A Deep Learning Model for Molecular Fingerprinting

Mattia Cordioli

Department of Electrical, Computer and Biomedical Engineering

Master's Degree in Bioengineering

Advisor:
Prof. Riccardo Bellazzi

Co-advisor:
Prof. Blaž Zupan

biolab

Univerza *v Ljubljani*
Fakulteta *za računalništvo*
in informatiko

bmi BIO-MEDICAL INFORMATICS
"Mario Stefanelli"

UNIVERSITÀ DI PAVIA

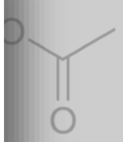
Outline

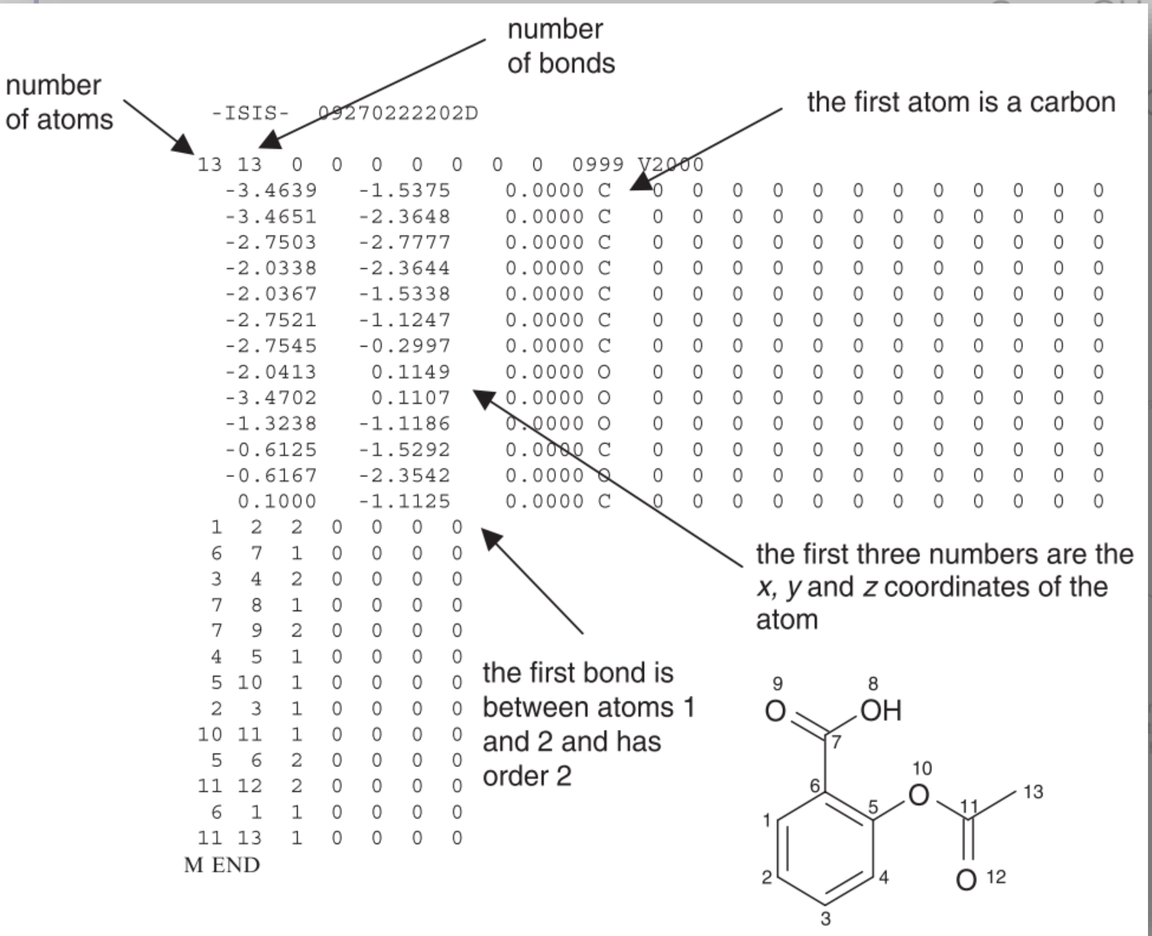
- Introduction to Chemoinformatics
  - Molecule representation techniques
  - Fingerprints
- A novel approach: Deep Learning fingerprints
  - Convolutional Neural Networks
- Results on different datasets
- Software development
- Conclusions and future developments

Molecular Representations

- Standard representations:
  - **not sig**
  - **2D stru**
  - **molecu**
- Molecular
  - **connec**
  - encode
- Linear nota
  - **more c**
  - useful molecul
  - **SMILE** Entry Specification

Aspirin - $C_9H_8O_4$





number of atoms

number of bonds

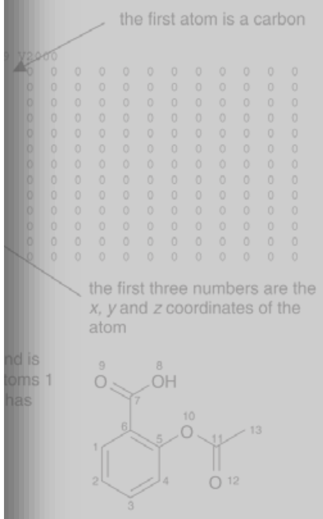
the first atom is a carbon

```
  -ISIS-  09270222202D

 13 13  0  0  0  0  0  0  0  0999 V2000
   -3.4639   -1.5375    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4651   -2.3648    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7503   -2.7777    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0338   -2.3644    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0367   -1.5338    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7521   -1.1247    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.7545   -0.2997    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.0413    0.1149    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -3.4702    0.1107    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3238   -1.1186    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6125   -1.5292    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.6167   -2.3542    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.1000   -1.1125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  6  7  1  0  0  0  0
  3  4  2  0  0  0  0
  7  8  1  0  0  0  0
  7  9  2  0  0  0  0
  4  5  1  0  0  0  0
  5 10  1  0  0  0  0
  2  3  1  0  0  0  0
 10 11  1  0  0  0  0
  5  6  2  0  0  0  0
 11 12  2  0  0  0  0
  6  1  1  0  0  0  0
 11 13  1  0  0  0  0
M END
```

