delta_2 & \dots & -a_{22} \\ \dots & \dots & \dots & \dots \\ -a_{n1} & -a_{n2} & \dots & \delta_n \end{pmatrix}$$

Собствени стойности на матрица

За дадена матрица $M = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1n} \\ m_{21} & m_{22} & \dots & m_{2n} \\ \dots & \dots & \dots & \dots \\ m_{n1} & m_{n2} & \dots & m_{nn} \end{pmatrix}$

Векторът $\mathbf{x} = (x_1, x_2, \dots, x_n)$ е ‘собствен’ със съответна ‘собствена’ стойност λ, ако

$$Mx = \lambda x$$

Собствени стойности на матрица

Собствените стойности на матрицата: $\lambda_1, \lambda_2, \ldots, \lambda_n$ състявляват така наречения спектър на матрицата:

$$Sp(M) = \{\lambda_1, \lambda_2, ..., \lambda_n\}$$

$$Sp(M) = \{Sp(M)_1, Sp(M)_2, ..., Sp(M)_n\}$$

Собствени стойности на матрица

Собствените стойности на матрицата могат да се получат с помощта на числени итеративни методи.

Не всяка мартица има реални собствени стойности.

За симетрична матрица може да се гарантира, че **всичките собствени стойности са реални**

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$$

Индекс на Уинер

Оригинална дефиниция на Уинер за ациклични алкани

$$W = \sum_{e_{ij} \in E(G)} N_i N_j$$

N_i е броят на върховете от страната на атом i

Универсална дефиниция на Хосоя

$$W(G) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} [D(G)]_{ij}$$

Ad hoc индекси

Терминът **ad hoc** се използва за индекси, които са конструирани с определена цел, но в последствие се намира по-широко приложение на тези индекси.

Индексът на Уинер **W** е типичен случай на ad hoc индекс.

Индексът на Уинер корелира много добре с температурите на кипене на алканите и описва адекватно степента на разклоненост на структурите.

В по-грубо приближение, **W** е пропорционален на молекулния обем.

Индекс на Хосоя

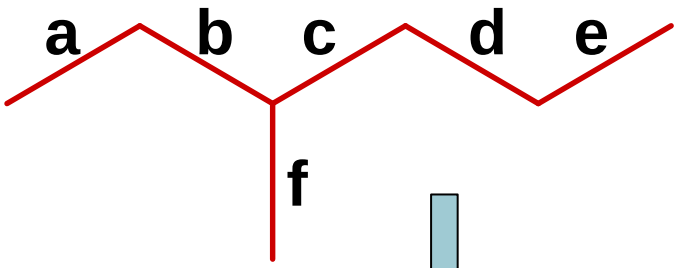
$$Z = \sum_{k=0}^{L} Z_k = \sum_{k=0}^{L} m(G,k)$$

Z_k=m(G,k) е броят на **k съответствията** за графа G.

k съотъветствие е комбинация от k връзки в графа, които не са съседни (ицидентни).

По дефиниция: Z_0 = 1, Z_1 = броят връзките

Индекс на Хосоя



2-съответствия:

(a,c) (a,d) (a,e) (a,f) (b,d)
(b,e) (c,e) (f,d) (f,e)

$Z_2 = 9$

3-съответствия:

(a,c,e) (a,f,d) (a,f,e)

$Z_3 = 3$

4-съответствия няма

$Z_4 = 0, Z_5 = 0, \ldots$

$$Z = Z_0+Z_1+Z_2+Z_3 = 1 + 6 + 9 + 3 = 19$$

Индекс на Рандич

$$\chi = \sum_{\substack{всички \\ връзки}\ (i,j)} \frac{1}{\sqrt{\delta_i\ \delta_j}}$$

δ_i е степента на i-тия атом от структурата

Индекс на Рандич

Индексът на Рандич е създаден с цел неговите стойности да корелират с някои свойства на алканите.

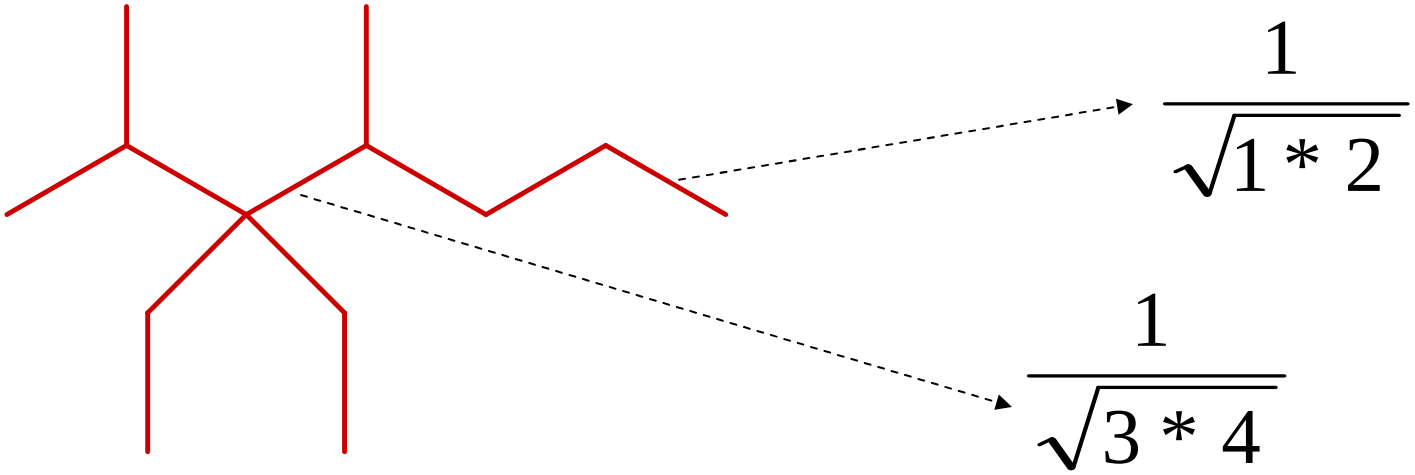
Връзките се класифицират според степените на атомите **(δ_i, δ_j)**.

