

# Kemometrija

QSP(A)R  
"Quantitative Structure Property (Activity) Relationship"

Mag 1.I Kemija, 2019-20

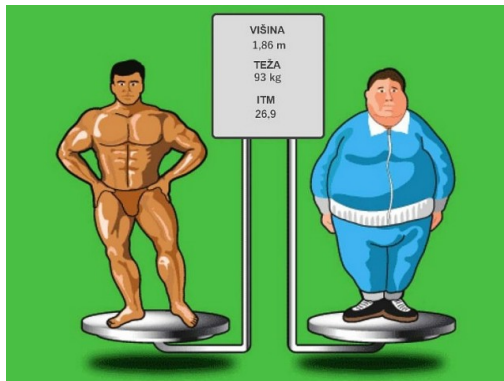
# Indeks telesne mase ITM ("Body Mass Index")

$$\text{ITM} = \frac{\text{telesna masa [kg]}}{\text{telesna višina}^2 [m]}$$



Vir: [dijetamesecevemene.com](http://dijetamesecevemene.com)

## Enak ITM obeh oseb?



# Indeks zamaščenih jeter IZJ ("Fatty Liver Index")

Ne da se ga določiti samo s krvnim testom.


## Določanje IZJ

ITM + obseg pasu + trigliceridi + GGT

(encim)

## Določanje IZJ

ITM + obseg pasu + trigliceridi + GGT  
(encim)



# Določanje IZJ

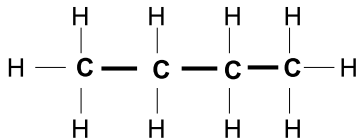
ITM + obseg pasu + trigliceridi + GGT

(encim)

**IZJ**

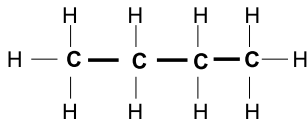


# Modeliranje ogljikovodikov

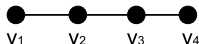


butan

# Modeliranje ogljikovodikov

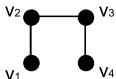
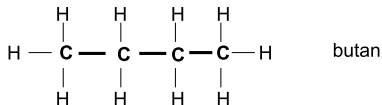


butan



graf

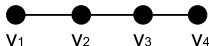
# Modeliranje ogljikovodikov



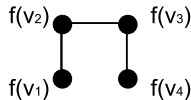
# Graf

Graf  $G$  je določen z množico vozlišč  $V(G)$  in množico povezav  $E(G)$ . Grafa  $G$  in  $H$  sta *izomorfna*, kadar obstaja bijektivna preslikava  $f : V(G) \mapsto V(H)$ , ki ohranja sosednost

$$v_i v_j \in E(G) \Leftrightarrow f(v_i) f(v_j) \in E(H).$$



G



H

Izomorfizem grafa samega vase je *avtomorfizem*.

# Wienerjev indeks

## H. Wiener (1947)

*Vsota razdalj med vsemi pari vozlišč grafa  $G$*

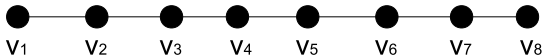
$$W(G) = \frac{1}{2} \sum_{u \in V(G)} \sum_{v \in V(G)} d(u, v)$$

Formula za določanje temperature vrelišča alkanov:

$$T_v(G) = a W(G) + b p + c$$

$p$ -polarno število,  $a$ ,  $b$ ,  $c$ -statistično določeni parametri

# Primer-oktan



$$W(G) = 2(1 \cdot 7 + 2 \cdot 6 + 3 \cdot 5) + 4 \cdot 4 = 84$$

$$p = 5$$

$$T_v(G) = 0.70 \cdot 84 + 5.04 \cdot 5 + 27.18 = 111.18$$

$$T_{eksp} = 125.1$$

## Izboljšani model (Seybold et al. 1987)

$$T_v(G) = a \frac{W(G)}{N^2} + bp + c$$

$$T_v(G) = 124.91$$

# Zaželjene lastnosti molekularnega deskriptorja (M. Randić, 1991)

- razloži strukturo molekule
- dobra korelacija z vsaj eno lastnostjo
- po možnosti loči med izomerami
- uporaben na lokalni strukturi
- posplošitev na kompleksnejše deskriptorje
- neodvisnost
- preprostost



## Zaželjene lastnosti molekularnega deskriptorja (M. Randić, 1991)

- ne temelji na lastnostih
- ni trivialno povezan z drugimi deskriptorji
- učinkovito konstruiranje
- uporablja znane strukturne koncepte
- se sorazmerno spreminja z velikostjo molekule
- odraža postopnost spremembe na strukturi