the first three numbers are the *x, y* and *z* coordinates of the atom

the first bond is between atoms 1 and 2 and has order 2



the first atom is a carbon

the first three numbers are the x, y and z coordinates of the atom

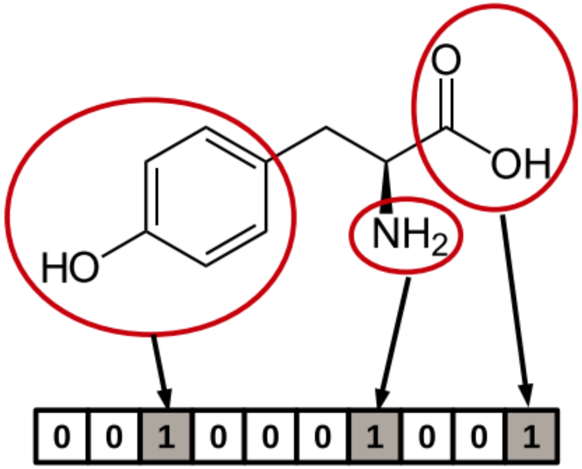
CC(=O)Oc1ccccc1C(=O)O

Fingerprints

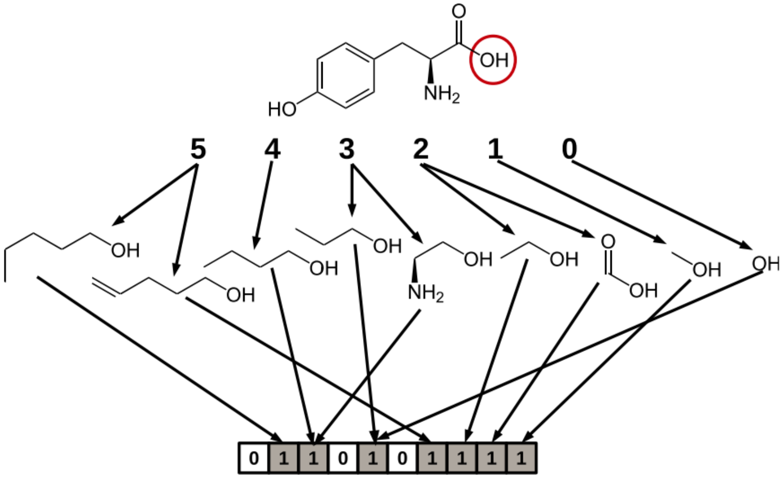
- SMILES are **not enough** in Chemoinformatics applications:
  - **Similarity** and **substructure search**
  - **Virtual screening**
  - **QSAR** and **machine learning** models

- **Fingerprints**: binary vectors of fixed length

Substructure keys-based



Topological / Hashed



Deep Learning for Fingerprinting

- Standard fingerprints **limits:**
  - Necessity of a **fragments dictionary** for substructure keys-based FPs
  - Topological FPs are usually **really long** (1024 – 2048 bits)
  - Binary, **not real-valued**
  - **Not trainable** for target-specific tasks

- Novel approach: molecular **embedding** through **Deep Learning**
  - **Deep Neural Networks** learn and abstract powerful representations of input data

- Literature approaches:
  - CNN for **molecular graphs convolution**
  - CNN applied directly to **SMILES strings**

Aims of the Thesis

- Development of a deep learning model for **molecular fingerprinting**:
  - **Simple CNN** architecture
  - **SMILES strings** as input

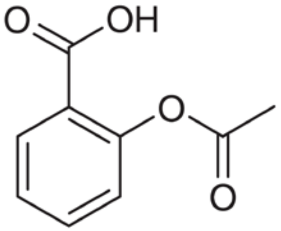
- Creation of a new **Chemoinformatics Add-on** for Orange:
  - To provide a tool for easily analyse and work with chemistry data
  - Implementation of the model in a tool for **molecular embedding**
  - Implementation of other useful tools, e.g. **molecules visualization**

Data Retrieval

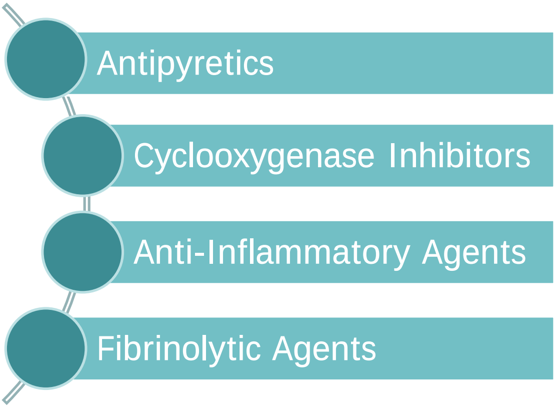
- Open chemistry database at NIH
  - **~92 million** compounds with information about structure, chemical/physical properties, identifiers, pharmacology, toxicity, patents, ...

- **MeSH** Ontology terms for **pharmacological actions**

- **PubChemAPI** Python library:
  - Programmatic access to PubChem to retrieve data
  - Linking to MeSH Ontology DB to retrieve associated pharmacological actions



Aspirin Pharma Actions:



- Antipyretics
- Cyclooxygenase Inhibitors
- Anti-Inflammatory Agents
- Fibrinolytic Agents

Data Preprocessing

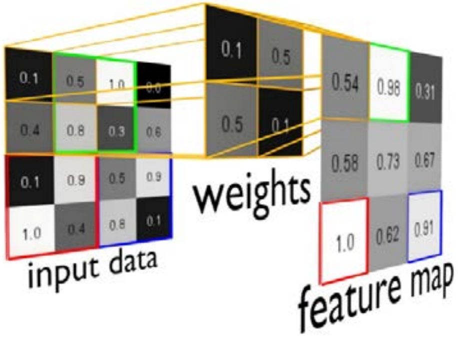
- **15 474 compounds** retrieved, annotated with **489 terms**
  - **CID, SMILES, name, formula, MeSH terms + tree numbers**

- **Preprocessing:**
  - **Duplicate rows** (same SMILES and terms, different names)
  - Terms appearing **<20 times**
  - Terms with tree number **not starting with ‘D27.505’**