Индексът представлява адитивна схема по връзките. На всяка връзка **(δ_i, δ_j)** се съпоставя инкремент:

$$\frac{1}{\sqrt{\delta_i \delta_j}}$$

Индекс на Рандич

Индексът на Рандич симулира приноса на различните връзки към молекулната повърхност:



По-вътрешните ‘закрити’ връзки (4,4), (3,4), (3,3) ... поулчават малки тегла: 0.25, 0.28, 0.33, ...

Периферните ‘открити’ връзки (1,2), (1,3), (1,4) ... поулчават големи тегла: 0.70, 0.57, 0.50, ...

Индекс на Балабан

Балабан предлага индекс, който е аналогичен на индекса на свързаност на Рандич

Индексът е адитивна схема с инкременти на връзките:

$$\frac{1}{\sqrt{DS_i \, DS_j}}$$

Индексът Балабан има много по-голяма дискриминантна способност от индекса на Рандич

Индекс на Балабан

$$J = \frac{m}{\mu + 1} \sum_{\substack{\text{всички} \\ \text{връзки}(i,j)}} \frac{1}{\sqrt{DS_i \, DS_j}}$$

m е броят на връзките

DS_i е сумата от елементите на i-тия ред в матрицата на разтоянията т.е. **$DS_i = d_{i1} + d_{i2} + \ldots + d_{in}$**

μ е цикломатичното число – минималният брой връзки, които като се премахнат превръщат структурата от циклична в ациклична

Индекси на сързаност от по-висок ред

Индексите на свързаност са обобщения на базовия индекс на **Рандич**

$$\chi = {}^{1}\chi$$

Аналогично се изчисляват индекси на свързаност от по-висок ред с теглова схема дефинирана за пътищата в графа с определена дължина **m**:

$${}^{m}\chi = \sum_{\substack{\text{всички} \\ \text{пътища} \\ (i_1, i_2, \ldots i_m)}} \frac{1}{\sqrt{\delta_{i_i}\,\delta_{i_2}\ldots\delta_{i_m}}}$$

Геометрични дескриптори

Геометричните дескриптори се изчисляват от 3D представянето на молекулата.

$$d = f(R_1, R_2, ..., R_n)$$

$$R_i = (x_i, y_i, z_i)$$

Геометричните дескриптори съдържат много повече информация от топологичните (2D) дескриптори.

Геометрични дескриптори

Изисква се оптимизация на геометрията → значително време за изчисления.

За конформационно гъвкавите молекули, значително се увеличава сложността на информацията (4D) и съответно се усложняват подходите за изчисляване и използване на дескрипторите.

Много геометрични дескриптори изискват ‘подравняване’ на молекулите в 3D пространството.