## Topološki indeksi, ki temeljijo na stopnji vozlišča

- Prvi Zagrebški indeks:

$$ZM_1(G) = \sum_{u \in V(G)} \delta^2(u)$$

- Drugi Zagrebški indeks:

$$ZM_2(G) = \sum_{uv \in E(G)} \delta(u)\delta(v)$$

- Randićev indeks:

$$\chi(G) = \sum_{uv \in E(G)} \frac{1}{\sqrt{\delta(u)\delta(v)}}$$

## Topološki indeksi, ki temeljijo na stopnji vozlišča

- Poglianijev indeks:

$$Dz(G) = \sum_{u \in V(G)} \delta^Z(u)$$

$Z$  = št. valen. elek.: glavno kvantno št.

- ABC ("atom-bond connectivity") indeks:

$$ABC(G) = \sum_{uv \in E(G)} \sqrt{\frac{\delta(u) + \delta(v) - 2}{\delta(u)\delta(v)}}$$

- indeks razvejanosti:

$$Ram(G) = \sum_{v \in V(G), d(v) \geq 3} (d(v) - 2)$$

- ...

# Geometrično-aritmetični indeksi

- splošna oblika

$$GA_{gen}(G) = \sum_{uv \in E(G)} \frac{2\sqrt{Q_u Q_v}}{Q_u + Q_v}$$

$Q_u$  je neka količina, ki jo lahko na enoličen način priredimo vozlišču

## Geometrično-aritmetični indeksi

- splošna oblika

$$GA_{gen}(G) = \sum_{uv \in E(G)} \frac{2\sqrt{Q_u Q_v}}{Q_u + Q_v}$$

$Q_u$  je neka količina, ki jo lahko na enoličen način priredimo vozlišču

- prvi, drugi, tretji, četrti GA indeks
- peti GA indeks:

$$GA_5(G) = \sum_{uv \in E(G)} \frac{2\sqrt{S_u S_v}}{S_u + S_v}$$

$$S_u = \sum_{v \in V(G)} \delta(v)$$

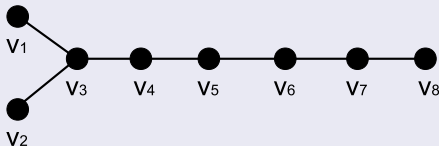
# Graovac-Pisanski indeks oz. modificiran-Wienerjev indeks

## A. Graovac & T. Pisanski (1991)

Meri povprečno razdaljo premika vozlišč z avtomorfizmi

$$GP(G) = \frac{|V(G)|}{2|Aut(G)|} \sum_{u \in V(G)} \sum_{f \in Aut(G)} d(u, f(u)).$$

## Primer

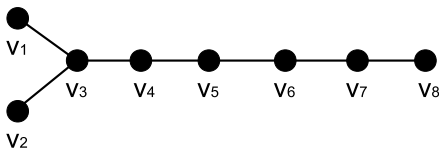


$$\begin{aligned} GP(G) &= \frac{8}{2 \cdot 2} \sum_{u \in V(G)} \sum_{f \in \text{Aut}(G)} d(u, f(u)) \\ &= 2(2 + 2 + 0 + 0 + 0 + 0 + 0 + 0) = 8 \end{aligned}$$

# Orbite grafa

V *orbiti* vozlišča  $u$  so vsa tista vozlišča grafa  $G$ , v katera se  $u$  preslika z nekim avtomorfizmom grafa  $G$

$$\{f(u) \mid f \in \text{Aut}(G)\}.$$



$$V_1 = \{v_1, v_2\},$$

$$V_2 = \{v_3\}, V_3 = \{v_4\}, V_4 = \{v_5\},$$

$$V_5 = \{v_6\}, V_6 = \{v_7\}, V_7 = \{v_8\}.$$



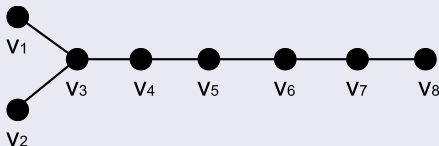
## Izračun Graovac-Pisanski indeksa s pomočjo orbit

