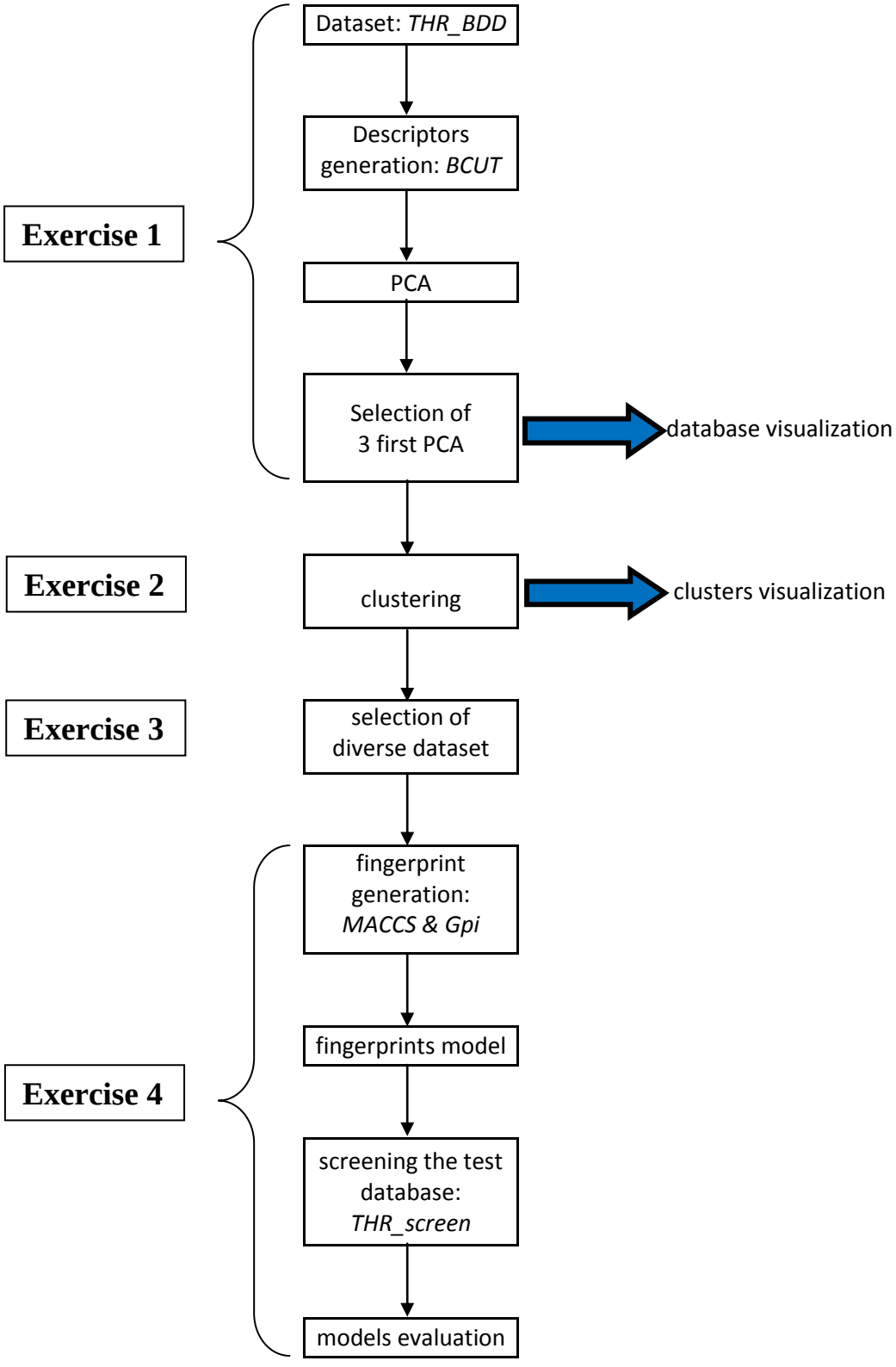
Tutorial on Chemical Similarity and Diversity

E. Lounkine and J. Bajorath (University of Bonn)
C. Muller and A. Varnek (University of Strasbourg)

Part 1:



Exercise 1. Projection of compounds into descriptor space

Calculating Descriptors

Open BindingDB Thrombin Inhibitors file: THR_BDB.mdb
File > open > Open in Database Viewer.

Compute > Descriptors > Calculate:

Choose "2D" as *descriptor Class*
Enter "BCUT" as *Filter*, select and calculate all BCUT descriptors

Principal Component Analysis

Compute > Descriptors > Principal Components

Select all BCUT descriptors
Set *Minimum Variance* to 95%
Click on "*Report*", then *OK*
New fields are created: PCA1 – PCA6

Showing a 3D Plot of first three PCs

Select Fields PCA1-3
Compute > Analysis > 3D Plot
Activity: Thrombin_nM
Click "*Plot*", look at the plot in the main MOE window
Click "Close"

Exercise 2. Descriptor based compound partitioning

Partitioning using principal components

MOE uses a partitioning scheme in order to assign cluster codes to molecules.

