# Sviluppo di modelli QSAR predittivi mediante l'uso estensivo di tecniche Machine Learning: Applicazione sugli inibitori della Lisina Istone Demetilasi 1 (KDM1A)


SAPIENZA
UNIVERSITÀ DI ROMA

Facoltà di Farmacia e Medicina
Corso di Laurea in Chimica e Tecnologia Farmaceutiche
Tesi Sperimentale in Chimica Farmaceutica
a.a. 2014/2015

Laureando: Alexandros Patsilinakos
Matricola: 317021

Relatore: Prof. Rino Ragno

# Chemical Space

by www.RCMD.it

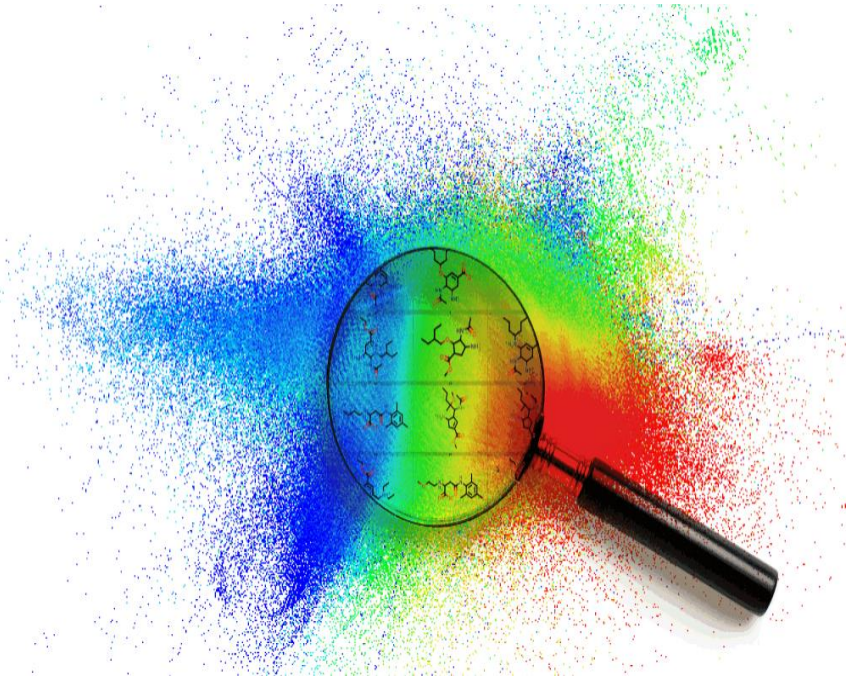
***"Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is".*** *Douglas Adams*

There are something like **$10^{22}$ to $10^{24}$ stars in the Universe**.[1]

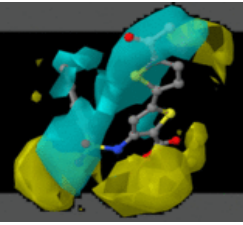
What about the "Chemical Universe"?

The total number of possible small organic molecules that populate "chemical space" has been estimated to exceed **$10^{60}$**.[2]

This is an amount so vast when compared to the number of such molecules we have made, or indeed could ever hope to make, that it might as well be infinite.[2]




1. http://www.esa.int/Our_Activities/Space_Science/Herschel/How_many_stars_are_there_in_the_Universe
2. Kirkpatrick, Peter, and Clare Ellis. "Chemical space." Nature 432.7019 (2004): 823.



# Available Molecular Databases

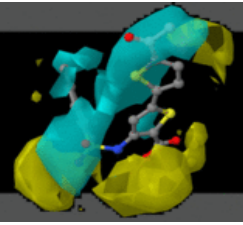
by www.RCMD.it

**Exploration of chemical space has so far been extremely limited.**

- The largest current public database of molecules so far synthesized, PubChem, contains around 50 orders of magnitude fewer (~70 million)

- The Chemical Universe Database GDB-17 lists 166.4 billion molecules of 17 atoms or less, which is four order of magnitude more than the number of known molecules in that size range.

| DATABASE | DESCRIPTION | SIZE |
|---|---|---|
| DrugBank | approved and investigational drugs | 7593 |
| SuperScent | scents from literature | 2300 |
| Flavornet | volatile compounds from literature | 738 |
| SuperSweet | carbohydrates and artificial sweeteners | 642 |
| BitterDB | bitter cpds from literature and Merck index | 606 |
| PubChem | NIH repository of molecules | ~70M |
| ZINC | commercial small molecules | 22 724 825 |
| ZINC.FL | fragrance-like subset of ZINC | 69 724 |
| BindingDB | small molecules annotated with bioactivity data | 453 657 |
| ChEMBL | small molecules annotated with bioactivity data | 1411 786 |
| GDB-11 | molecules of up to 11 atoms of C, N, O, and F | 26 434 571 |
| GDB-13 | molecules of up to 13 atoms of C, N, O, S, and Cl | 977 468 314 |
| GDB-13.subset | simplicity-selected GDB-13 molecules | 43 729 989 |
| GDB-13.FL | fragrance-like subset of GDB-13 | 59482 898 |
| GDB-17 | molecules of up to 17 atoms of C, N, O, S, and halogens | 166 443 860 262 |



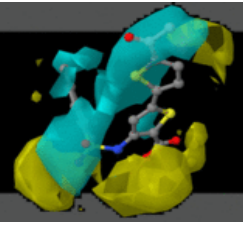
# Challenges in Drug Discovery



How the enormous chemical space of over $10^{60}$ conceivable compounds can be filtered to a manageable number that can be synthesized, purchased, and tested?

- **Virtual screening** (VS) is a computational technique used in drug discovery to search libraries of small molecules in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme.

- Thanks to the application of **Machine Learning** algorithms, the accuracy of the method has increased.



# What is Machine Learning?

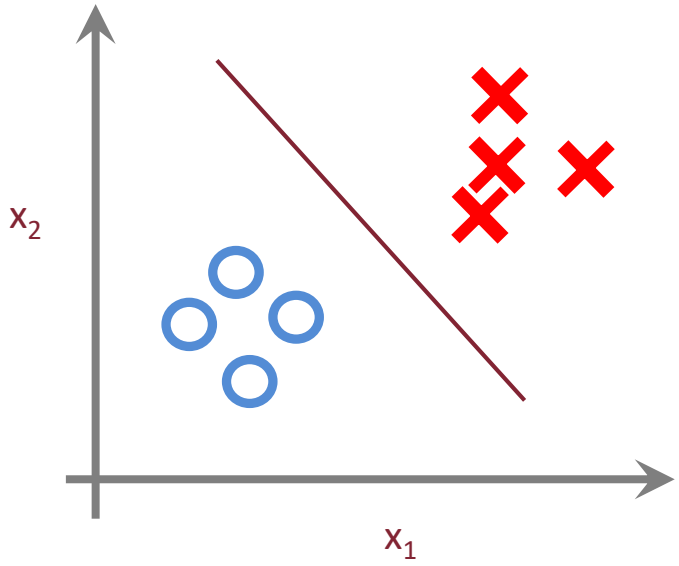


"Field of study that gives computers the ability to learn without being explicitly programmed"
(1959 – Arthur Samuel)