- **Final dataset:**
  - **9 174 records**
  - **191 terms**

Embedding Model: CNNs

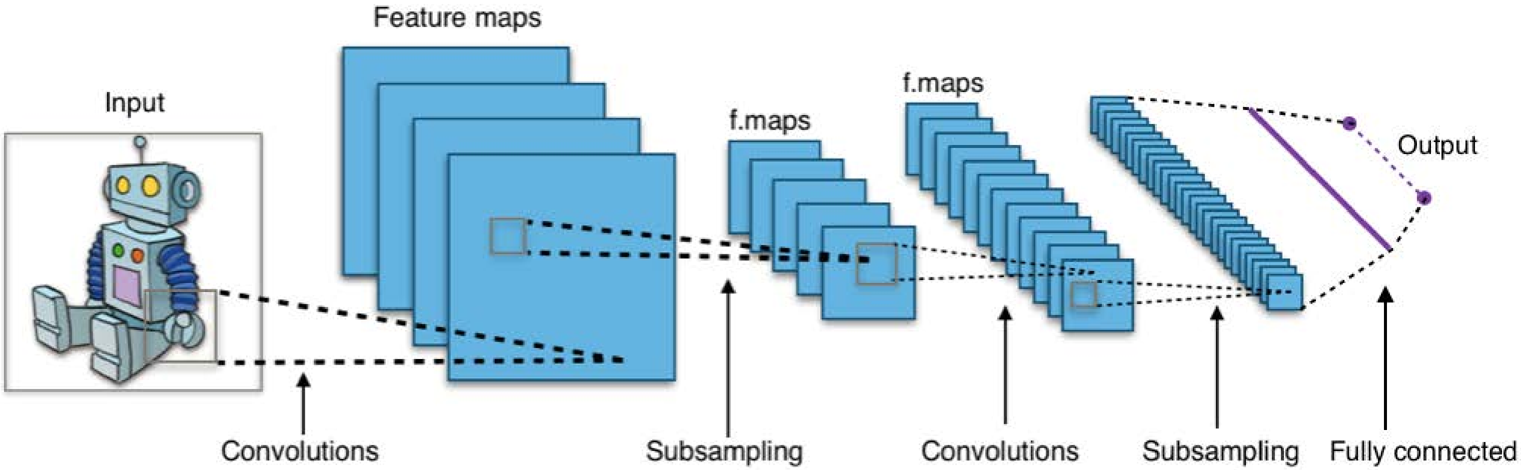
- **Convolutional Neural Networks:**
  - Input data convoluted with **kernels** to obtain **feature maps**
  - **MAX Pooling** layers to reduce dimensionality
  - **Fully connected** layers on top
  - **Dropout** to prevent overfitting



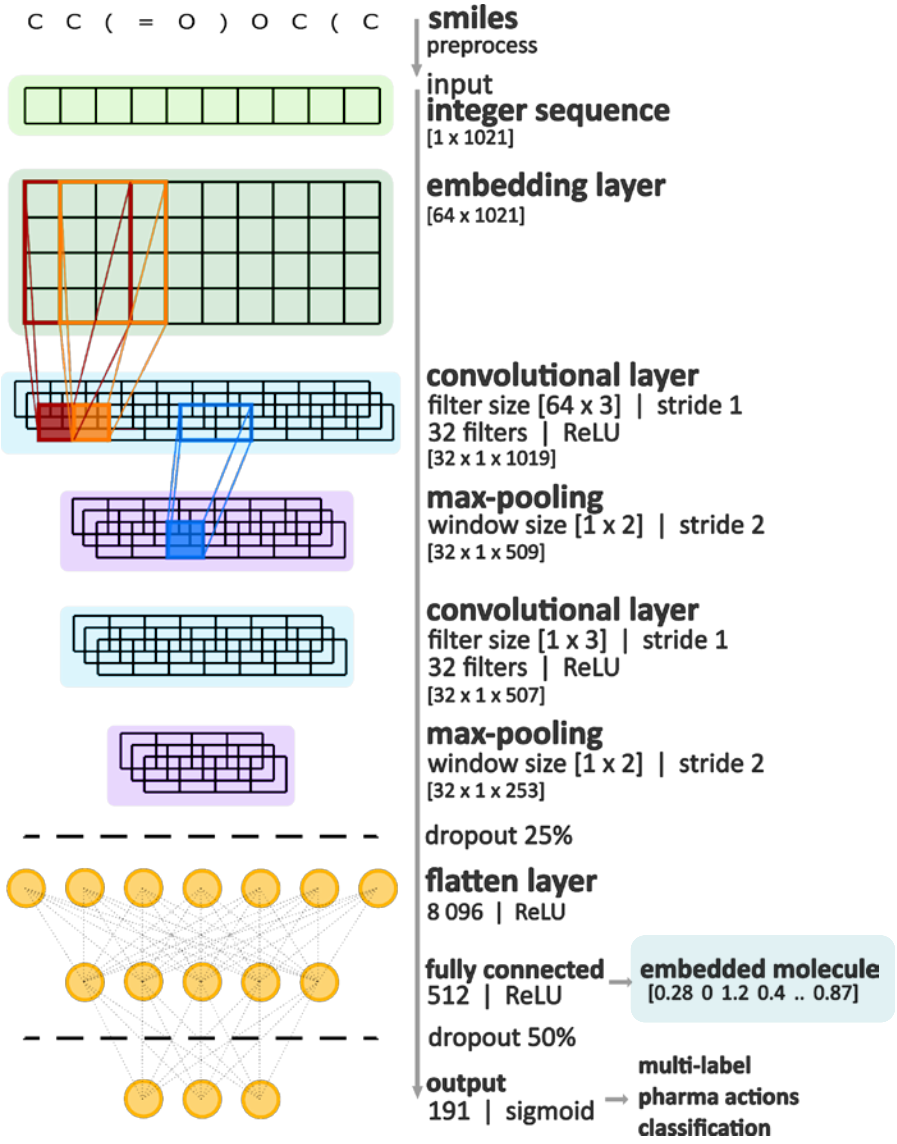
0.1	0.5	1.0	0.0
0.4	0.8	0.3	0.6
0.1	0.9	0.5	0.9
1.0	0.4	0.8	0.1