Compute > Descriptor > Clusters
We have already done PCA
Uncheck "*De-correlate Descriptors*"
"*Equiprobable subdivisions*" creates equally populated partitions.
Select PCA1-3
Set *Code Count* to 3. Thus, each axis will be divided into 3 parts, leading to 27 (3 x 3 x 3) possible cluster codes
Click "OK"
Select Field "$CLUSTER"
Compute > Sort > Select Unique Entries $CLUSTER
This will select 27 entries

Visualize clusters:

Compute > Analysis > 3D Plot

X,Y,Z: PCA1-3, *Activity*: $CLUSTER, *Threshold*: Unique

Select Unique $CLUSTER (as described above)

Display > Entry > Hide unselected

Look at the distribution in the main window

Render > Atoms > Ball and Stick

Display > Entry > Show all

Exercise 3. Calculation of a diverse subset

Diverse Subset

Diverse subsets are calculated based on Euclidean distances for each cluster. To determine, which unranked entry is farthest from all already-ranked entries, the distance between each unranked entry and each ranked entry is calculated. For each unranked entry, the minimum of its distances to each ranked entry is found. The entry with the largest such minimum distance is the farthest.

Compute > Diverse Subset > Method: Descriptors

Cluster Field: $CLUSTER

Select PCA 1-6

Output limit: 0 (no limit)

This will calculate a $DIVPRIO field.

Deselect all molecules *Edit > Clear > Entry selection*

Edit > Select

and *Add to Entry selection* **$DIVPRIO** *matches* **<= 2** (select two diverse representatives of each cluster)

This will select 54 compounds representing the structural BCUT space covered by the whole thrombin antagonist set.

Select the fields “mol” and “Thrombin_nM”

File\Save As

MOE molecular database (mdb)

Selected Fields only

Selected entries only

Export to **THR_BDB_BCUT_subset.mdb**

Now, we have a diverse reference set of active Thrombin antagonists.

Exercise 4. Molecular Fingerprints

Calculating Fingerprints

Open the Thrombin reference set.

Compute > Fingerprints > Calculate
FP:MACCS (keyed fingerprint, 166 keys)
FP:GpiDAPH3 (2D Pharmacophore)

Fingerprint Model

A fingerprint model saves information about the fingerprints, the similarity metric and the search strategy to use. Create two separate fingerprint models for the MACCS and the GpiDAPH3 fingerprints.

Compute > Model > Fingerprint
Score: Maximum (corresponds to nearest neighbor, meaningful since we have a diverse representative set)
Save the two fingerprint models under **THR_MACCS.fpt** and **THR_Gpi.fpt**

Multiple ligand similarity searching

The two fingerprint models can be used in order to sort database compounds according to fingerprint Tanimoto similarity.

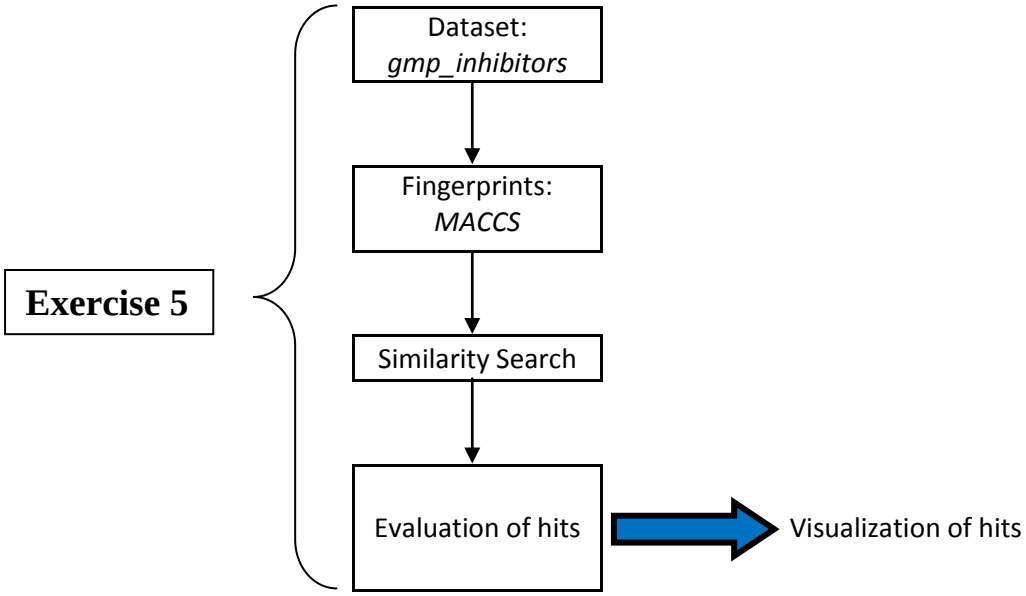
Open Thrombin screening data: **THR_screen.mdb**