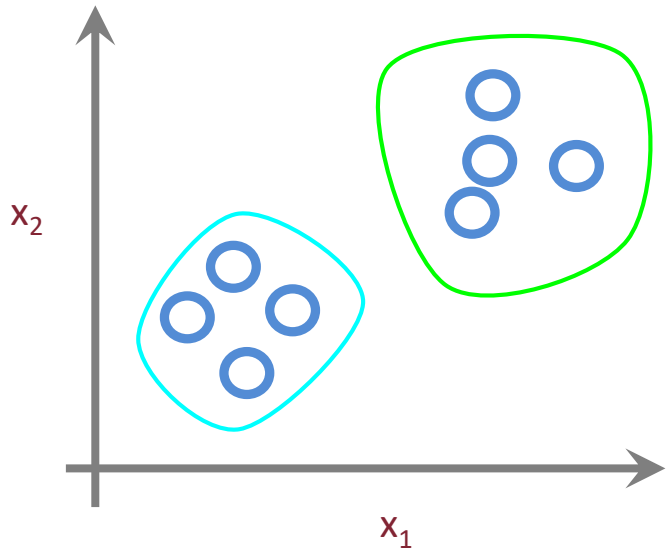
- Evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- Explores the study and construction of algorithms that can learn from and make predictions on data.
- Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Machine learning tasks are typically classified into **two broad categories**, depending on the nature of the learning "signal" or "feedback" available to a learning system.