0.1	0.5
0.5	0.1

0.54	0.98	0.31
0.58	0.73	0.67
1.0	0.62	0.91



Embedding Model: Architecture



smiles
preprocess

input
integer sequence
[1 x 1021]

embedding layer
[64 x 1021]

convolutional layer
filter size [64 x 3] | stride 1
32 filters | ReLU
[32 x 1 x 1019]

max-pooling
window size [1 x 2] | stride 2
[32 x 1 x 509]

convolutional layer
filter size [1 x 3] | stride 1
32 filters | ReLU
[32 x 1 x 507]

max-pooling
window size [1 x 2] | stride 2
[32 x 1 x 253]

dropout 25%

flatten layer
8 096 | ReLU

fully connected → **embedded molecule** [0.28 0 1.2 0.4 .. 0.87]
512 | ReLU

dropout 50%

output → **multi-label pharma actions classification**
191 | sigmoid

- **Keras** to design and train the model
- **Supervised** learning
  - Pharmacological actions **multi-label classification**
- Penultimate layer activations
  - **512** bits real-valued **fingerprint**
- Training
  - **on the entire dataset**
  - **GeForce GTX TITAN X GPU**
- Model saved to be used as **embedder**

Validation: CNN Performance

- Assessing CNN performance:
  - **70/30 train/test split**

- Metric:
  - Area Under ROC Curve (**AUC**) **for each term** separately

Minimum AUC	**Maximum AUC**	**Mean AUC**
0.62	0.99	0.87

Comparison: MeSH Terms Prediction

- Comparison with **ECFP**:
  - Circular/topological fingerprint
  - **Standard in QSAR**
  - **512-bits** version

- **Pharmacological Actions** prediction:
  - **Logistic Regression**
  - **One-Vs-All** approach
  - 10-Fold Cross Validation
  - AUC

Fingerprint	Mean AUC
CNNFP	**0.99**
ECFP	0.92

- **Non-Pharma** terms prediction:
  - 1 091 compounds discarded in the preprocessing phase

Fingerprint	Mean AUC
CNNFP	0.83
ECFP	**0.92**

Comparison: Other QSAR Datasets

- Comparison on datasets obtained from MoleculeNet:
  - **Logistic Regression** and **Random Forest**
  - CNNFP, ECFP, CNNFP+ECFP
  - 10-Fold CV AUC

- **ClinTox**:
  - **1 491** compounds
  - Clinical Trial toxicity
  - FDA approval status

Task	Classifier	ECFP	CNNFP	CNNFP+ECFP
CT Toxicity	LR	0.72	**0.93**	**0.95**
	RF	0.74	**0.94**	**0.96**
FDA Approval	LR	0.74	**0.92**	**0.95**
	RF	0.74	**0.94**	**0.97**

Comparison: Other QSAR Datasets - 2

- **BACE**:
  - **1 522** β-secretase-1 inhibitors
  - Binding results

- **BBBP:**
  - **2 000** compounds
  - Blood-brain barrier permeability

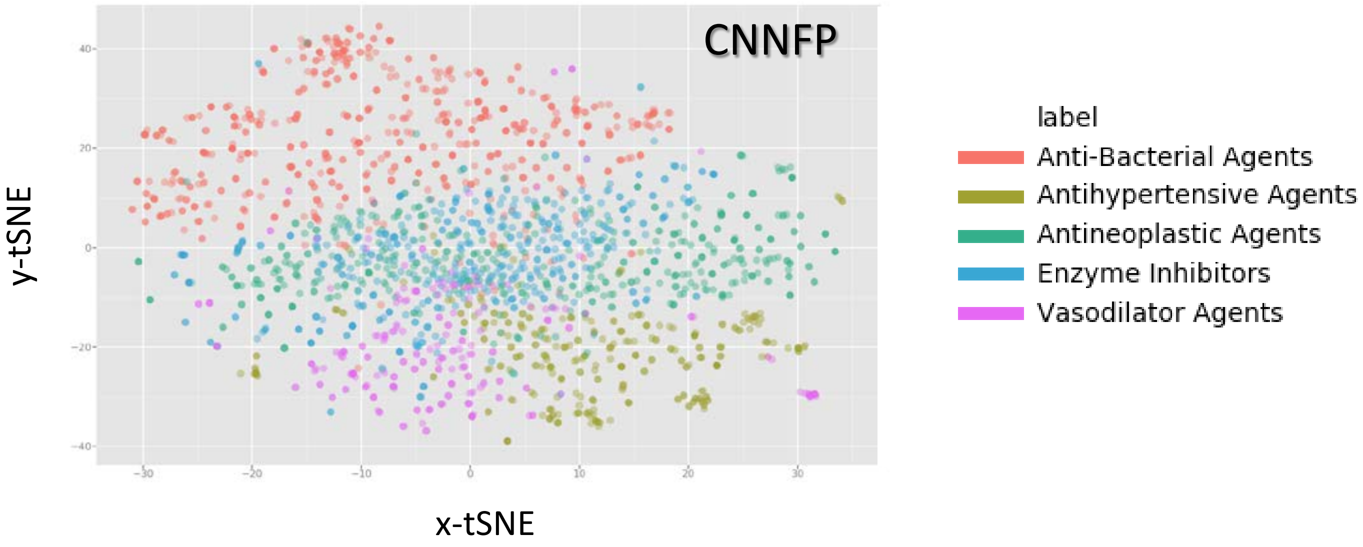
Dataset	Classifier	ECFP	CNNFP	CNNFP+ECFP
BACE	LR	**0.85**	0.72	0.81
	RF	**0.87**	0.79	0.86
BBBP	LR	**0.83**	0.79	**0.85**
	RF	0.88	**0.89**	**0.91**