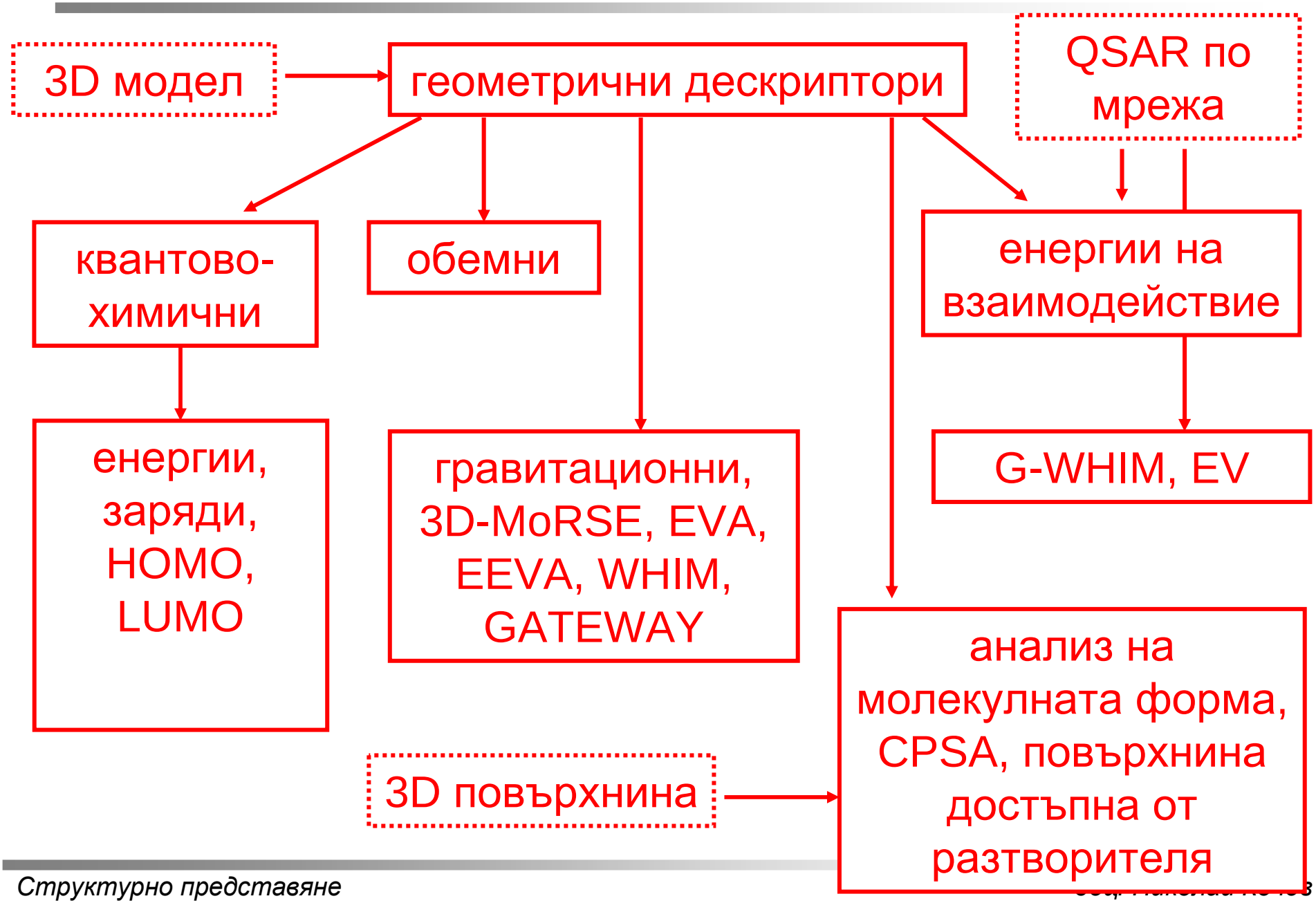
Употреба на дескрипторите

0D, 1D и 2D дескрипторите се използват за скрининг на големи бази данни и търсене по подобие.

3D дескрипторите се използват за реализация на модели за сложни молекули свойства и биологична активност.

QSAR моделите експлоатират високо информационното съдържание на геометричните дескриптори.

Класификация на 3D дескрипторите



Молекулен обем и лице на повърхнина

Молекулният обем се получава като от сумата на Ван дер ваалсовите обеми се извадят сеченията на сферите:

$$V_{VdW} = \sum_{i} V_{VdW}^{(i)} - \sum_{i,j} V_{i,j}$$

Аналогично се изчислява молекулната Ван дер Ваалсова повърхност:

$$S_{VdW} = \sum_{i} S_{VdW}^{(i)} - \sum_{i,j} S_{i,j}$$

Център на масите (гравитационен център)

Точка в пространството със следните координати:

$$R_C = \frac{1}{M}\sum_{i=1}^{n} m_i R_i$$

m_i е масата на атом **i**; **R_i** са координатите на атом **i**; **M** е сумата от масите на всички атоми: **$M = m_1+m_2+\ldots+m_n$**

$$x_C = \frac{1}{M}\sum_{i=1}^{n} m_i x_i \qquad z_C = \frac{1}{M}\sum_{i=1}^{n} m_i z_i$$

$$y_C = \frac{1}{M}\sum_{i=1}^{n} m_i y_i$$

Матрица на 3D разстоянията

$$G = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}$$

нарича се още матрица на молекулната геометрия или *топографска матрица*

геометричното разстояние между атоми **i** и **j** е:

$$r_{ij} = \left\| R_i - R_j \right\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

$$r_{ii} = 0$$

Топографски индекси

Топографските индекси са множество дескриптори, които се изчисляват от матрицата на 3D разстоянията.

Топографските индекси са аналози на топологичните индекси, които се изчисляват въз основа на матрицата на топологичните разстояния.

Геометрична степен на атом (GDD)

$$G_{\sigma_i} = \sum_{j=1}^{n} r_{ij}$$

сумата от елементите на i-тия ред в топографската матрица – атомен индекс, който описва ‘централността’ на атома

средна геометрична степен (AGDD):

$$G_{\overline{\sigma}} = \frac{1}{n}\sum_{j=1}^{n} G_{\sigma_i} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}$$

3D аналози на топологични индекси

3D индекс на свързаност:

$$^{3D}\chi = \sum_{\substack{\text{всички} \\ \text{връзки } (i,j)}} \frac{1}{\sqrt{G_{\sigma_i} G_{\sigma_j}}}$$

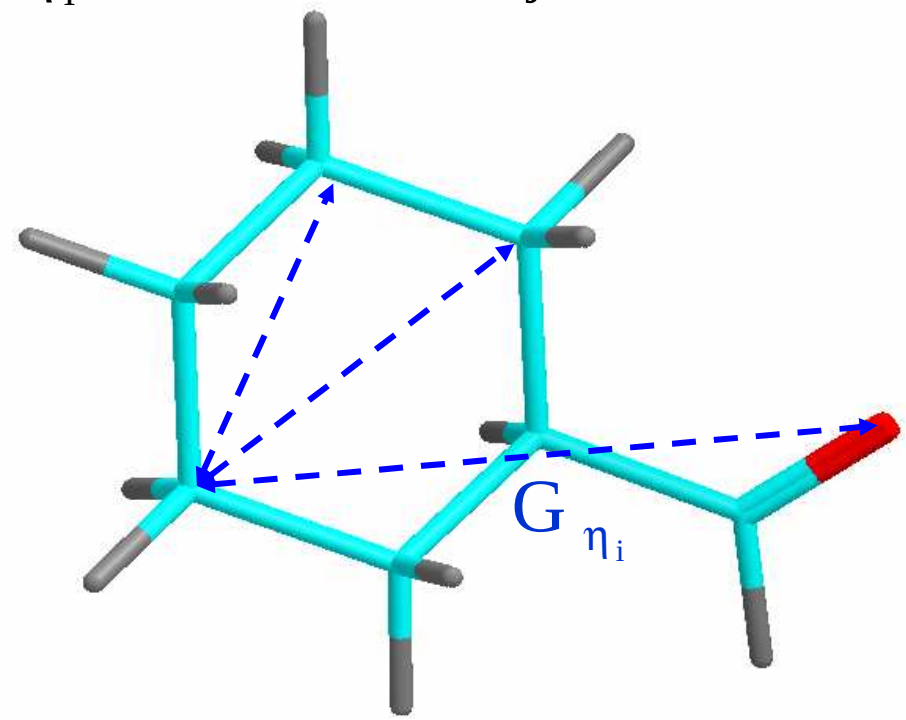
3D индекс на Уинер:

