Generative Topographic Mapping: universal tool for chemical space analysis

Alexandre Varnek
*University of Strasbourg*

*BigChem school, Modena, 25 October 2017*
Chemical universe: Big Data problem

- $10^8$ compounds are currently available
- $10^{33}$ drug-like molecules could be synthesized *

How to represent this huge chemical space and to navigate in this space?

Encoding chemical structures by molecular descriptors

Molecular graph

Descriptors

Descriptor vector

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁</td>
<td>a₁</td>
</tr>
<tr>
<td>D₂</td>
<td>a₂</td>
</tr>
<tr>
<td>....</td>
<td>...</td>
</tr>
<tr>
<td>Dᵢ</td>
<td>aᵢ</td>
</tr>
<tr>
<td>....</td>
<td>...</td>
</tr>
</tbody>
</table>

> 5000 types of descriptors are used
Chemography: cartography of chemical space

Data visualization =>
dimensionality reduction problem

Data space
(N-dimensional)

Latent space
(2-dimensional)
Dimensionality reduction requirements

- minimal information loss,
- topology preservation,
- distance preservation
Dimensionality Reduction: information loss

Greenland 2.2 M km²
Arabian Peninsula 3.5 M km²
Australia 7.7 M km²
Taxonomy of Dimensionality Reduction Techniques

Dimensionality Reduction

Linear
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)
  - Sammon Mapping (SM)
  - Isomap

Nonlinear
- Distance Preserving Methods
  - Kernel Principal Component Analysis (KPCA)
- Topology Preserving Methods
  - Isomap
- Predefined Lattice
  - Self-Organizing Maps (SOMs)
  - Generative Topographic Mapping (GTM)
- Data Driven
  - Locally Linear Embedding (LLE)
  - Isotop
Dimensionality reduction methods

Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives

- Multi-Dimensional Scaling
- Canonical Correlation Analysis
- Independent Component Analysis
- Exploratory Factor Analysis
- Sammon map
- PCA
- Kernel PCA (RBF kernel)
- Kernel PCA (polynomial kernel)
- Isomap
- Locally Linear Embedding
- Laplacian Eigenmaps
- t-SNE
- Autoencoder dimensionality reduction
- SOM
- GTM
Generative Topographic Mapping approach
Teuvo Kohonen

- Nonlinear unsupervised approach
- Simple interpretation
- Topology preservation
- Can be used for classification purposes
Limitations of SOMs

- the absence of a cost function to be optimized in training;
- the lack of a theoretical basis for choosing methods parameters;
- the absence of any general proofs of convergence;
- models do not define a probability density

GTM overcomes most of limitations of SOMs without introducing disadvantages.
Generative Topographic Mapping: algorithm

Insertion of a manifold

Projection of data points onto the manifold

Unbending manifold

2D latent space
GTM generates a data probability distribution in both initial and latent data spaces.
This opens an opportunity to use GTM not only to visualize the data but also for structure-property modeling tasks.

- C. M. Bishop *Pattern Recognition and Machine Learning*, 2006 Springer
Projection of an object on GTM is described by the probability distribution (responsibilities) over the lattice nodes.

Probabilities (responsibilities) of acetylcholinesterase ligand projected into latent space
GTM descriptors for molecules and datasets

Map resolution: $N_{nodes} = K^2$

Standard setting: $K = 25$, $N_{grid} = 625$

**Molecule** ➞ responsibilities’ vector $\{R_{tk}\}$ of $N_{nodes}$ length

**Dataset** ➞ normalized cumulated responsibilities’ vector of $N_{nodes}$ length
GTM: areas of application

- Data visualisation and analysis
- Libraries comparison
- Structure-Activity modeling
- Extraction of privilege patterns

Conformational space analysis
Virtual screening
Library design
Inverse QSAR
Chemical data analysis and activity prediction
Properties mapping

- Political map
- Physical map
- CO₂ emission
- Population density
GTM property (activity) landscape

\[ \bar{A}_k = \frac{\sum_i A_i R_{ik}}{\sum_i R_{ik}} \]

- Activity value assigned to a node \( k \)

Expectation of activity \( \bar{A}_k \) in node \( k \)

**Dataset** $\longrightarrow$ property landscape vector \( \{\bar{A}_k\} \) of \( N_{\text{nodes}} \) length
Stability of Lu$^{3+}$ complexes with organic molecules

Activity landscape for Lu\(^{3+}\) complexation

Responsibilities’ distribution of a test set cmpd

Predictions on a test set

$$\hat{A}_j \text{(test)} = \sum_k \bar{A}_k \text{(training)} R_{jk} \text{(test)}$$
Performance of GTM-based QSAR models

Regression models for LogK (Lu$^{3+}$) using ISIDA descriptors

GTM-based models perform similarly to those obtained with popular machine-learning methods

Class landscape can be used to predict a class ("active"/"inactive") for any new compound.