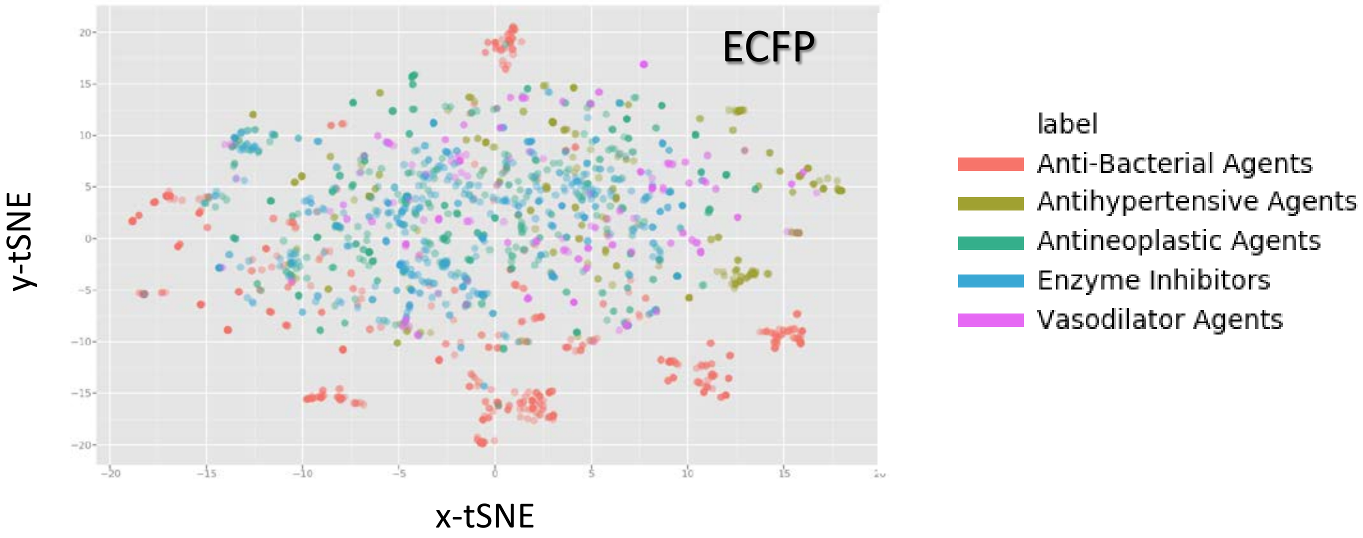
Comparison: t-SNE Visualization

- **Data**:
  - Compounds related to the 5 most frequent MeSH Terms
  - ClinTox
  - BACE
  - BBBP
- **t-SNE**:
  - Non-linear dimensionality reduction
  - **PCA** (**100 components**) applied before

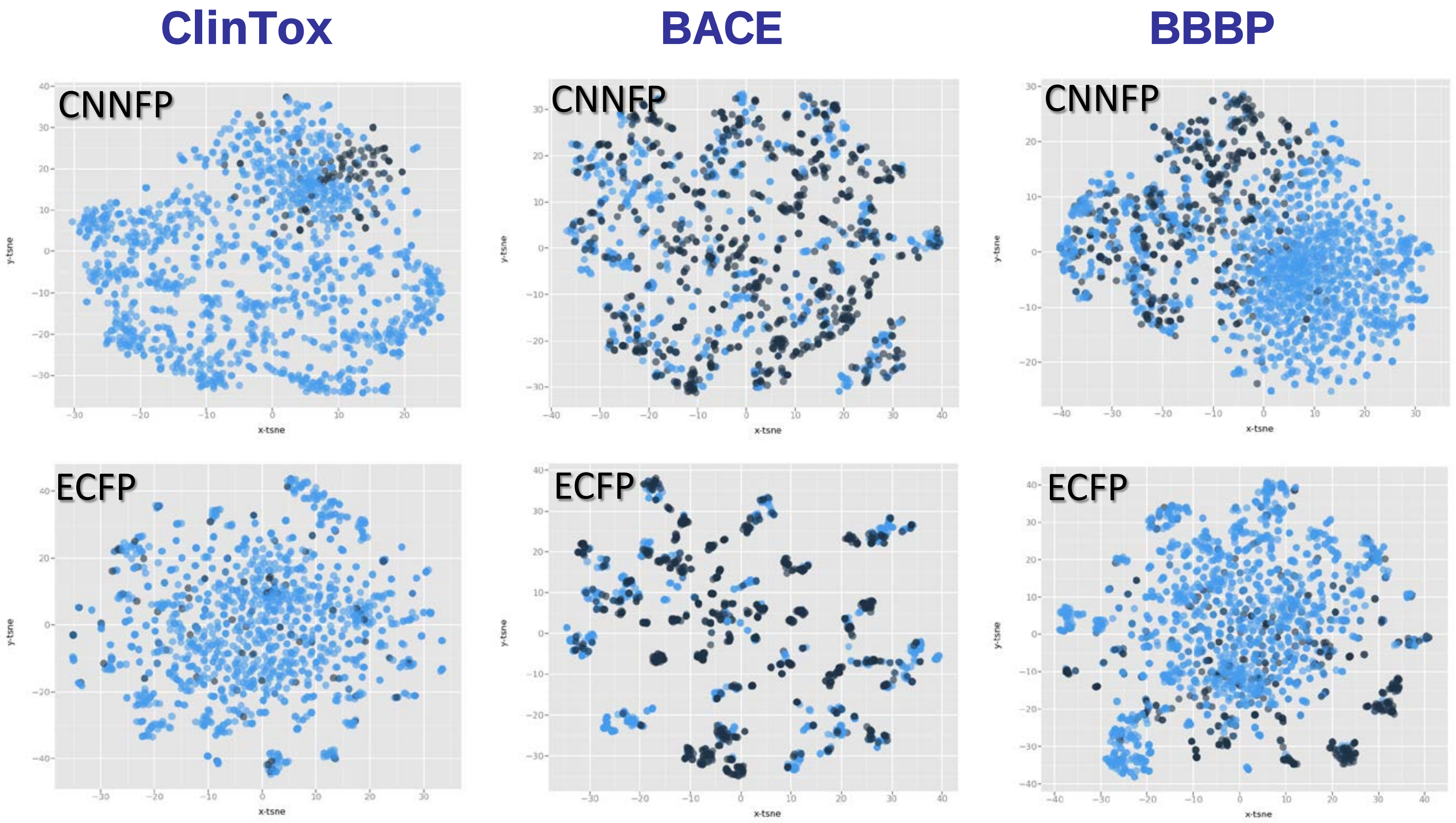
	Explained Variance	
	ECFP	**CNNFP**
MeSH Terms	52%	**91%**
ClinTox	65%	**90%**
BACE	84%	**94%**
BBBP	67%	**87%**