**Supervised learning:**
The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

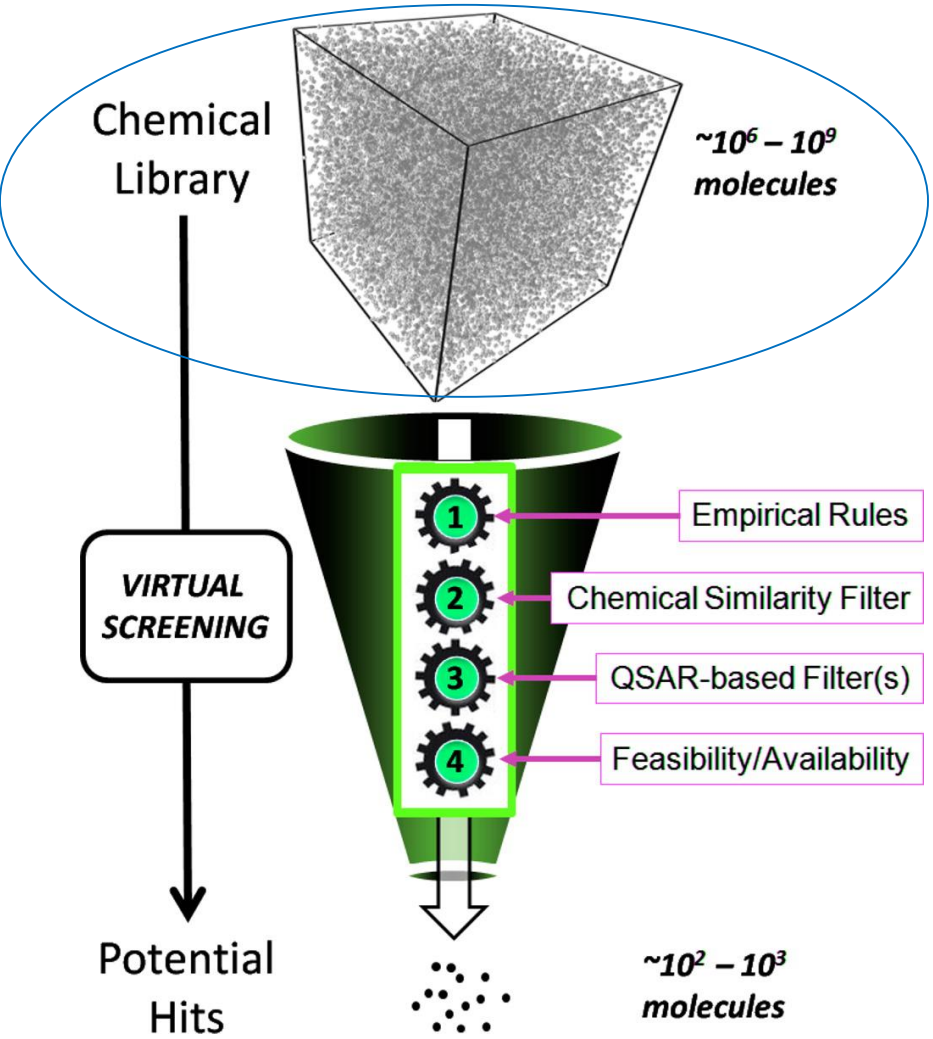



**Unsupervised learning:**
The computer discovers hidden patterns in data.



# Virtual Screening of Chemical Libraries

by www.RCMD.it




PubChem

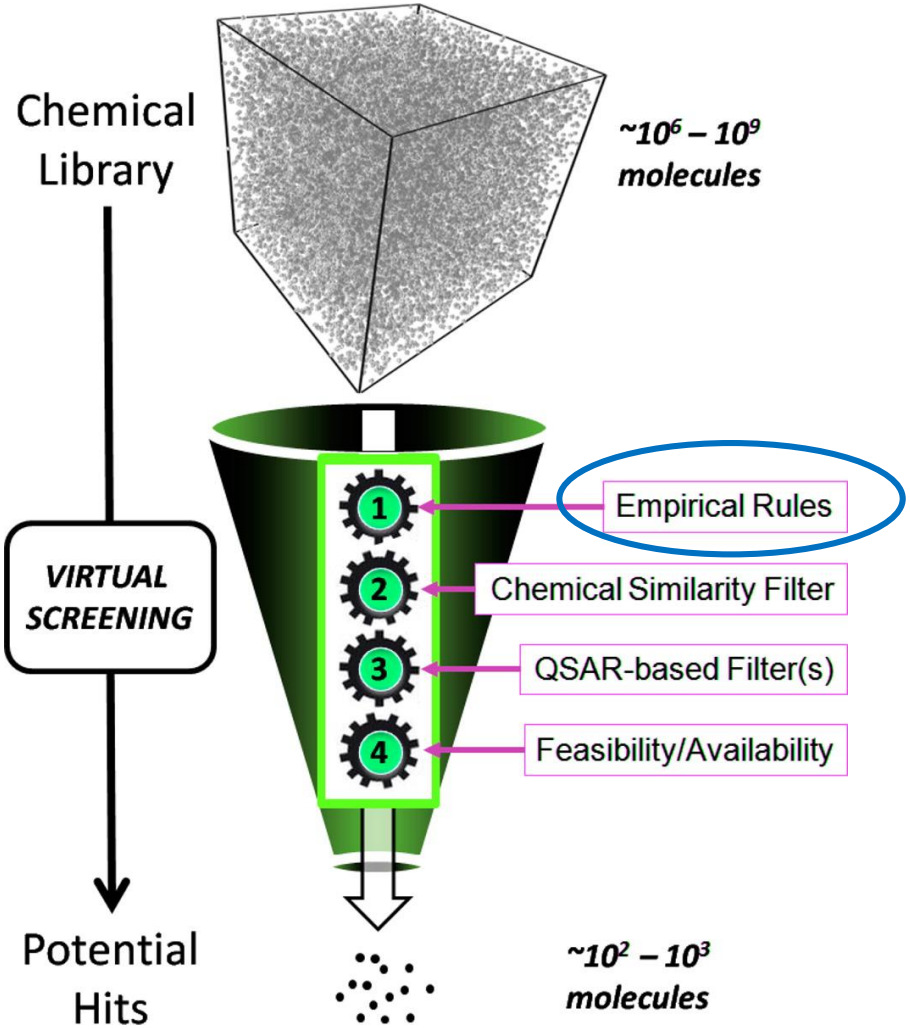
ZINC

ChemSpider
The free chemical database

ChEMBL

CHERKASOV, Artem, et al. QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 2014, 57.12: 4977-5010.

# Screening chemical libraries




**Lipinski rule of five**

MW <= 500

log*P* <= 5

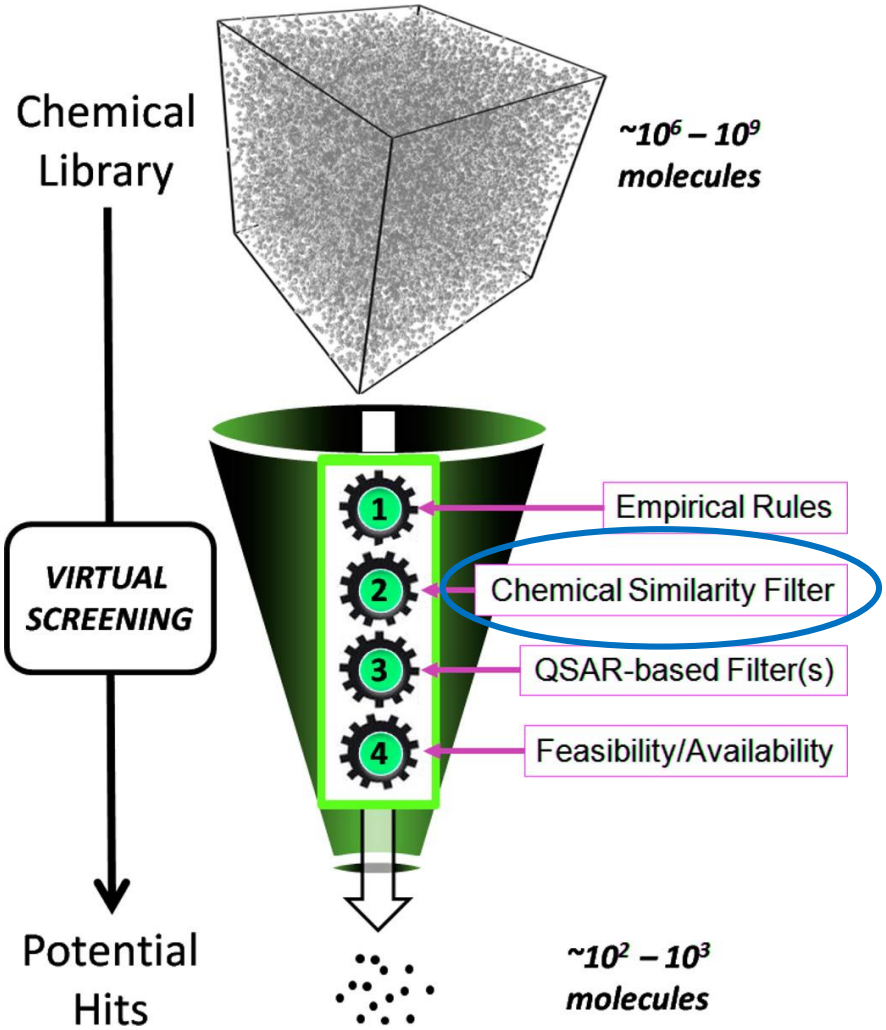
HBD <= 5

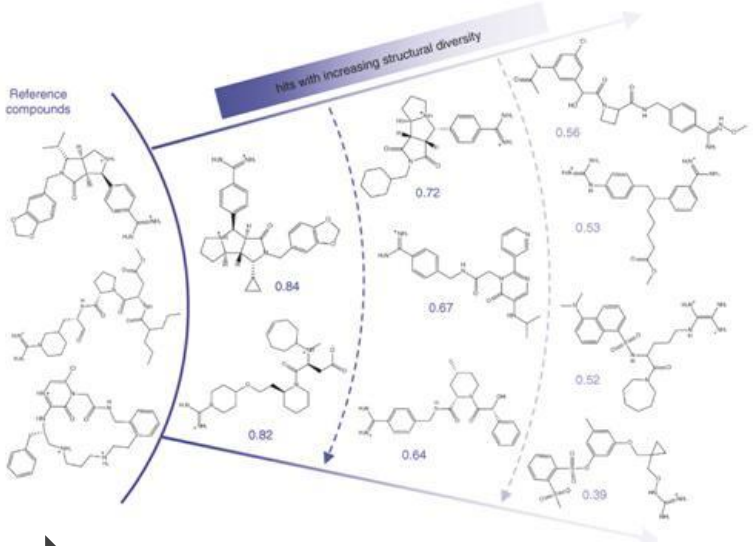
HBA <= 10

CHERKASOV, Artem, et al. QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 2014, 57.12: 4977-5010.