$$^{3D}W = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}$$

Атомна геометрична ‘ексцентричност’

Локалнен атомен дексриптор, който описва най-голямото разстояние за даден атом (маскималният елемент от даден ред в топографската матрица):

$$G_{\eta_i} = \max\{ r_{ij} \mid j = 1,2,..., n \}$$



Геометричен радиус и диаметър

Геометричен радиус – минималната геометрична ексцентричност

$$^{G}R = \min\{ G_{\eta_i} \mid j = 1,2,..., n\}$$

Геометричен диаметър – максималната геометрична ексцентричност

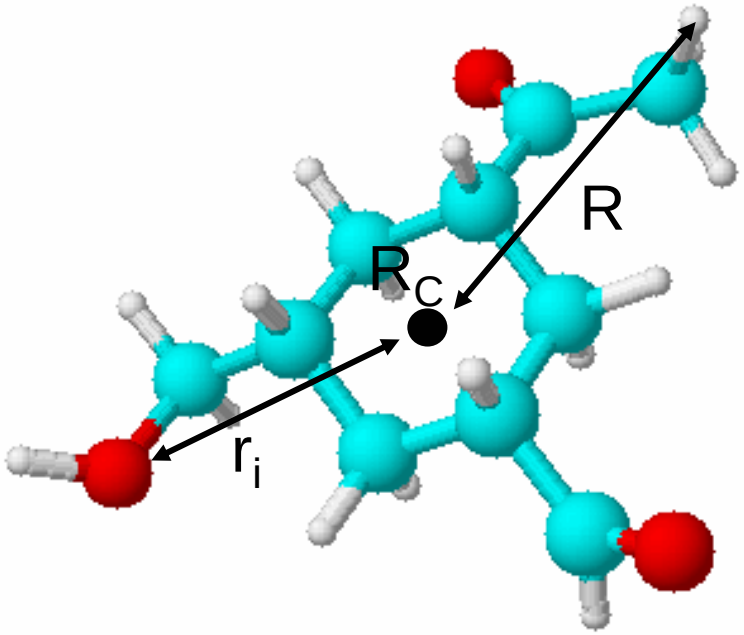
$$^{G}D = \max\{ G_{\eta_i} \mid j = 1,2,..., n\}$$

Коефициент на геометричната форма

$$I_3 = \frac{^{G}D - ^{G}R}{^{G}R}$$

Размах на молекула

Изчислява се вектор с разстоянията на атомите до центъра на масите – $(r_1, r_2, \ldots, r_n)$

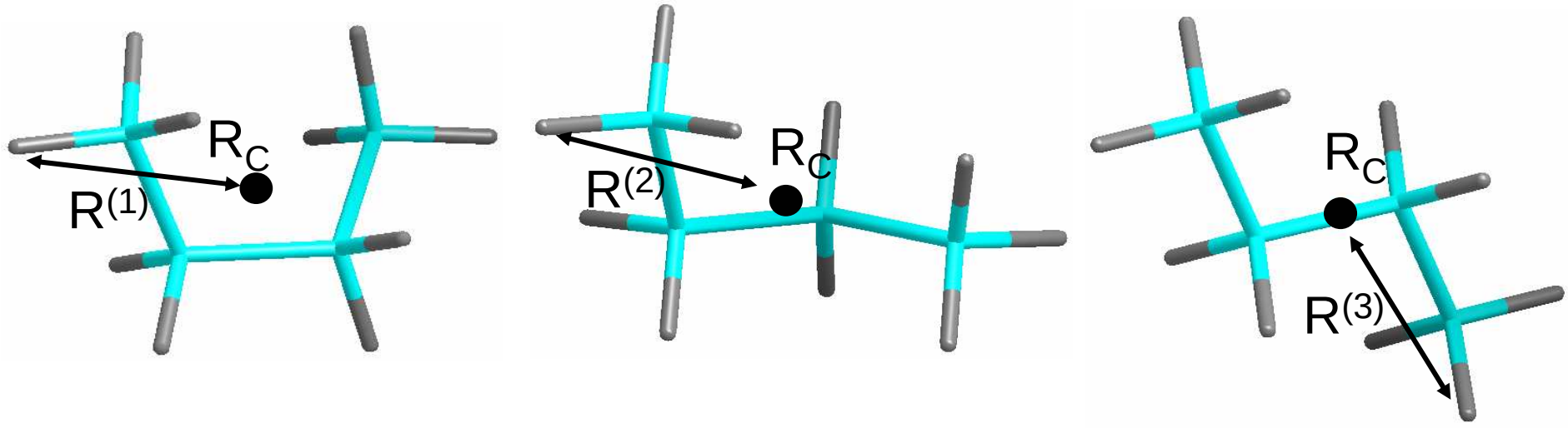


Размаха на молекулата е разстоянието между центроида на молекулата и най-отдалеченият атом:

$$R = \max\{r_i\}$$

Среден размах на молекула

Изчислява се средна стойност на размаха на молекулата за всички конформации:



$$\overline{R} = \frac{1}{N_{conf}} \sum_{i}^{N_{conf}} R^{(i)}$$

Гравитационни индекси

$$G_1 = \sum_{i<j}^{N_A} \frac{m_i m_j}{r_{ij}^2}$$

$$G_2 = \sum_{\substack{\text{всички} \\ \text{връзки } (i,j)}} \frac{m_i m_j}{r_{ij}^2}$$