### A. Graovac & T. Pisanski (1991)

Če so  $V_1, \dots, V_t$  orbite grafa  $G$ , tedaj je

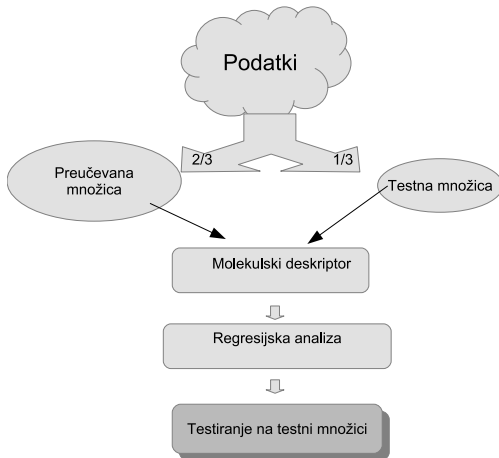
$$GP(G) = |V(G)| \sum_{i=1}^t \frac{1}{|V_i|} W(V_i).$$

## Primer-izračun s pomočjo orbit



$$\begin{aligned}
 GP(G) &= 8 \sum_{i=1}^7 \frac{1}{|V_i|} W(V_i) \\
 &= 8\left(\frac{2}{2} + 0 + 0 + 0 + 0 + 0 + 0\right) = 8
 \end{aligned}$$

# Shema



# "International Academy of Mathematical Chemistry"

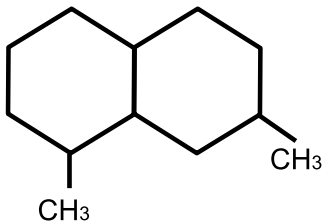
Spletna stran:

▶ <http://www.iamc-online.org>

## Priprava podatkov

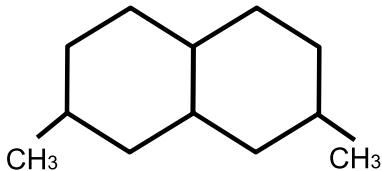
- molekula - "mol" datoteka - matrika sosednosti molekulskega grafa
  - ▶ Cheminfo
  - ▶ Smiles koda
- program (programski jezik) za izračun topološkega indeksa (E-Dragon; Sage, CoCalc; C++, Python,...)

# Graovac-Pisanski indeks dimetil-naftalena



1,7-dimetil-naftalen

$$GP(G_1) = 0$$



2,7-dimetil-naftalen

$$GP(G_2) = 108$$

## Nenavadno obnašanje tališča ogljikovodikov

| molekula                    | tališče °C | vrelišče °C |
|-----------------------------|------------|-------------|
| naftalen                    | 81         | 218         |
| 1-metil-naftalen            | -22        | 245         |
| 2-metil-naftalen            | 35         | 241         |
| 1-etil-naphthalene          | -14        | 259         |
| 2-etil-naphthalene          | -7         | 258         |
| 2-6-dimetil-naftalen        | 110        | 262         |
| <b>2-7-dimetil-naftalen</b> | <b>97</b>  | 262         |
| <b>1-7-dimetil-naftalen</b> | <b>-14</b> | 263         |
| 1-5-dimetil-naftalen        | 80         | 269         |
| 1-2-dimetil-naftalen -      | 4          | 271         |
| 1-3-7-trimetil-naftalen     | 14         | 280         |
| 2-3-5-trimetil-naftalen     | 101        | 286         |
| anthracene                  | 16         | 340         |

