Molecular descriptors based on automorphism data

Kurt Varmuza* [1,2], Matthias Dehmer [3,4], Peter Filzmoser [1]

[1] Vienna University of Technology, Institute of Statistics and Mathematical Methods in Economics | Computational Statistics
[2] Vienna University of Technology, Institute of Chemical, Environmental and Bioscience Engineering | Sustainable Techn. & Process Simulation
[3] UMIT Tirol - Private University for Health Sciences and Health Technology, Hall in Tyrol, Austria
[4] Swiss Distance University of Applied Sciences, Department of Computer Science, Brig, Switzerland

Basic definitions / Overview

Chemical structures are represented by mathematical **graphs** [1,2]. Graph vertices: atoms; graph edges: bonds; hydrogen-depleted; connected coloured graphs; no atomic charge info; no 3D info.

Topologically (constitutionally) **equivalent atoms (bonds)** have identical neighborhoods in terms of connectivity as described by the graph - considering the whole graph (molecular structure).

The **automorphism group** of a graph describes all mappings of the graph onto itself - preserving the connectivity (not cutting bonds). It contains data about **topologically equivalent atoms (bonds)**. Asymmetric structures: only a single (trivial) mapping exists; highly symmetric structures: several (many) mappings.

Size of the automorphism group: number of possible mappings of a graph onto itself; it is a **symmetry measure** for the graph.

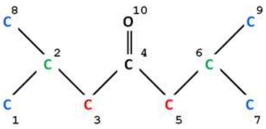
Orbits (in graph theory)
Atom (vertex) orbit: set of topologically equivalent atoms
Bond (edge) orbit: set of topologically equivalent bonds

Molecular descriptors [3] can be are derived from the size of the automorphism group and from the frequencies of the different orbit sizes [4]. They are evaluated here together with other descriptors [5] for multivariate QSPR models: quantitative structure property relationships for the prediction of melting points.

Continues a **similar study with exhaustive sets of alkanes** [6].

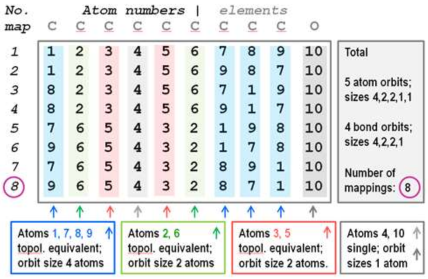
Automorphism with a demo structure

4-heptanone, 2,6-dimethyl-
$C_9H_{18}O$
H-depleted: 10 atoms (vertices)
9 bonds (edges)



Software. SubMat [6,7],
functions in R [8]

Automorphism mappings



No. map	C	C	C	C	C	C	C	C	C	O
1	1	2	3	4	5	6	7	8	9	10
2	1	2	3	4	5	6	9	8	7	10
3	8	2	3	4	5	6	7	1	9	10
4	8	2	3	4	5	6	9	1	7	10
5	7	6	5	4	3	2	1	9	8	10
6	9	6	5	4	3	2	1	7	8	10
7	7	6	5	4	3	2	8	9	1	10
8	9	6	5	4	3	2	8	7	1	10

Atom numbers | elements

Total

5 atom orbits; sizes 4,2,2,1,1

4 bond orbits; sizes 4,2,2,1

Number of mappings: 8

Atoms 1, 7, 8, 9 topol. equivalent; orbit size 4 atoms

Atoms 2, 6 topol. equivalent; orbit size 2 atoms

Atoms 3, 5 topol. equivalent; orbit size 2 atoms.

Atoms 4, 10 single; orbit sizes 1 atom

Examples of molecular descriptors based on automorphism data

- **Number of orbits**; separately for atom types (C, N, O) and bond types (single, double, triple, aromatic).
- **Size of automorphism group** (number of mappings); absolute, logarithm, normalized by number of atoms, bonds.
- Number of asymmetric C-atoms.
- Based on **frequencies of orbits with sizes 1, 2, 3**, …; e. g., entropy, symmetry index [3,9], root of orbit polynomial [4]; separately for atom and bond types.