Гравитационните индекси отразяват разпределението на масите в дадена молекула.

Индексите G_1 и G_2 може да се реализират и с други атомни свойства: поляризуемост, електроотрицателност

Ротационен радиус

$$R_G = \sqrt{\sum_{i=1}^{n} \frac{m_i r_i^2}{M}}$$

Ротационният радиус отразява разпределението на масите около центроида на молекулата.

Индекс на овалност

Индексът на овалност отразява факта, че от всички 3D обекти с един и същ обем, най-малка повърхина има сферата.

Индексът на овалност е отношението на лицето на молекулната повърхност S_A и лицето на повърхнината на сферата, която има обем равен на молекулния обем.

$$O = \frac{S_A}{S_{A\,0}}$$

Индекс на овалност

Референтната сфера има радиус $\mathbf{R_0}$ и обем $\mathbf{V_{VdW}}$. Радиуса се определя от обема:

$$O = \frac{S_A}{S_{A\,0}} = \frac{S_A}{4\pi R_0^{\,2}}$$

$$V_{VdW} = \frac{4}{3}\pi R_0^{\,3} \Rightarrow R_0 = \left(\frac{3V_{VdW}}{4\pi}\right)^{1/3}$$

Индексът на овалност става функция на молекулния обем и лицето на молекулната повърхност:

$$O = \frac{S_A}{4\pi\left(\frac{3V_{VdW}}{4\pi}\right)^{2/3}}$$

WHIM дескриптори

WHIM дескрипторите се получават от статистически индекси изчислени въз основа на проекциите на атомите върху главните оси.

WHIM дескрипторите отразяват съществена информация за размера, формата, симетрията и разпределението на атомите спрямо инвариантни оси.

WHIM дескрипторите се делят на два вида:

- дирекционни
- глобални

GETAWAY дескриптори

GETAWAY (**GE**ometry, **T**opology, and **A**tom-**W**eighted Assembl**Y** – клас дескриптори, които комбинират геометрични данни, топологични данни и теглова схема базирана на атомни свойства.

Дескрипторите са изчислени от специална матрица **MIM** (Molecular Influence Matrix) - матрица на молекулното влияние, която се отбелязва още с **H**

Матрицата **H**, е квадратна симетрична матрица с размер $\mathbf{n = N_A}$ и е инвариантна спрямо ротация на молекулата.

Автокорелационни дескриптори

Обща дефиниция за автокорелация на функция f(x):

$$AC_k = \int_a^b f(x) \cdot f(x+k) \cdot dx$$

AC_k характеризира как се променят стойностите на функцията при интервал с дължина **k**

При дискретна променлива $x=x_1, x_2, \ldots, x_n$:

$$AC_k = \sum_{i=1}^{n-k} f(x_i) \cdot f(x_{i+k})$$

Автокорелационни дескриптори

Автокорелация на топологичната структура:

$$ATS_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i \cdot w_j \cdot \delta(k, d_{ij})$$

където **{ w_i }** теглова атомна схема отразяваща дадено свойство; **δ(k,d)** е делта-фунцкията на Дирак:

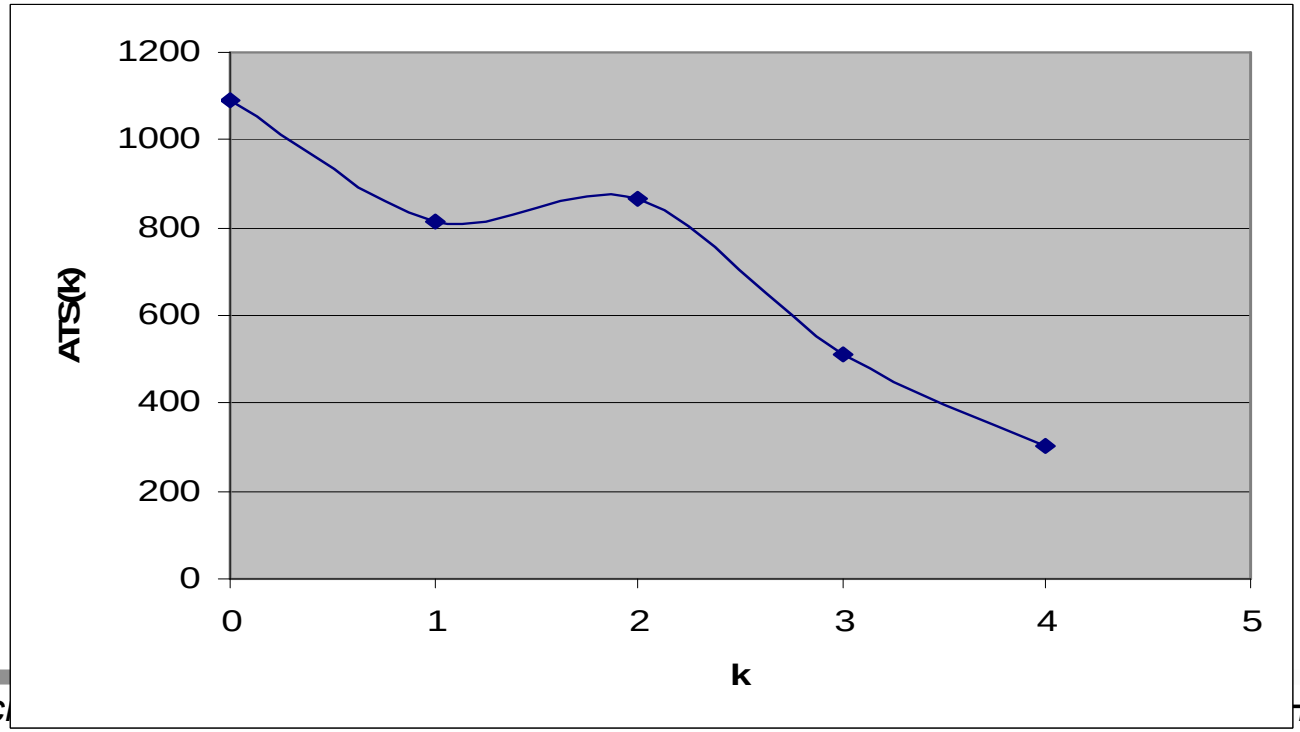
$$\delta(k, d) = \begin{cases} 1 & k = d \\ 0 & k \neq d \end{cases}$$