Comparison: t-SNE Visualization - 2



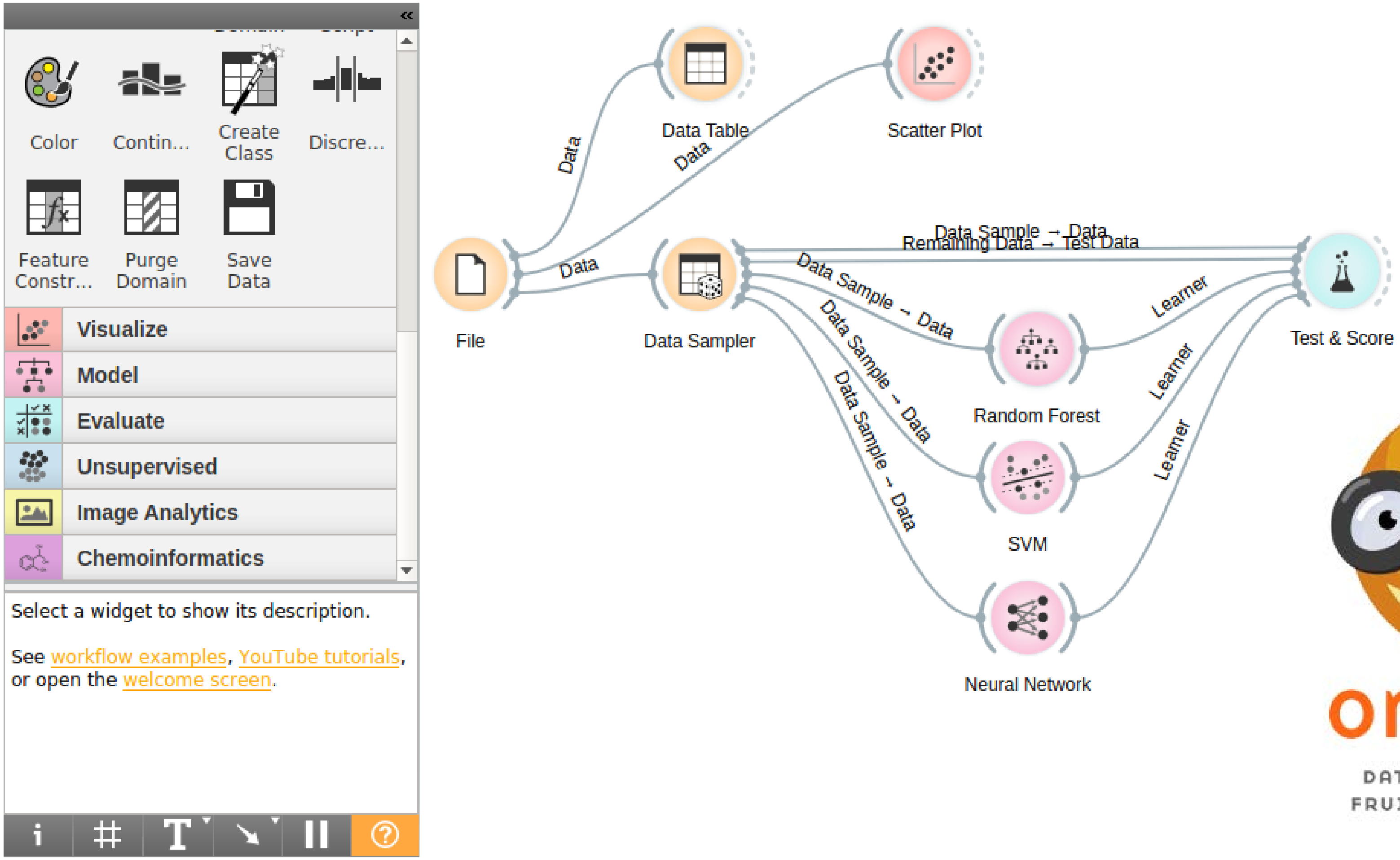


Comparison: t-SNE Visualization - 3



Orange

- **Machine Learning and Data Mining suite** developed by **Bioinformatics Lab** at **University of Ljubljana**



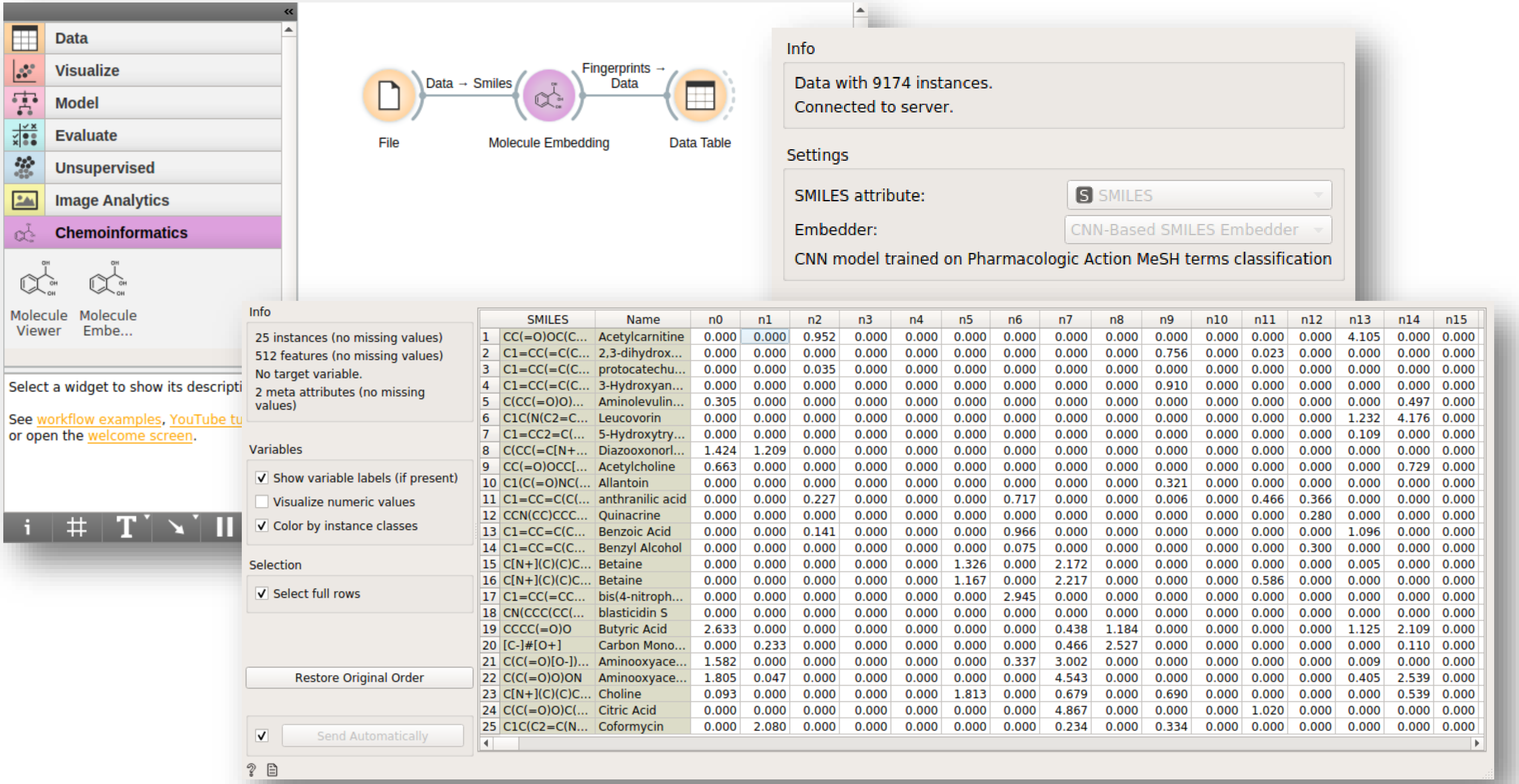
Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