# Screening chemical libraries






Tanimoto similarity

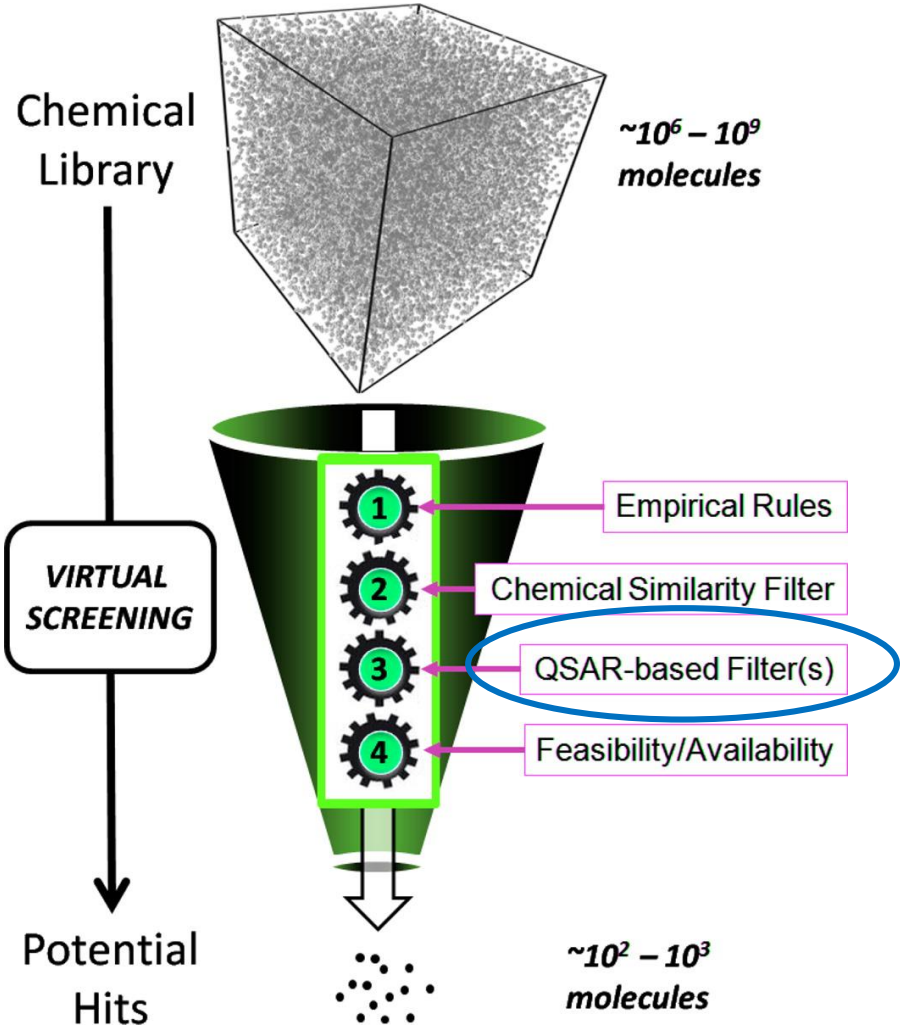
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

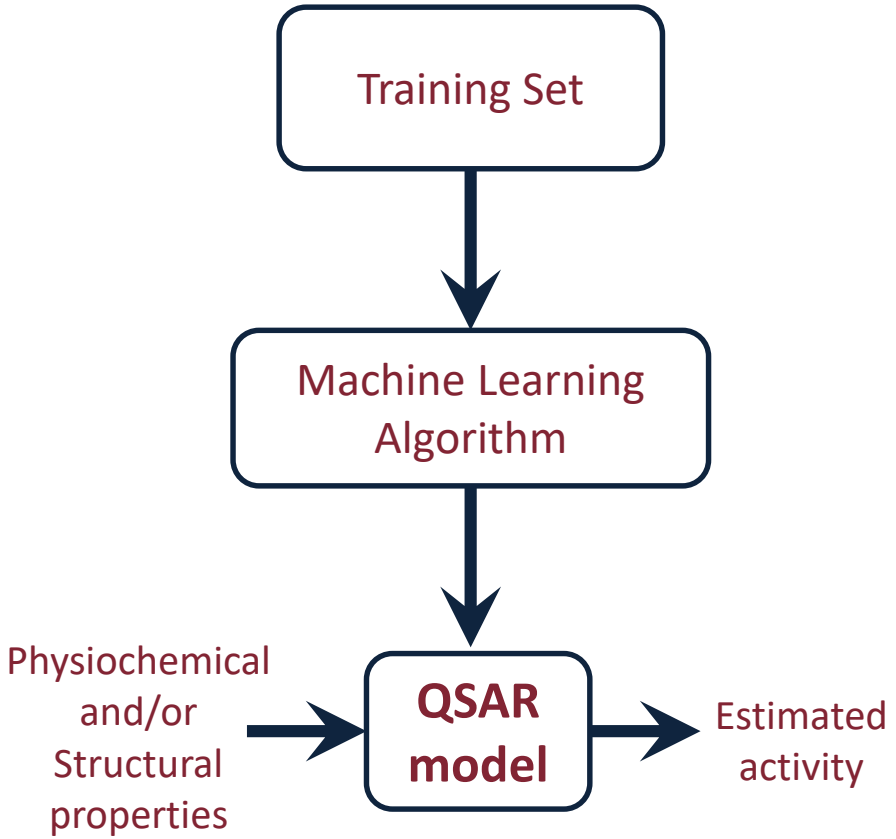
Euclidean distance

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

CHERKASOV, Artem, et al. QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 2014, 57.12: 4977-5010.