Автокорелационни дескриптори

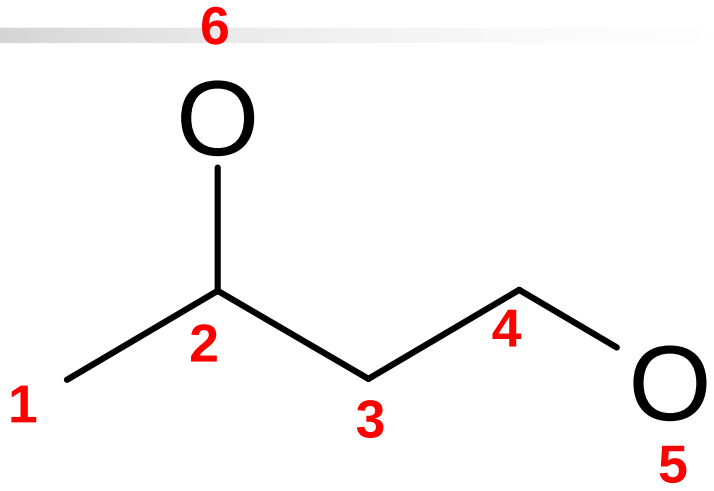
ATS векторът се дефинира като:

$$(ATS_0, ATS_1, ATS_2, ..., ATS_D)_W$$

Графиката съотвеставаща на вектора ATS се нарича ‘автокорелограма’



Автокорелационни дескриптори



$$ATS_0 = w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2$$

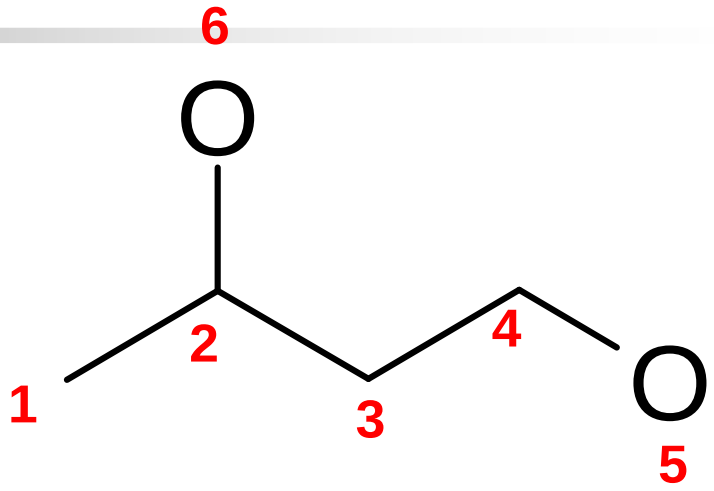
$$ATS_1 = w_1w_2 + w_2w_3 + w_3w_4 + w_4w_5 + w_2w_6$$

$$ATS_2 = w_1w_3 + w_1w_6 + w_2w_4 + w_3w_5 + w_3w_6$$

$$ATS_3 = w_1w_4 + w_2w_5 + w_2w_4 + w_4w_6$$

Автокорелационни дескриптори

тегло $w_i = m_i$ (атомната маса)



$$ATS_0 = 4*12^2 + 2*14^2 = 1088$$

$$ATS_1 = 4*12\cdot 12 + 2*12\cdot 16 = 816$$

$$ATS_2 = 2*12\cdot 12 + 3*12\cdot 16 = 864$$

$$ATS_3 = 12\cdot 12 + 2*12\cdot 16 = 528$$

3D автокорелационни дескриптори

Междуатомните разстояния се разглеждат в няколко интервала: $[r_1,r_2]$ $[r_2,r_3]$, …, $[r_s,r_{s+1}]$

За всеки интервал се дефинира делта-функция:

$$\delta(r; r_k, r_{k+1}) = \begin{cases} 1 & r \in [r_k, r_{k+1}] \\ 0 & r \notin [r_k, r_{k+1}] \end{cases}$$

делта-фунцкията е 1 когато междуатомното разстояние е в определен интервал т.е. ще се отчетат теглата на двойките атоми с разстояние между тях в дадения интервал

3D автокорелационни дескриптори

$$AC(r_k, r_{k+1}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i \cdot w_j \cdot \delta(r_{ij}, r_k, r_{k+1})$$

AC векторът е уникален профил (fingerprint /‘пръстов-опечатък’) на молекулата и е чувствителен към конформационни промени.