orange
DATA MINING
FRUITFUL&FUN

Chemoinformatics Add-on

- **Molecule Embedding Widget**
  - Embedding on Orange server, using Keras and TITAN X GPU



Select a widget to show its descripti

See workflow examples, YouTube tu
or open the welcome screen.

Info

Data with 9174 instances.
Connected to server.

Settings

SMILES attribute: SMILES

Embedder: CNN-Based SMILES Embedder

CNN model trained on Pharmacologic Action MeSH terms classification

Info

25 instances (no missing values)
512 features (no missing values)
No target variable.
2 meta attributes (no missing values)

Variables

☑ Show variable labels (if present)
☐ Visualize numeric values
☑ Color by instance classes

Selection

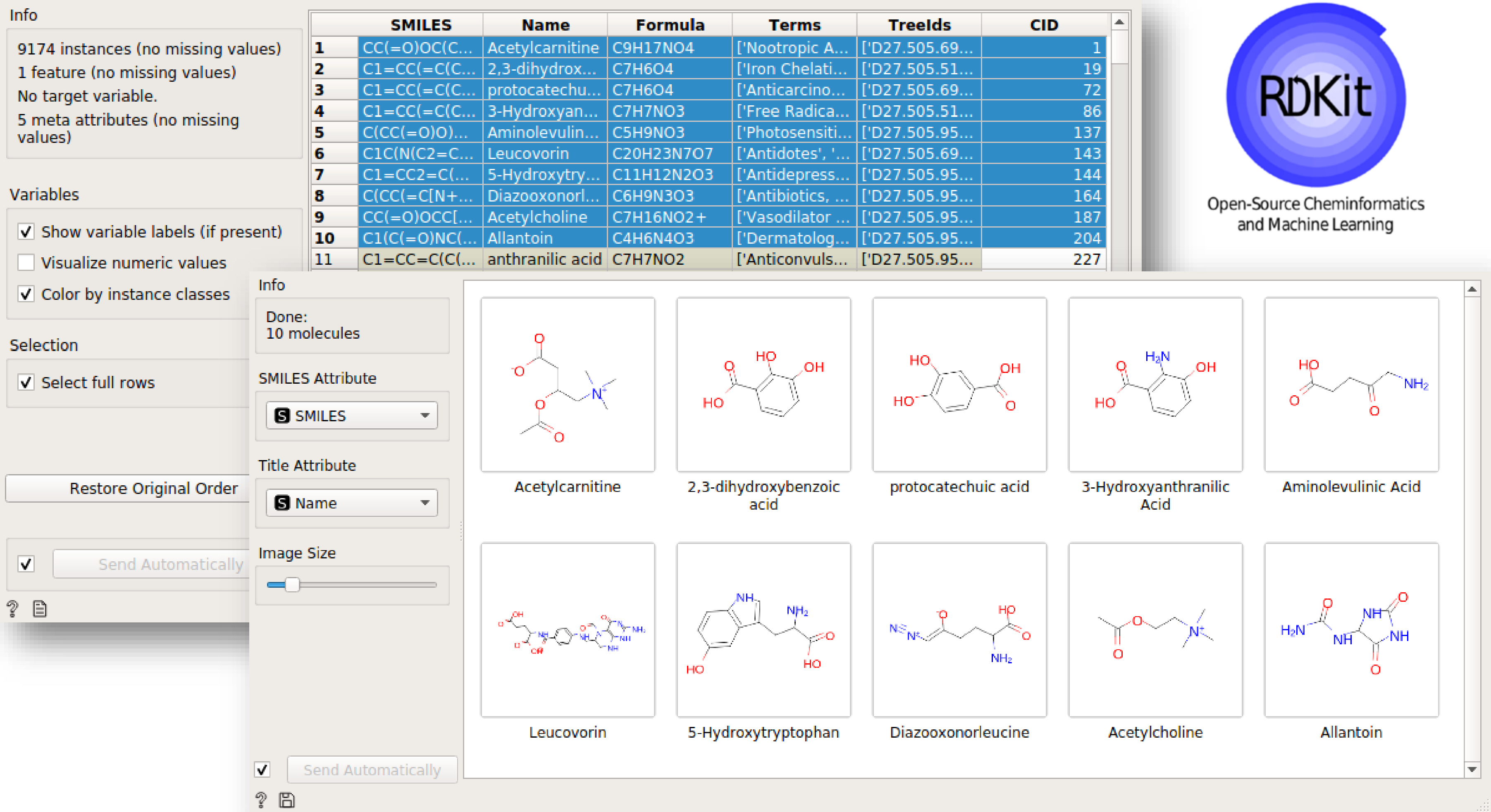
☑ Select full rows

☑ Send Automatically

	SMILES	Name	n0	n1	n2	n3	n4	n5	n6	n7	n8	n9	n10	n11	n12	n13	n14	n15
1	CC(=O)OC(C...	Acetylcarnitine	0.000	0.000	0.952	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.105	0.000	0.000
2	C1=CC(=C(C...	2,3-dihydrox...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.756	0.000	0.023	0.000	0.000	0.000	0.000
3	C1=CC(=C(C...	protocatechu...	0.000	0.000	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	C1=CC(=C(C...	3-Hydroxyan...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.910	0.000	0.000	0.000	0.000	0.000	0.000
5	C(CC(=O)O)...	Aminolevulin...	0.305	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.497	0.000
6	C1C(N(C2=C...	Leucovorin	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.232	4.176	0.000
7	C1=CC2=C(...	5-Hydroxytry...	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.109	0.000	0.000
8	C(CC(=C[N+...	Diazooxonorl...	1.424	1.209	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	CC(=O)OCC[...	Acetylcholine	0.663	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.729	0.000
10	C1(C(=O)NC(...	Allantoin	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.321	0.000	0.000	0.000	0.000	0.000	0.000
11	C1=CC=C(C(...	anthranilic acid	0.000	0.000	0.227	0.000	0.000	0.000	0.717	0.000	0.000	0.006	0.000	0.466	0.366	0.000	0.000	0.000
12	CCN(CC)CCC...	Quinacrine	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.280	0.000	0.000	0.000
13	C1=CC=C(C...	Benzoic Acid	0.000	0.000	0.141	0.000	0.000	0.000	0.966	0.000	0.000	0.000	0.000	0.000	0.000	1.096	0.000	0.000
14	C1=CC=C(C...	Benzyl Alcohol	0.000	0.000	0.000	0.000	0.000	0.000	0.075	0.000	0.000	0.000	0.000	0.000	0.300	0.000	0.000	0.000
15	C[N+](C)(C)C...	Betaine	0.000	0.000	0.000	0.000	0.000	1.326	0.000	2.172	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000
16	C[N+](C)(C)C...	Betaine	0.000	0.000	0.000	0.000	0.000	1.167	0.000	2.217	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000
17	C1=CC(=CC...	bis(4-nitroph...	0.000	0.000	0.000	0.000	0.000	0.000	2.945	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	CN(CCC(CC(...	blasticidin S	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	CCCC(=O)O	Butyric Acid	2.633	0.000	0.000	0.000	0.000	0.000	0.000	0.438	1.184	0.000	0.000	0.000	0.000	1.125	2.109	0.000