# Screening chemical libraries

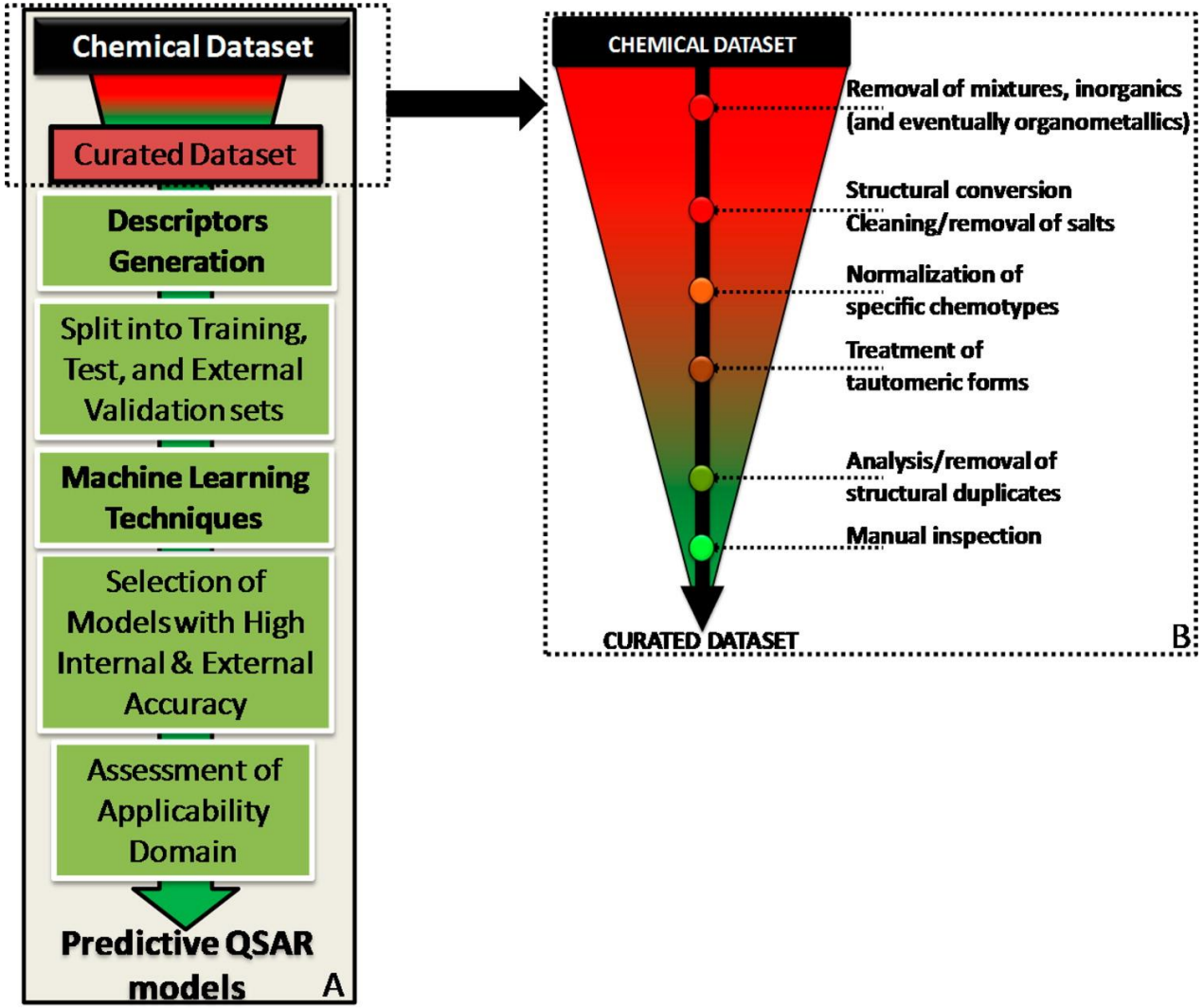


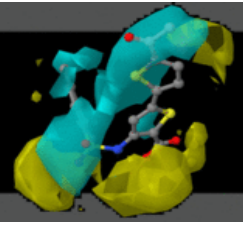




CHERKASOV, Artem, et al. QSAR modeling: where have you been? Where are you going to?. *Journal of medicinal chemistry*, 2014, 57.12: 4977-5010.