Подходящ е за търсене по подобие.

RDF дескриптори

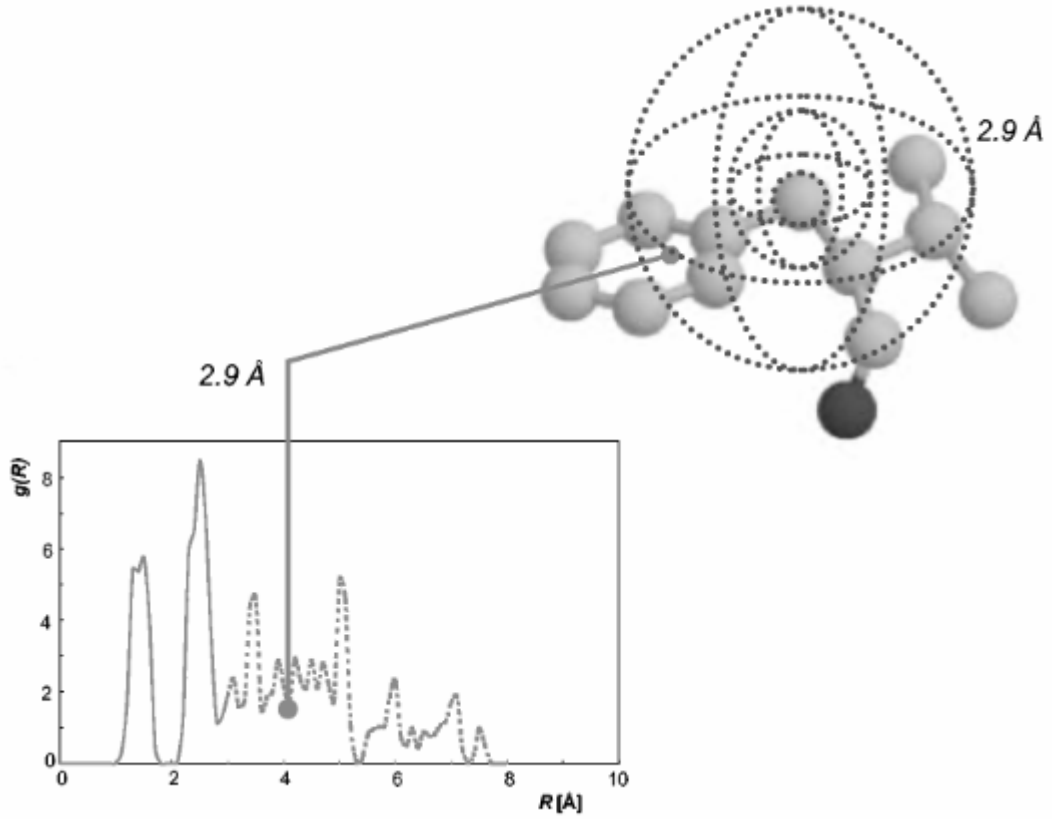
RDF(Radial Distribution Function) – изчислява се информация за молекулата, използвайки радиални функции на разпределение:

$$g(r) = f \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i \cdot w_j \cdot e^{-\beta(r-r_{ij})^2}$$

f е фактор на разпръскване, **β** е изглаждащ параметър, който може да се интерптертира като температурен фактор.

