Molecular Descriptors
Theory and tips for real-world applications

Francesca Grisoni
University of Milano-Bicocca, Dept. of Earth and Environmental Sciences, Milan, Italy
ETH Zurich, Dept. of Chemistry and Applied Biosciences, Zurich, Switzerland
francesca.grisoni@unimib.it
Presentation Outline

• Introduction
• Molecular representation and Molecular description
• Classical vs Fingerprint approach
• Tips and tricks
“It is obvious that there must exist a relation between the chemical constitution and the physiological action of a substance […], but as yet scarcely any attempts have been made to discover what this relation is. […] it might be supposed that a careful examination and comparison of known facts would lead to the discovery of some empirical law by means of which we could deduce the action from the chemical constitution.”

“... the final result of a logical and mathematical procedure that transforms chemical information of a molecule, such as structural features, into useful numbers or the result of standardized experiments.”

\[ P = f \left( \begin{array}{c}
\end{array} \right) \]

\[ P = f \left( 0.1, 1, 0, 3, 3.5, 100, 2, ... \right) + \varepsilon \]

“... the final result of a logical and mathematical procedure that transforms chemical information of a molecule, such as structural features, into useful numbers or the result of standardized experiments.”

- Atom counts
- Molecular weight
- Atomic properties

- Fragment counts
- Fragment presence

- Topo-structural
- Topo-chemical

- Geometrical
- Atomic coordinates

- Grid-based
- Ensemble-based
“Make things as simple as possible, but not simpler.”
Classical MDs

Fingerprints

VS

1

\[
\begin{array}{c}
\text{100010} \\
\vdots \\
01100
\end{array}
\]

\( n \)
Molecular Weight

\[
\text{MW} \geq 206.31
\]

\[
\begin{align*}
\text{C} & \quad 12.01 \times 13 + \\
\text{H} & \quad 1.01 \times 18 + \\
\text{O} & \quad 16.00 \times 2 = \\
\end{align*}
\]
Matrix-based descriptors

\[ \lambda_{\text{MAX}} = 11.21 \]
GEometry, Topology, and Atom-Weights Assembly (GETAWAY)

\[ H = M \times (M^T \times M)^{-1} \cdot M^T \]

\[ HIC = -\sum_{i=1}^{nAT} \frac{h_i}{D} \cdot \log_2 \frac{h_i}{D} \]
Binary Fingerprints
Extended Connectivity FP

Radius = 0

Radius = 1

Radius = 2

https://docs.chemaxon.com/display/docs/Extended+Connectivity+Fingerprint+ECFP
Path FP

- Path length = 0
- Path length = 1
- Path length = 2
- Path length = 3

https://docs.eyesopen.com/toolkits/python/graphsimtk/fingerprint.html
FP settings

- Radius/path length
- Number of bits
- FP length

- Molecular Information
- Bit collision
- Darkness

Darkness (av. 40-50%, max 80%)
Chemically Advanced Template Search (CATS)

Which approach?

Classical MDs

- Amount of encoded information
- Interpretability
- Require pre-treatment

Fingerprints

- Quick similarity calculations
- No need for pre-treatment
- Modelling approaches for binary data
Structure Activity Landscapes

- **Gently rolling hills**
- **Rugged landscapes**

\[ SALI_{i,j} = \frac{|A_i - A_j|}{1 - \text{sim}(i, j)} \]
Tips and tricks

"If this works, it’ll change everything. We could open a casino.”
Tips and tricks

0. Attention to structure representation

SMILES: CC1=CC=CC=C1  \( nBM = 3 \)

SMILES: Cc1ccccc1  \( nBM = 6 \)
Tips and tricks

1. Know your purpose

Tips and tricks

2. Reduce Dimensionality (if possible)

---

2.1. Reduce Dimensionality (if possible)
Tips and tricks

2. Reduce Dimensionality (if possible)

PCA = Principal Component Analysis

Tips and tricks

2. Reduce Dimensionality (if possible)

K-means clustering

- Assign variables randomly to a set of $k$ clusters
- Compute cluster centroids
- Re-assign variables to the cluster with the closest centroid
Tips and tricks

3. Mind the measuring unit

- **Auto-scaling** (Gaussian normalization)
  \[ x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \]

- **Range-scaling** (minMax normalization)
  \[ x'_{ij} = \frac{x_{ij} - \min_j}{\text{Max}_j - \min_j} \]

## Tips and tricks

### 4. Consider other similarity measures

<table>
<thead>
<tr>
<th>Distance</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>$D_{xy}^{EUC} = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$</td>
</tr>
<tr>
<td>Manhattan or city-block</td>
<td>$D_{xy}^{MAN} = \sum_{j=1}^{p}</td>
</tr>
<tr>
<td>Lagrange</td>
<td>$D_{xy}^{LAG} = \max_j</td>
</tr>
<tr>
<td>Minkowski</td>
<td>$D_{xy}^{MIN} = \left[\sum_{j=1}^{p}</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>$D_{xy}^{MAH} = \sqrt{(x - y)^T \cdot S^{-1} \cdot (x - y)}$</td>
</tr>
</tbody>
</table>