# QSAR Work Flow

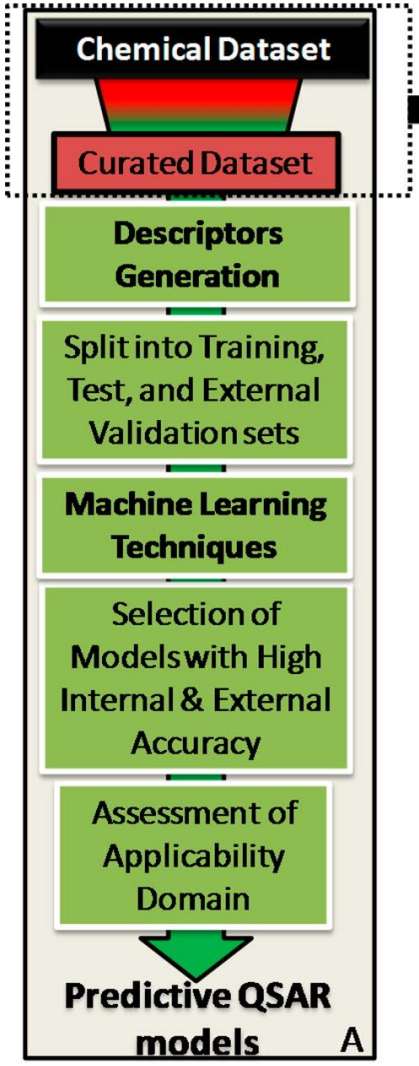


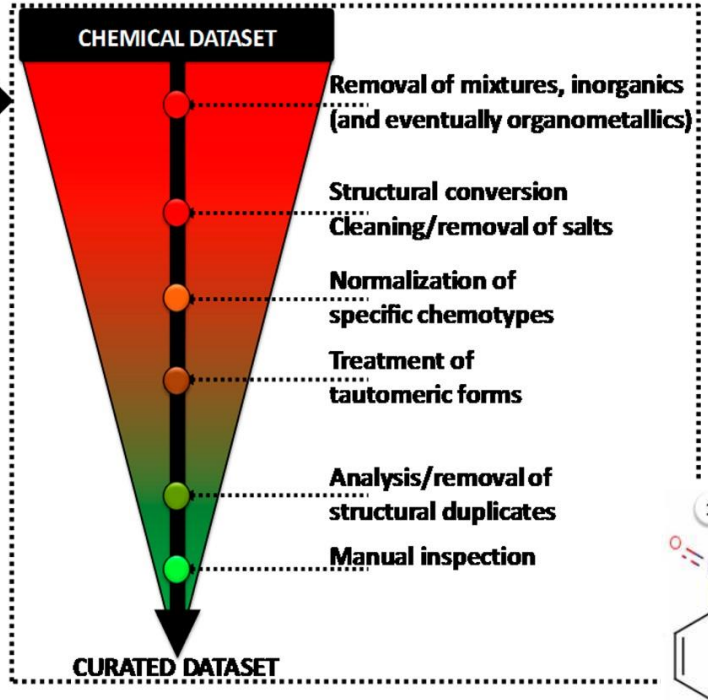




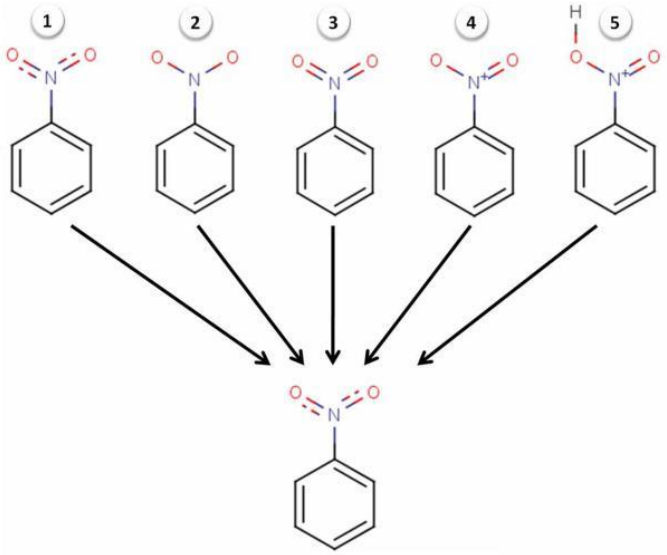
# QSAR Work Flow

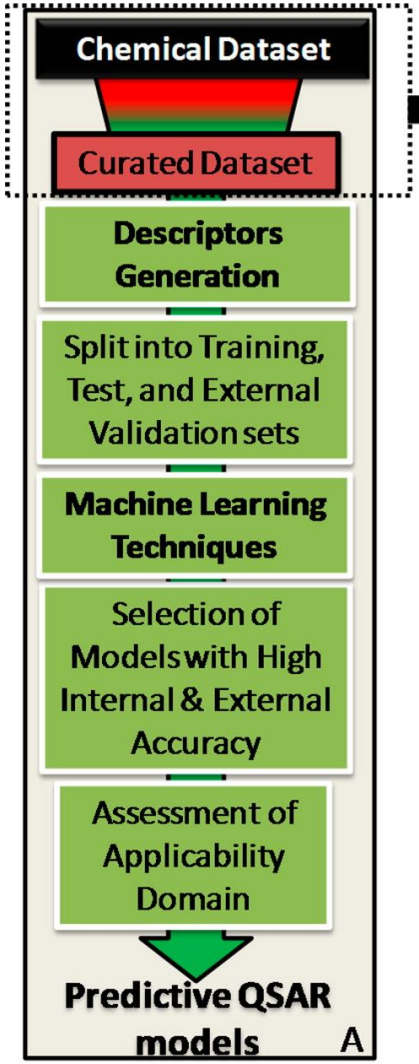
by www.RCMD.it

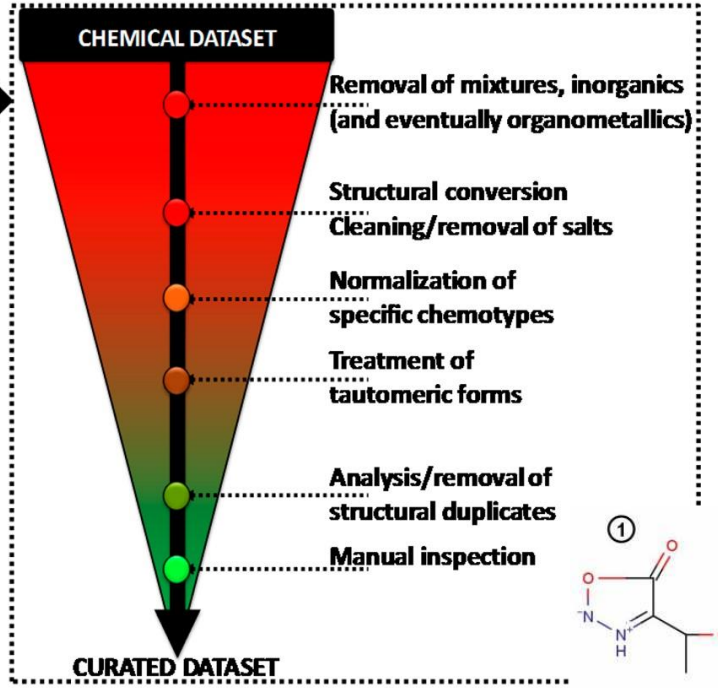


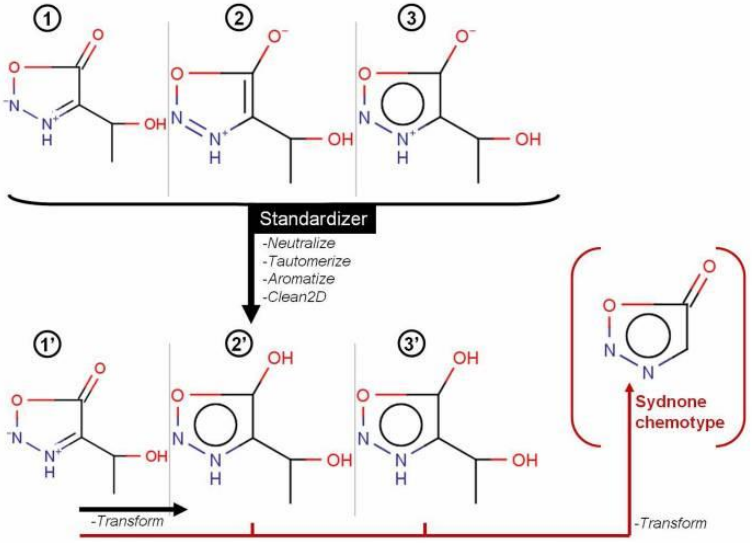

**Normalization** of five types of nitro group representations

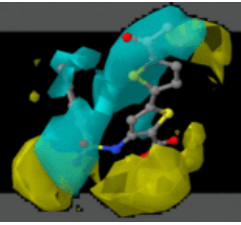




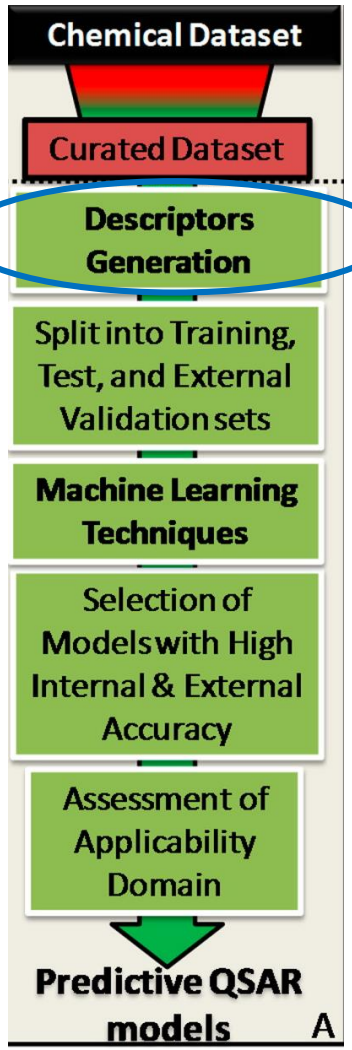

**Normalization** of three compounds possessing the sydnone chemotype





# QSAR Work Flow

by www.RCMD.it




## Molecular Descriptors Representation

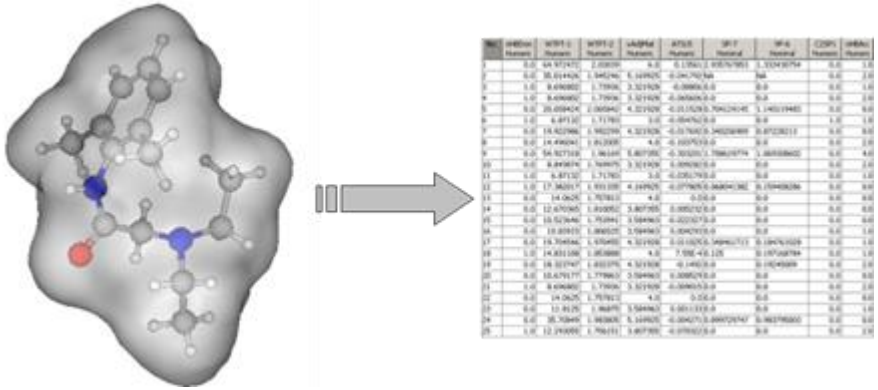
**0D**: bond counts, molecular weight, atom counts