Indeksi okoli nas

Topološki indeksi oz. molekularni deskriptorji

QSP(A)R- "Quantitative Structure Property(Activity) Relationship

QSP(A)R pregled

Vaja

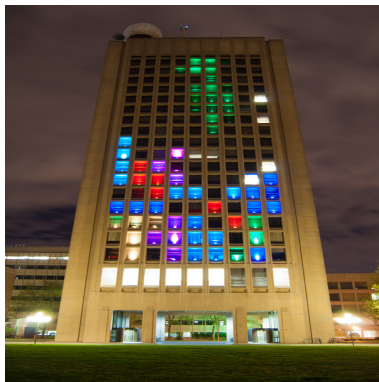
Podatki

Izračun novih indeksov

Simetrija molekul in tališče

Regresijska analiza na primeru alkanov

# Igra Tetris





## Boljše zlaganje "simetričnih" molekul



LAŽJE

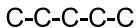


TEŽJE

## Bolše zlaganje molekul $\Rightarrow$ višje tališče



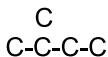
LAŽJE



pentan:  $T_{\text{T}} = -95^{\circ}\text{C}$



TEŽJE



metilbutan:  $T_{\text{T}} = -154^{\circ}\text{C}$

## Zadeva ni tako enostavna...

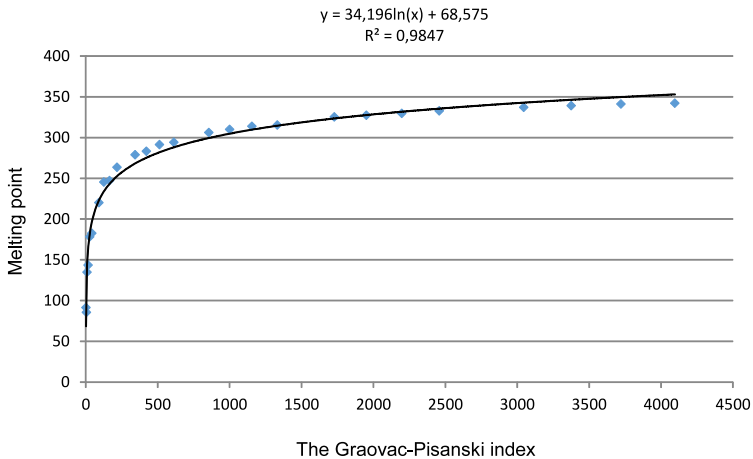
simetrije & sferi podobna oblika molekule  $\Rightarrow$  višje tališče

# Temperatura tališča alkanov

- 26 preučevanih alkanov in 5 alkanov v testni množici
- Nelinearna regresija

| Alkane           | GP index | MP     |
|------------------|----------|--------|
| ethane           | 1        | 91,39  |
| propane          | 3        | 85,45  |
| butane           | 8        | 134,75 |
| pentane          | 15       | 143,35 |
| hexane           | 27       | 178,15 |
| heptane          | 42       | 182,54 |
| <b>octane</b>    | 64       | 216,3  |
| nonane           | 90       | 220,15 |
| decane           | 125      | 245,25 |
| undecane         | 165      | 247,15 |
| dodecane         | 216      | 263,55 |
| <b>tridecane</b> | 273      | 268,15 |
| tetradecane      | 343      | 278,65 |
| pentadecane      | 420      | 283,05 |
| hexadecane       | 512      | 291,15 |

| Alkane            | GP index | MP     |
|-------------------|----------|--------|
| <b>octadecane</b> | 729      | 302,15 |
| nonadecane        | 855      | 306,15 |
| icosane           | 1000     | 309,85 |
| henicosane        | 1155     | 313,65 |
| docosane          | 1331     | 315,15 |
| <b>tricosane</b>  | 1518     | 322,15 |
| tetracosane       | 1728     | 325,15 |
| pentacosane       | 1950     | 327,15 |
| hexacosane        | 2197     | 329,55 |
| heptacosane       | 2457     | 332,65 |
| <b>octacosane</b> | 2744     | 337,65 |
| nonacosane        | 3045     | 336,85 |
| triacontane       | 3375     | 338,95 |
| hentriacontane    | 3720     | 341,05 |
| dotriacontane     | 4096     | 342,15 |



## Testna množica

| Alkan     | $GP$ index | $T_{eks}$ | $T_{izr}$ | Napaka | % Napaka |
|-----------|------------|-----------|-----------|--------|----------|
| $C = 8$   | 64         | 216,3     | 210,792   | 5,508  | 2,546    |
| $C = 13$  | 273        | 268,15    | 260,396   | 7,754  | 2,891    |
| $C = 18$  | 729        | 302,15    | 293,984   | 8,166  | 2,703    |
| $C = 23$  | 1518       | 322,15    | 319,066   | 3,084  | 0,957    |
| $C = 28$  | 2744       | 337,65    | 339,311   | -1,661 | 0,492    |
| povprečje |            |           |           |        | 1,918    |