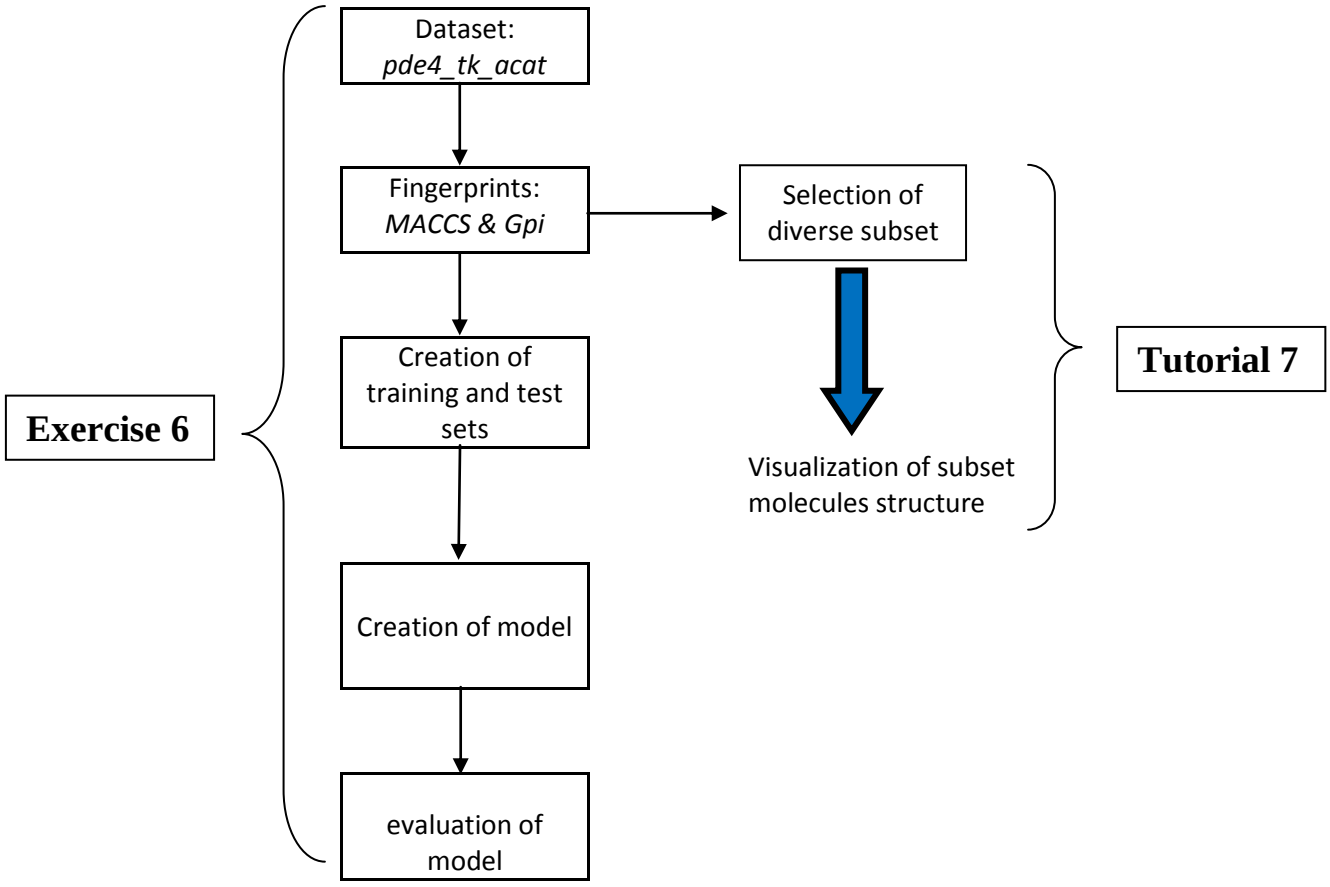
Apply the two models:
Compute > Model > Evaluate
Model File: your .fpt file
Field: $PRED_MACCS or $PRED_Gpi

Evaluate the performance of each model:
Sort descending by $PRED (*Compute > Sort*)
Select first 100 entries
Edit > Select: Intersect Entry selection **Active** *matches* **= 1**.

This gives the hit rate for each of the models.