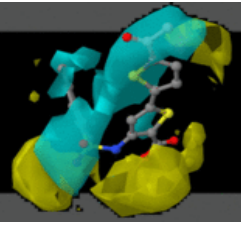
**1D**: fragment counts, H-Bond acc/don, Crippen, PSA, SMARTS

**2D**: topological descriptors (Balaban, Randic, Wiener, BCUT, kappa, chi)

**3D**: geometrical descriptors, surface properties, COMFA

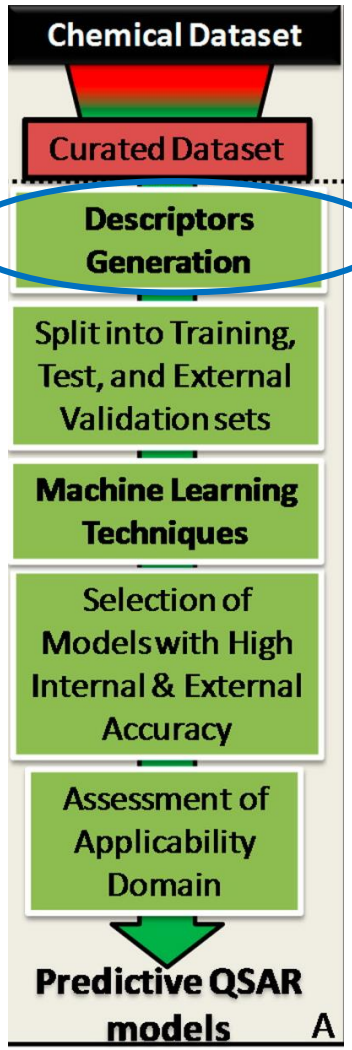
**4D**: 3D coordinates + conformations



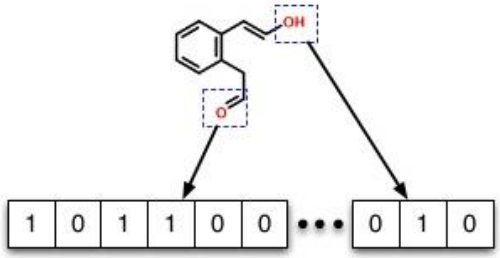


# QSAR Work Flow

by www.RCMD.it



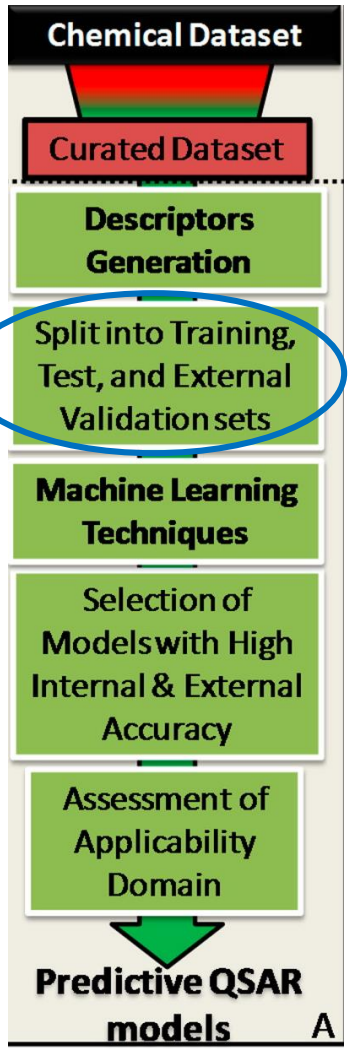

## Fingerprint Representation

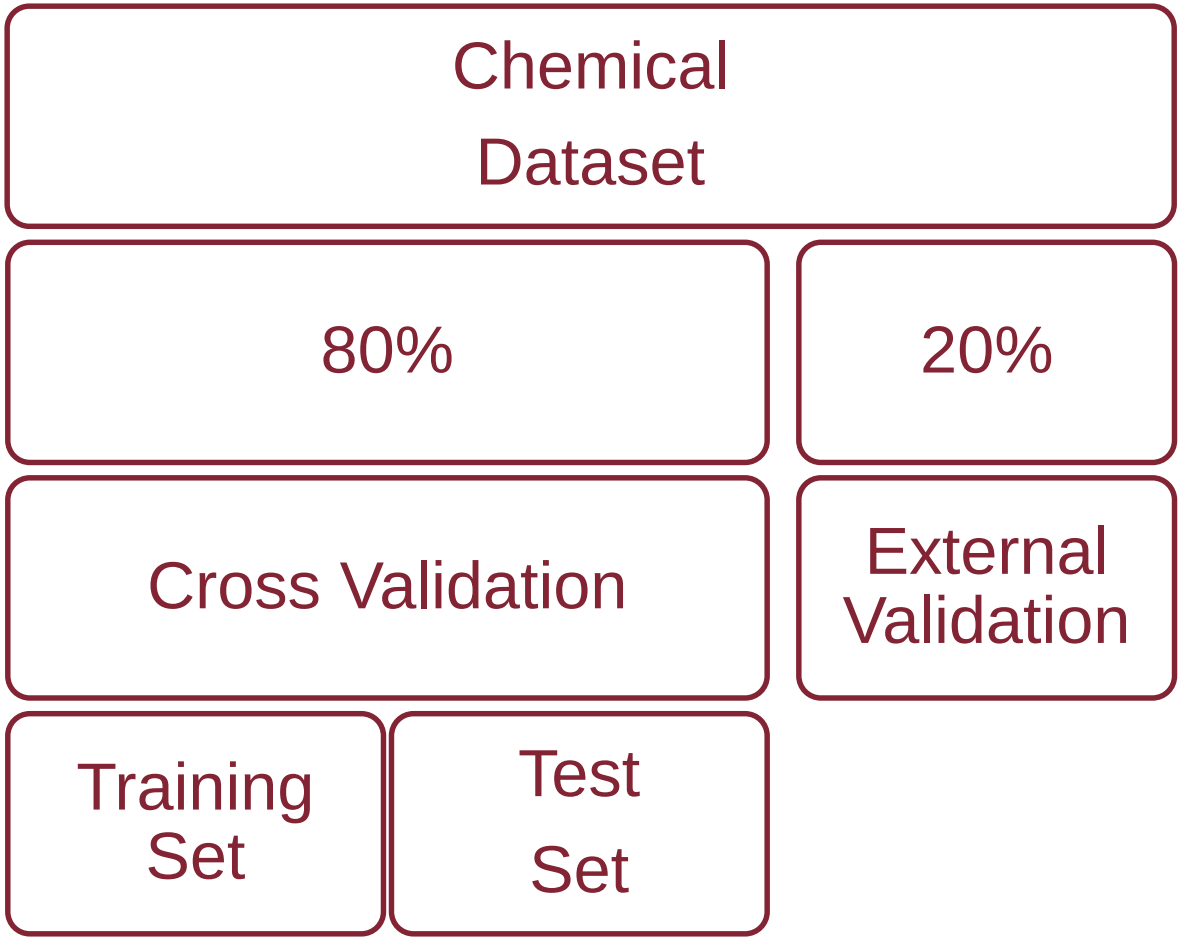


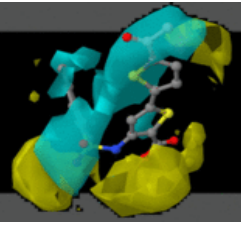

- Lots of types of fingerprints