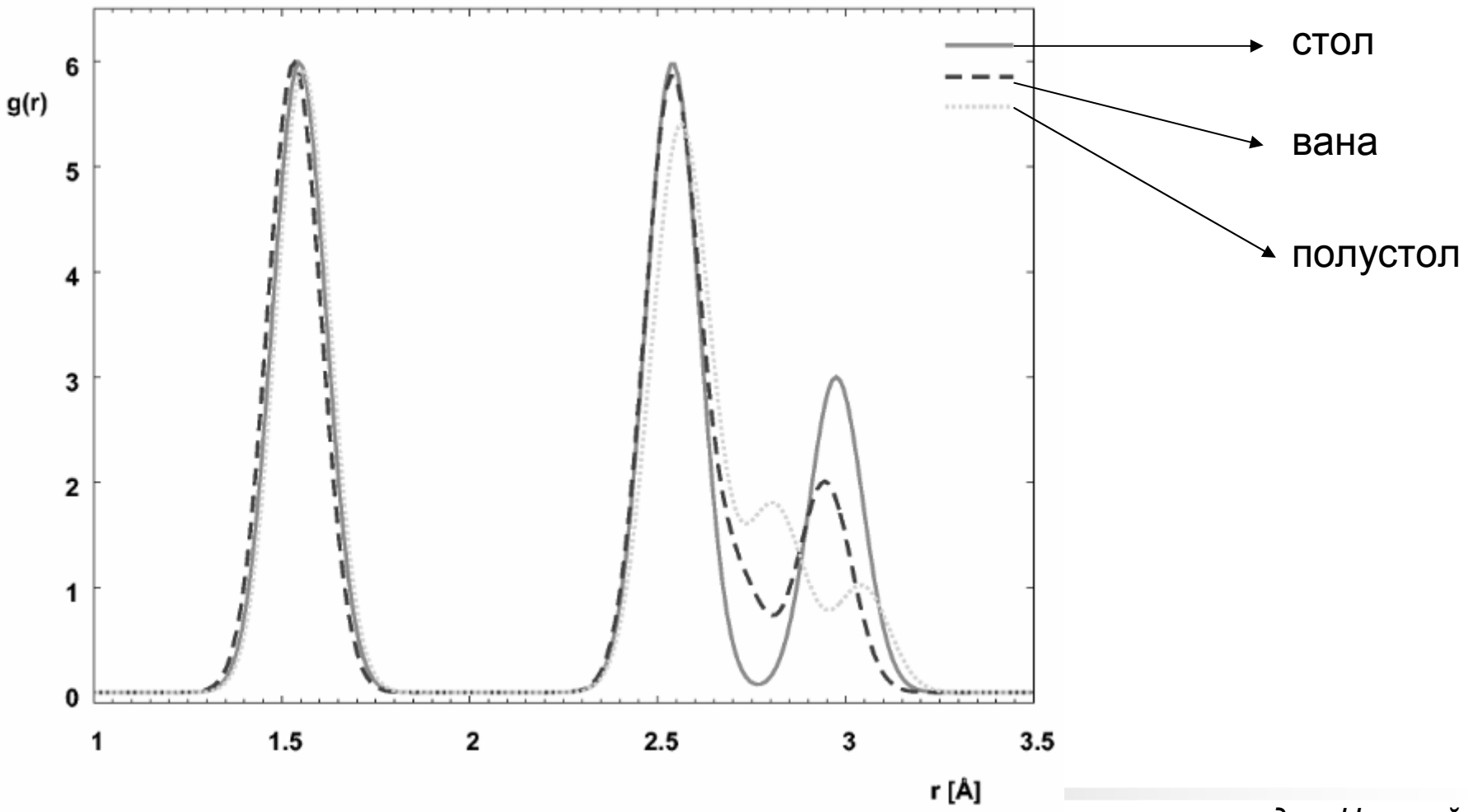
RDF дескриптори

RDF функцията **g(r)** се сканира през определена стъпка **Δr**:



RDF дескриптори

Разграничаване на 3-те конформации на циклохексана



Електростатични дескриптори

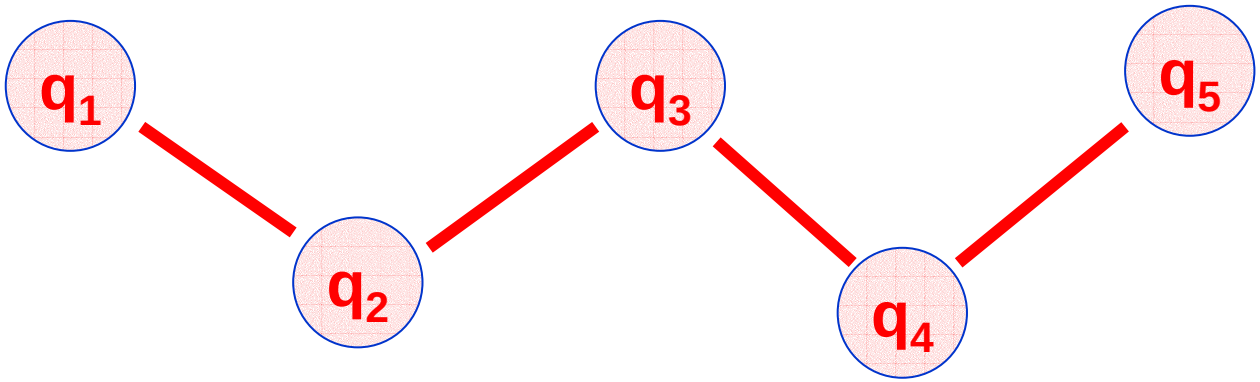
Разпределението на електроните е един от най-важните фактори, който определя физичните, химичните и биологичните свойства на молекулите.

Детайлна информация за разпределението на електроните може да се получи чрез квантово-химични методи.

Няколко различни подхода за бързо изчисляване на разпределението са базирани на частични атомни заряди.

Електростатични дескриптори

В химичнта информатика се използват адитивни подходи, при които се асоциира частичен заряд за всеки атом:



частични заряди на Гащайгер-Марсили

частични заряди на Зефиров

частични заряди на Муликен

Електростатични дескриптори

Най-положителен и най-отрицателен частичен атомен заряд: **Q_{max} и Q_{min}**.

Диполен момент **μ**

Параметри на поляризуемост: **P, P', P"**

$$P = Q_{max} - Q_{min}$$

$$P' = \frac{Q_{max} - Q_{min}}{R_{mm}}$$

$$P'' = \frac{Q_{max} - Q_{min}}{R_{mm}^{2}}$$

R_{mm} е разстоянието между атомите с минималния и максималния заряд

Литература:

1. Handbook of Chemoinformatics: From Data To Knowledge Vol. 1, Ed. J. Gasteiger, WILEY-VCH Verlag GmbH & Co., 2003.

2. Handbook of Chemoinformatics: From Data To Knowledge Vol. 3, Ed. J. Gasteiger, WILEY-VCH Verlag GmbH & Co., 2003.

3. Handbook of Molecular Descriptors; R. Todeschini, V. Consonni; WILEY-VCH Verlag GmbH & Co., 2000.

4. Topological Indices. M. Randich; Encyclopedia of Computational Chemistry, p.3018-3030.

5. CODESSA-Pro Classes of descriptors; http://codessa-pro.com/