Application example [QSPR model for melting point]

Data

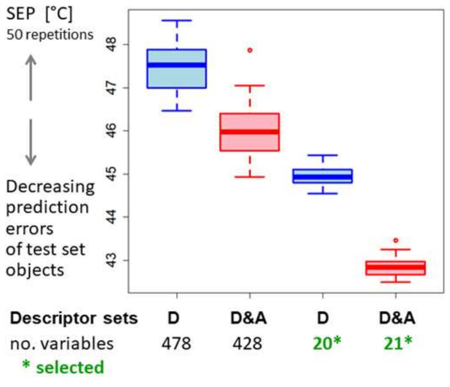
Origin: Tetko et al., 2014 [10]; Bradley J.C., 2014 [11]
Selected:
n = 1161 chemical compounds ($C_{5\text{-}30}$, H, N, O only),
$\boldsymbol{y}$, **melting point** 20 – 300 °C; sd(y) = 61
$\boldsymbol{X}$, m = 478 **molecular descriptors**;
m_1 = 428 (**D**ragon) [5]; m_2 = 50 (**A**utomorphism);
autoscaled.
$\boldsymbol{X}_{\text{SEL}}$, m = 21, **selected stepwise** with BIC criterion [12];
$m_{1,\text{SEL}}$ = 15 (**D**ragon); $m_{2,\text{SEL}}$ = 6 (**A**utomorphism)

Multivariate models

Linear regression with **PLS**; strategy **repeated double cross validaton (rdCV)** separates optimization of model complextity (no. of PLS components) and estimation of prediction performance [13, 14].

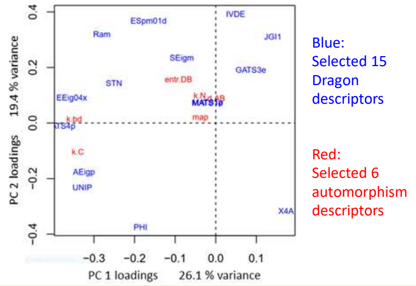
Performance criterion: **SEP = standard deviation of prediction errors for test set objects** (boxplot for 50 repetitions).

Comparison of descriptors sets without/with automorphism descriptors



Tentative conclusion. Perhaps a useful complement to other descriptors. More tests required.

PCA loading plot of selected descriptors

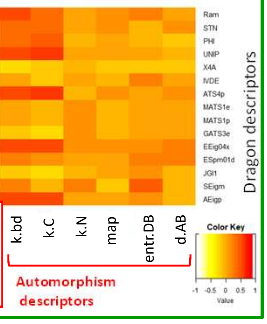


Blue: Selected 15 Dragon descriptors

Red: Selected 6 automorphism descriptors

Heat map for correlation coefficients

between selected Dragon and automorphism descriptors



map, size of automorphism group
entr.DB, entropy, orbits with doubl bd.
d.AB, root of orbit polynomial (orbits with aromatic bonds)
k.bd, k.C, k.N, no. orbits (bonds, C, N)

[1] Engel T., Gasteiger J. (Eds.): *Chemoinformatics - Basic concepts and methods*, Wiley VCH, Weinheim, Germany, **2018**
[2] Trinajstic N.: *Chemical graph theory,* CRC Press, Boca Raton, FL, USA, **1992**
[3] Todeschini R. et al.: *Molecular descriptors for chemoinformatics*, Wiley-VCH, Weinheim, **2009**
[4] Dehmer M. et al.: *IEEE Access* **2020**, *8*, 36100
[5] Dragon, Software for molecular descriptor calculation, vers. 6.0 (**2010**), www.talete.mi.it; https://chm.kode-solutions.net/
[6] Varmuza K. et al.: *Croatica Chemica Acta*, **2021**, *94*, 47
[7] Varmuza K. et al.: *Croatica Chemica Acta*, **2005**, *78*, 141
[8] R, A language and environment for statistical computing, http://www.r-project.org, Vienna, Austria, **2023**
[9] Mowshowitz A., Dehmer M.: *Symmetry: Culture and Science*, 21 (**2010**) 321
[10] Tetko I.V. et al.: *J. Chem. Inf. Model*. **2014**, *54*, 3320
[11] Bradley J.C.: https://figshare.com/articles/dataset/Jean_Claude_Bradley_Double_Plus_Good_Highly_Curated_and_Validated_Melting_Point_Dataset/1031638 (**2014**)
[12] Filzmoser P., Varmuza K.: *R package chemometrics*, Vienna, Austria, **2010**, http://cran.at.r-project.org/web/packages/chemometrics/index.html
[13] Varmuza K., Filzmoser P.: Introduction to multivariate statistical analysis in chemometrics, CRC Press, Boca Raton, FL, USA, **2009**
[14] Filzmoser P. et al.: *J. Chemometrics*, *23*, 160 (**2009**)

240521

Poster at **Math/Chem/Comp 2024**, 3-7 June 2024, Dubrovnik, Croatia