20	[C-]#[O+]	Carbon Mono...	0.000	0.233	0.000	0.000	0.000	0.000	0.000	0.466	2.527	0.000	0.000	0.000	0.000	0.000	0.110	0.000
21	C(C(=O)[O-])...	Aminooxyace...	1.582	0.000	0.000	0.000	0.000	0.000	0.337	3.002	0.000	0.000	0.000	0.000	0.000	0.009	0.000	0.000
22	C(C(=O)O)ON	Aminooxyace...	1.805	0.047	0.000	0.000	0.000	0.000	0.000	4.543	0.000	0.000	0.000	0.000	0.000	0.405	2.539	0.000
23	C[N+](C)(C)C...	Choline	0.093	0.000	0.000	0.000	0.000	1.813	0.000	0.679	0.000	0.690	0.000	0.000	0.000	0.000	0.539	0.000
24	C(C(=O)O)C(...	Citric Acid	0.000	0.000	0.000	0.000	0.000	0.000	0.000	4.867	0.000	0.000	0.000	1.020	0.000	0.000	0.000	0.000
25	C1C(C2=C(N...	Coformycin	0.000	2.080	0.000	0.000	0.000	0.000	0.000	0.234	0.000	0.334	0.000	0.000	0.000	0.000	0.000	0.000

Chemoinformatics Add-on - 2

- ## Molecule Viewer Widget



Info

9174 instances (no missing values)
1 feature (no missing values)
No target variable.
5 meta attributes (no missing values)

Variables

☑ Show variable labels (if present)
□ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

	SMILES	Name	Formula	Terms	TreeIds	CID
1	CC(=O)OC(C...	Acetylcarnitine	C9H17NO4	['Nootropic A...	['D27.505.69...	1
2	C1=CC(=C(C...	2,3-dihydrox...	C7H6O4	['Iron Chelati...	['D27.505.51...	19
3	C1=CC(=C(C...	protocatechu...	C7H6O4	['Anticarcino...	['D27.505.69...	72
4	C1=CC(=C(C...	3-Hydroxyan...	C7H7NO3	['Free Radica...	['D27.505.51...	86
5	C(CC(=O)O)...	Aminolevulin...	C5H9NO3	['Photosensiti...	['D27.505.95...	137
6	C1C(N(C2=C...	Leucovorin	C20H23N7O7	['Antidotes', '...	['D27.505.69...	143
7	C1=CC2=C(...	5-Hydroxytry...	C11H12N2O3	['Antidepress...	['D27.505.95...	144
8	C(CC(=C[N+...	Diazooxonorl...	C6H9N3O3	['Antibiotics, ...	['D27.505.95...	164
9	CC(=O)OCC[...	Acetylcholine	C7H16NO2+	['Vasodilator ...	['D27.505.95...	187
10	C1(C(=O)NC(...	Allantoin	C4H6N4O3	['Dermatolog...	['D27.505.95...	204
11	C1=CC=C(C(...	anthranilic acid	C7H7NO2	['Anticonvuls...	['D27.505.95...	227

Info

Done:
10 molecules

SMILES Attribute: SMILES

Title Attribute: Name

Acetylcarnitine

2,3-dihydroxybenzoic acid

protocatechuic acid

3-Hydroxyanthranilic Acid

Aminolevulinic Acid

Leucovorin

5-Hydroxytryptophan

Diazooxonorleucine

Acetylcholine

Allantoin

RDKit

Open-Source Cheminformatics and Machine Learning

Conclusions and Future Developments

- **Conclusions:**
  - Novel deep learning model for molecular fingerprinting
  - Short real-valued fingerprint (512 bits) with high representative power
  - Simple architecture, using simple input representation
  - Good capability of generalitazion
  - Trainable for target specific applications

- **Future Developments:**
  - Optimize a tool for pharmacological actions prediction
    - Drug repurposing
  - Chemical interpretation of the learned features
  - Extension of Chemoinformatics Add-on functionalities

Grazie per l’attenzione