Part 2


Exercise 5
Dataset:
gmp_inhibitors
Fingerprints:
MACCS
Similarity Search
Evaluation of hits
Visualization of hits


Exercise 6
Dataset:
pde4_tk_acat
Fingerprints:
MACCS & Gpi
Creation of training and test sets
Creation of model
evaluation of model
Selection of diverse subset
Visualization of subset molecules structure
Tutorial 7

Exercise 5. GMP Inhibitors Similarity Search

Open the database file: $MOE/sample/mol/**gmp_inhibitors.mdb**
File > open > Open in Database Viewer.

Save the database
File > Save As
Export to **my_gmp.mdb**

Fingerprints calculation

Compute > Fingerprints > Calculate
Fingerprint: MACCS keys
Press *OK*

Fingerprints can be used to compute the similarity between two or more molecules. The molecular similarity can be assessed using the degree of overlap using various similarity metrics: Tanimoto, Euclidian …

Choosing a reference molecule

Sort the pIC50 from low to high to place the most active inhibitor at the top of the list.

Open the panel *Compute > Sort*.
Sort by *Field* pIC50
Select *Descending*.
Press *OK*.

Copy the most active molecule into the main MOE window, it will be the reference molecule for the similarity search.

Right click on the structure and *Send To MOE*.

Similarity Search

Open the MOE-Similarity Search panel
Compute > Fingerprint > Search
Visibility: Nothing
Selection: Select Hits
Overlap: 50
Click on *Set Fingerprint*…
Select FP:MACCS
OK
Search

How many molecules are considered similar at less 50% to the reference molecule?

Clear the entry selection
Entry > Clear entry selection

Compute > Fingerprint > Search
Visibility: Nothing
Selection: Select Hits
Overlap: 70

How many molecules are considered similar at less 70% to the reference molecule?

Exercise 6. Fingerprint Modeling

The Example database contains active inhibitors for three target class:
- Phosphodiesterase (PDE4, 33 entries)
- Tyrosine Kinase (TK, 44 entries)
- Acyl Coenzyme A: cholesterol acyltransferase (ACAT, 58 entries)

These active inhibitors are combined with 2837 random drug-like structures.

Open the database file: $MOE/sample/mol/**pde4_tk_acat.mdb**
File > open > Open in Database Viewer.

Save the database
File > Save As
Export to **multi_class.mdb**

Fingerprints calculation

Compute > Fingerprints > Calculate
Fingerprint: GpiDAPH3 and MACCS keys
Press *OK*

Create a Training and Test set

Compute > Calculator
Click on *RAND*
Destination Field: Rand
Press *Evaluate*
Then *Close*

Edit > Select: Add To Entry selection: ***ACTIVE*** *matches >* ***0****.*
Edit > Select: Intersect Entry selection: ***Rand*** *matches >* ***0.8****.*
This last procedure selects ~20% of the active compounds.

File > Save As
Selected entries only
Select *New*
Export to **fp_training.mdb**

In the original dataset (multi_class.mdb) delete the training set compounds.

Edit > Delete > Selected Entries

This database is now the test set.

Creation of model

In the **fp_training.mdb**:

Compute > Model > Fingerprint

Fingerprint Type: FP:MACCS
Score: Average
Save

Name the fingerprint model **MACCS_average.fpt**
OK

Fingerprint Type: FP:MACCS
Score: Maximum
Save

Name the fingerprint model **MACCS_maximum.fpt**
OK
Cancel

Evaluation of models

In the **multi_class.mdb**:

Compute > Model > Evaluate

Model File: MACCS_average.fpt
Field: $AVE_PRED
OK

Compute > Model > Evaluate

Model File: MACCS_maximum.fpt
Field: $MAX_PRED
OK

The compounds in the test database are compared to the fingerprint model and if the entry meets the overlap criteria, it is a hit.

Comparing MAX and AVE similarity scores

Compute > Calculator

Click the *index* button
Destination Field: INDEX
Close

*Edit > Select: Add To Entry selection: **ACTIVE** matches > **0**.*
Active compounds in the Test dataset are now selected.

Compute > Analysis > Correlation Plot

Select **INDEX** and **$AVE_PRED** to plot along the X and Y-axis, respectively.

Compute > Analysis > Correlation Plot

Select **INDEX** and **$MAX_PRED** to plot along the X and Y-axis, respectively.

Which score provides a better separation for this dataset when comparing active and inactive compounds?

Exercise 7. Diverse subset selection

Compute > Diverse Subset
Output Limit: 10
Method: Fingerprint
Click *Set Fingerprint*
Fingerprint: FP:MACCS
OK
OK

Compute > Sort $DIVPRIO
Select the 10 most diverse compounds and copy them to MOE.
MOE | *Compute > 2D Molecules*