Class landscape for antiviral activity

Responsibilities’ distribution of a test set cmpd

Probabilities to be “active” or “inactive”) for a new compound is estimated using the Bayes equation

\[ P(c_i | x_k) = \frac{P(x_k | c_i) \times P(c_i)}{\sum_i P(x_k | c_i) \times P(c_i)} \]

Toward « universal » map of chemical space
Map of the chemical space should:

- be representative with respect to the variety of known chemotypes;
- be able to distinguish different activity classes and different chemotypes;
- be able to accommodate novel structures and activities in agreement with the neighborhood behavior principle.
As any machine-learning method, GTM is limited by the data size. For very large data sets, only part of molecules can be used for the manifold construction. Thus, a representative subset (a “Frame” Set) must be selected for this purpose.

Examples of Frame Sets

Non-representative

Representative

Hecataeus of Miletus (c. 550 – 476 BCE)

contour lines world map
Choice of descriptors – a vital issue for chemical space construction

- Positioning of objects in the initial and latent spaces depends on the choice of molecular descriptors
- Is there a way to select some « optimal » descriptors for maps construction?
DUD dataset: *GTM as a function of type of descriptors*

Γ-score – measures the ability of a model to produce clustering of similar structures in latent space. For ideal classes separation, $\Gamma = 1$. 
Optimal descriptors are supposed to provide with the best GTM-based regression or classification models built on some « scoring » dataset(s).
ISIDA fragment descriptors

Sequences containing $2 < N < 15$ atoms

Augmented Atoms: selected atoms with their closest neighbours
Several hundreds types of fragment descriptors can be generated for one same data set
Toward universal map(s) of chemical space

More activities are used, more universal is a map

<table>
<thead>
<tr>
<th>Activity</th>
<th>Model’s performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Act₁</td>
<td>BA₁</td>
</tr>
<tr>
<td>Act₂</td>
<td>BA₂</td>
</tr>
<tr>
<td>......</td>
<td></td>
</tr>
<tr>
<td>SCORE</td>
<td>&lt;BA&gt;</td>
</tr>
</tbody>
</table>
Toward universal map(s) of chemical space

<table>
<thead>
<tr>
<th>Descriptors type</th>
<th>MAP</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descr₁</td>
<td>Map₁</td>
<td>SCORE₁</td>
</tr>
<tr>
<td>Descr₂</td>
<td>Map₂</td>
<td>SCORE₂</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
<td>......</td>
</tr>
<tr>
<td>Descrₙ</td>
<td>Mapₙ</td>
<td>SCOREₙ</td>
</tr>
</tbody>
</table>

Decraptors leading to largest score are selected

**Basic assumption:**
If the manifold is trained on activity predictions of ligands for >100 different biological targets, it may also accommodate other biological activities and compounds.
Chemical space of druglike compounds

Universal manifold:
- Optimized for 144 sets (GPCR, kinases, ...)
- Validated on > 450 sets
ChEMBL: chemical space of antiviral compounds

3 “universal” maps based on different types of ISIDA descriptors

Privilege patterns extraction
From responsibility vector to binned pattern

Responsibility Vector

<table>
<thead>
<tr>
<th>Node Number ( n )</th>
<th>...</th>
<th>183</th>
<th>184</th>
<th>185</th>
<th>186</th>
<th>187</th>
<th>188</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of Residence ( R_n )</td>
<td>0.0</td>
<td>0.005</td>
<td>0.020</td>
<td>0.022</td>
<td>0.031</td>
<td>0.021</td>
<td>0.002</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node Number ( n )</th>
<th>...</th>
<th>183</th>
<th>184</th>
<th>185</th>
<th>186</th>
<th>187</th>
<th>188</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binned “Pattern” vector</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
3 different maps are needed in order to extract the privilege patterns for main classes of antivirals.

3 “universal” maps based on different types of ISIDA descriptors.
GTM: extraction of privilege structural patterns

ChEMBL: class landscape for GPCR ligands

Evolution of privileged structural motifs of GPCR ligands in ChEMBL

Big Data challenge
**GDB-17** – computer-generated virtual molecules containing up to 17 heavy atoms *

- whole set: $1.66 \times 10^{11}$ virtual molecules
- « lead-like » subset used in this study: **10 M** molecules

**PubChem-17** – subset of **10.8 M** real molecules containing up to 17 heavy atoms extracted from the PubChem database

* L. Ruddigkeit et al. J Chem Inf Model 2012, 52, 2864–2875
Comparison of databases: GDB vs PubChem

21*10^6 cmpds

How can we select unique GDB molecules?
PubChem-17 vs GDB-17

21*10^6 cmpds

0.65*10^6 cmpds

2.5*10^3 cmpds

Chemography: hierarchical GTM

No analogues in PubChem
Databases comparison and properties profiling
Comparison of suppliers databases

~ 2.2 M cmpds from 37 subsets
(samples from the catalogs of 36 suppliers + NCI) *

GTM for the Suppliers DB (> 2 M cmpds)