## 4. Consider other similarity measures

<table>
<thead>
<tr>
<th>Similarity coefficient</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sokal–Michener, Simple Matching</td>
<td>$S_{xy}^{SM} = \frac{a+d}{p}$</td>
</tr>
<tr>
<td>Rogers–Tanimoto</td>
<td>$S_{xy}^{RT} = \frac{a+d}{p+b+c}$</td>
</tr>
<tr>
<td>Jaccard–Tanimoto</td>
<td>$S_{xy}^{JT} = \frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>Gleason–Dice–Sorensen</td>
<td>$S_{xy}^{GDL} = \frac{2a}{2a+b+c}$</td>
</tr>
<tr>
<td>Russell–Rao</td>
<td>$S_{xy}^{RR} = \frac{a}{p}$</td>
</tr>
<tr>
<td>Forbes</td>
<td>$S_{xy}^{FOR} = \frac{pa}{(a+b)(a+c)}$</td>
</tr>
<tr>
<td>Simpson</td>
<td>$S_{xy}^{SIM} = \frac{a}{\min{a+b,a+c}}$</td>
</tr>
<tr>
<td>Braun-Blanquet</td>
<td>$S_{xy}^{BB} = \frac{a}{\max{a+b,a+c}}$</td>
</tr>
<tr>
<td>Driver–Kroeber–Ochiai cosine</td>
<td>$S_{xy}^{DK} = \frac{a}{\sqrt{(a+b)(a+c)}}$</td>
</tr>
<tr>
<td>Baroni-Urbani–Buser</td>
<td>$S_{xy}^{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad+a+b+c}}$</td>
</tr>
<tr>
<td>Kulczynski</td>
<td>$S_{xy}^{KUL} = \frac{1}{2} \cdot \left[ \frac{a}{a+b} + \frac{a}{a+c} \right]$</td>
</tr>
</tbody>
</table>

### 4. Consider other similarity measures

<table>
<thead>
<tr>
<th>Similarity coefficient</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sokal–Michener, Simple Matching</td>
<td>$S_{xy}^{SM} = \frac{a+b}{a+b+c}$</td>
</tr>
<tr>
<td>Rogers–Tanimoto</td>
<td>$S_{xy}^{RT} = \frac{a+b}{a+b+c}$</td>
</tr>
<tr>
<td>Jaccard–Tanimoto</td>
<td>$S_{xy}^{RT} = \frac{a}{a+b+c}$</td>
</tr>
<tr>
<td>Gleason–Dice–Sorensen</td>
<td>$S_{xy}^{GLE} = \frac{2a}{2a+b+c}$</td>
</tr>
<tr>
<td>Russell–Rao</td>
<td>$S_{xy}^{RR} = \frac{a}{p}$</td>
</tr>
<tr>
<td>Forbes</td>
<td>$S_{xy}^{FOR} = \frac{pa}{(a+b)(a+c)}$</td>
</tr>
<tr>
<td>Simpson</td>
<td>$S_{xy}^{SIM} = \frac{a}{\min((a+b),(a+c))}$</td>
</tr>
<tr>
<td>Braun-Blanquet</td>
<td>$S_{xy}^{BB} = \frac{a}{\max((a+b),(a+c))}$</td>
</tr>
<tr>
<td>Driver–Kroebcr–Ochiai cosine</td>
<td>$S_{xy}^{DK} = \frac{a}{\sqrt{a+b(a+c)}}$</td>
</tr>
<tr>
<td>Baroni-Urbani–Buser</td>
<td>$S_{xy}^{BU1} = \frac{\sqrt{ad}+a}{\sqrt{ad+a+b+c}}$</td>
</tr>
<tr>
<td>Kuczynski</td>
<td>$S_{xy}^{KUL} = \frac{1}{2} \left( \frac{a}{a+b} + \frac{a}{a+c} \right)$</td>
</tr>
</tbody>
</table>

• Descriptors are numbers that capture particular molecular features
• The best descriptors set depends on the problem
• Different types of descriptors require different type of pre-treatment
• Molecular similarity is not an absolute concept
Additional Reading

- **Molecular descriptor theory**

- **Tutorial on descriptors processing and use**

- **Automated data pre-processing**
# Software (some examples)

<table>
<thead>
<tr>
<th>Software</th>
<th>No. descr.</th>
<th>Description</th>
<th>Free</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMEWORKS ModelBuilder</td>
<td>≈ 400</td>
<td>Physicochemical, topological, geometrical, and electronic properties derived from the molecular structure</td>
<td></td>
</tr>
<tr>
<td>BlueDesc</td>
<td>174</td>
<td>Descriptors from JOELib2 and CDK sources, works only with 3D structures.</td>
<td></td>
</tr>
<tr>
<td>CODESSA</td>
<td>≈ 1,500</td>
<td>Constitutional, topological, geometrical, charge-related, quantum-chemical and thermodynamic descriptors.</td>
<td></td>
</tr>
<tr>
<td>Dragon</td>
<td>&gt; 5,200</td>
<td>Benchmark software for calculating 0- to 3D descriptors and binary fingerprints.</td>
<td></td>
</tr>
<tr>
<td>E-Dragon</td>
<td>&gt; 3,000</td>
<td>Free, electronic remote version of DRAGON.</td>
<td>yes</td>
</tr>
<tr>
<td>MOE - Molecular Operating Environment</td>
<td>≈ 300</td>
<td>Topological indices, structural keys, E-state indices, physical properties.</td>
<td></td>
</tr>
<tr>
<td>PaDel</td>
<td>&gt; 1,875</td>
<td>Open source. Based on CDK with additional 2D and 3D descriptors.</td>
<td>yes</td>
</tr>
<tr>
<td>ISIDA Fragmentor</td>
<td>/</td>
<td>Molecular fragments from a Structure-Data File (SDF).</td>
<td></td>
</tr>
</tbody>
</table>