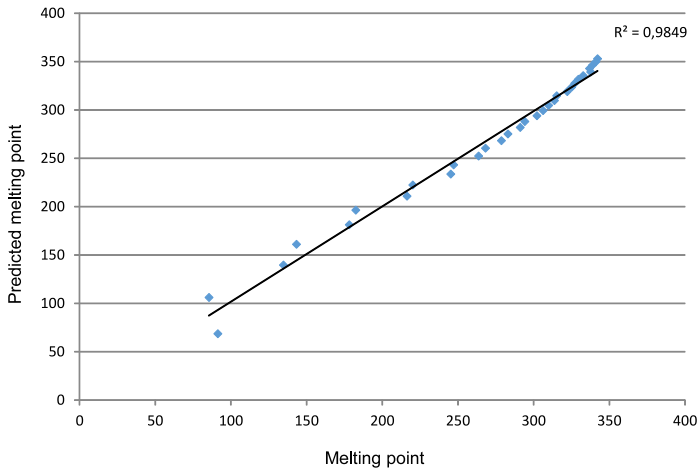


| Alkane      | GP index | MP     | $\widehat{MP}$ | Residual | % Residual |
|-------------|----------|--------|----------------|----------|------------|
| ethane      | 1        | 91,39  | 68,575         | 22,815   | 24,964     |
| propane     | 3        | 85,45  | 106,143        | -20,693  | 24,217     |
| butane      | 8        | 134,75 | 139,684        | -4,935   | 3,661      |
| pentane     | 15       | 143,35 | 161,179        | -17,829  | 12,438     |
| hexane      | 27       | 178,15 | 181,279        | -3,129   | 1,757      |
| heptane     | 42       | 182,54 | 196,388        | -13,848  | 7,586      |
| octane      | 64       | 216,3  | 210,792        | 5,508    | 2,546      |
| nonane      | 90       | 220,15 | 222,450        | -2,300   | 1,045      |
| decane      | 125      | 245,25 | 233,684        | 11,566   | 4,716      |
| undecane    | 165      | 247,15 | 243,178        | 3,972    | 1,607      |
| dodecane    | 216      | 263,55 | 252,388        | 11,162   | 4,235      |
| tridecane   | 273      | 268,15 | 260,396        | 7,754    | 2,891      |
| tetradecane | 343      | 278,65 | 268,202        | 10,448   | 3,749      |

| Alkane      | GP index | MP     | $\widehat{MP}$ | Residual | % Residual |
|-------------|----------|--------|----------------|----------|------------|
| hexadecane  | 512      | 291,15 | 281,901        | 9,249    | 3,177      |
| heptadecane | 612      | 294,15 | 288,002        | 6,148    | 2,090      |
| octadecane  | 729      | 302,15 | 293,984        | 8,166    | 2,703      |
| nonadecane  | 855      | 306,15 | 299,436        | 6,714    | 2,193      |
| icosane     | 1000     | 309,85 | 304,793        | 5,057    | 1,632      |
| hencosane   | 1155     | 313,65 | 309,720        | 3,930    | 1,253      |
| docosane    | 1331     | 315,15 | 314,570        | 0,580    | 0,184      |
| tricosane   | 1518     | 322,15 | 319,066        | 3,084    | 0,957      |
| tetracosane | 1728     | 325,15 | 323,497        | 1,653    | 0,509      |
| pentacosane | 1950     | 327,15 | 327,630        | -0,480   | 0,147      |
| hexacosane  | 2197     | 329,55 | 331,708        | -2,158   | 0,655      |
| heptacosane | 2457     | 332,65 | 335,533        | -2,883   | 0,867      |
| octacosane  | 2744     | 337,65 | 339,311        | -1,661   | 0,492      |

| Alkane         | <i>GP</i> index | <i>MP</i> | $\widehat{MP}$ | Residual | % Residual |
|----------------|-----------------|-----------|----------------|----------|------------|
| triacontane    | 3375            | 338,95    | 346,388        | -7,438   | 2,195      |
| hentriacontane | 3720            | 341,05    | 349,717        | -8,667   | 2,541      |
| dotriacontane  | 4096            | 342,15    | 353,009        | -10,859  | 3,174      |
| average        |                 |           |                |          | 4,025      |

## Korelacija med $T_{eks}$ in $T_{izr}$



▶ QSPR pregled

▶ Fenoli in disociacijska konstanta

S spletne strani <https://www.fkkt.um.si/ukemat/MatMag.php> naloži datoteko PAH\_Vaja.xlsx s podatki.

- 1 Poišči linearno regresijo med ABC indeksom in GA5 indeksom tako, da sestaviš preučevano množico (70 podatkov) in testno množico (12 podatkov). Na testni množici preveri dobljeno linearno regresijo in izračunaj povprečno napako (v %).
- 2 Izloči 3 osamelce in ponovi postopek iz prejšnje točke (preučevana množica 67 podatkov). Na testni množici preveri dobljeno linearno regresijo in izračunaj povprečno napako (v %).
- 3 Izvedi multilinerano regresijo odvisnosti ABC indeksa od GA5 in W indeksa ob enakih zahtevah kot v 1. točki.

## Literatura dostopna na

▶ Matematika na FKKT