Data density distributions built on responsibility vectors

Suppliers DB: GTM property landscapes

Molecular Weight

Chiral centers

Aqueous solubility (logS)
Suppliers DB: GTM property landscapes

High molecular weight & High chirality & Low solubility

Molecular Weight

Number of chiral centers

Aqueous solubility (logS)
Suppliers DB map: regions of interest

High molecular weight & High chirality & Low solubility

each library is described by GTM descriptors
Stargate GTM (S-GTM) setup

Descriptors space

Activities space

<table>
<thead>
<tr>
<th></th>
<th>Descri1</th>
<th>Descri2</th>
<th>Descri3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>0.22</td>
<td>5.43</td>
<td>1.12</td>
</tr>
<tr>
<td>Mol 2</td>
<td>7.18</td>
<td>0.96</td>
<td>-5.42</td>
</tr>
<tr>
<td>Mol 3</td>
<td>-1.01</td>
<td>7.41</td>
<td>10.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Act1</th>
<th>Act2</th>
<th>Act3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>6.24</td>
<td>9.35</td>
<td>7.11</td>
</tr>
<tr>
<td>Mol 2</td>
<td>7.32</td>
<td>10.22</td>
<td>8.53</td>
</tr>
<tr>
<td>Mol 3</td>
<td>6.03</td>
<td>6.78</td>
<td>7.11</td>
</tr>
</tbody>
</table>
Stargate GTM (S-GTM) setup

**Descriptors space**

<table>
<thead>
<tr>
<th></th>
<th>Descr1</th>
<th>Descr2</th>
<th>Descr3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>0.22</td>
<td>5.43</td>
<td>1.12</td>
</tr>
<tr>
<td>Mol 2</td>
<td>7.18</td>
<td>0.96</td>
<td>-5.42</td>
</tr>
<tr>
<td>Mol 3</td>
<td>-1.01</td>
<td>7.41</td>
<td>10.63</td>
</tr>
</tbody>
</table>

**Activities space**

<table>
<thead>
<tr>
<th></th>
<th>Act1</th>
<th>Act2</th>
<th>Act3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>6.24</td>
<td>9.35</td>
<td>7.11</td>
</tr>
<tr>
<td>Mol 2</td>
<td>7.32</td>
<td>10.22</td>
<td>8.53</td>
</tr>
<tr>
<td>Mol 3</td>
<td>6.03</td>
<td>6.78</td>
<td>7.11</td>
</tr>
</tbody>
</table>
S-GTM: prediction of pharmacological profile

Descriptors space

Activities space

Mol 1

<table>
<thead>
<tr>
<th></th>
<th>Descr1</th>
<th>Descr2</th>
<th>Descr3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>0.22</td>
<td>5.43</td>
<td>1.12</td>
</tr>
</tbody>
</table>
S-GTM: discovery of structures corresponding to a given pharmacological profile

![Diagram showing Descriptors space and Activities space]

<table>
<thead>
<tr>
<th></th>
<th>Act1</th>
<th>Act2</th>
<th>Act3</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>6</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Descr1</th>
<th>Descr2</th>
<th>Descr3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mol 1</td>
<td>??</td>
<td>??</td>
<td>??</td>
</tr>
<tr>
<td>Mol 2</td>
<td>??</td>
<td>??</td>
<td>??</td>
</tr>
<tr>
<td>Mol 3</td>
<td>??</td>
<td>??</td>
<td>??</td>
</tr>
</tbody>
</table>
### Case study: S-GTM models for a set of 8 GPCR activities

<table>
<thead>
<tr>
<th>ID</th>
<th>Target name</th>
<th>Mol 1</th>
<th>Mol 2</th>
<th>...</th>
<th>Mol 1325</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2a</td>
<td>Alpha-2a adrenergic receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>Dopamine D2 receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3</td>
<td>Dopamine D3 receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D4</td>
<td>Dopamine D4 receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1a</td>
<td>Serotonin 1a (5-H1a) receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2a</td>
<td>Serotonin 2a (5-H2a) receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>Serotonin 7 (5-HT7) receptor</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST</td>
<td>Serotonin transporter</td>
<td>AFFINITY</td>
<td>VALUES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
S-GTM: retrieval of structures corresponding to a given pharmacological profile

**Descriptors space** → **Activities space**

**EXP**

S-GTM: retrieval of structures corresponding to a given pharmacological profile

Activities space \rightarrow Descriptors space

\( A = \{6.88, 8.82, 6.76, 6.65, 7.67, 7.58, 7.22, 6.91\} \)

ISIDA-GTM Software

- **GTM Algorithms:**
  « classical », incremental and kernel,

- **Modeling:**
  regression and classification models

- **Visualization:**
  data points, data probability distribution, activity landscapes, chemical structures

H. Gaspar, G. Marcou, D. Horvath, A. Lin, F. Bonachera and A. Varnek, University of Strasbourg