- "Keyed" fingerprints indicate the presence or absence of a structural feature
- Length can vary from 166 to 4096 bits or more
- Fingerprints usually compared using the Tanimoto metric

# QSAR Work Flow

by www.RCMD.it





# QSAR Work Flow

by www.RCMD.it

Chemical Dataset

Curated Dataset

Descriptors Generation

Split into Training, Test, and External Validation sets

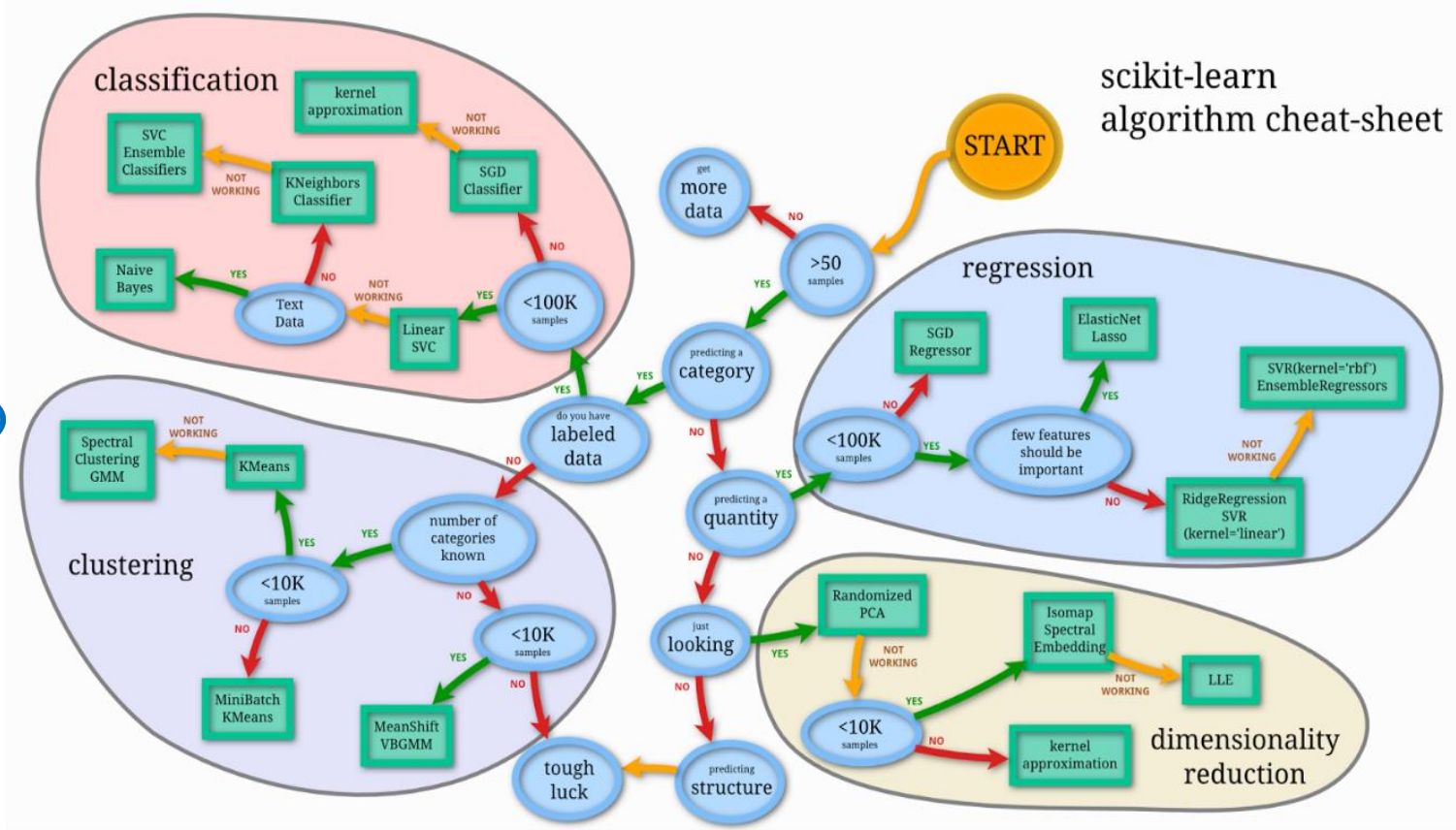
Machine Learning Techniques

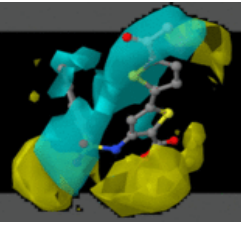
Selection of Models with High Internal & External Accuracy

Assessment of Applicability Domain

Predictive QSAR models

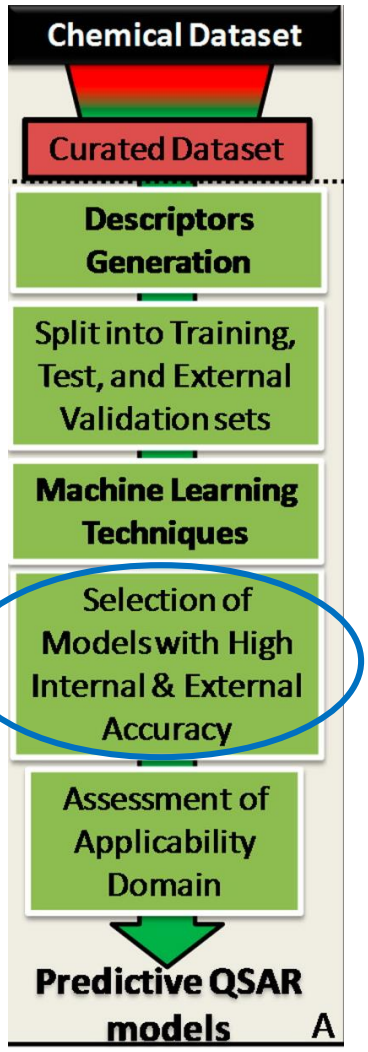
A

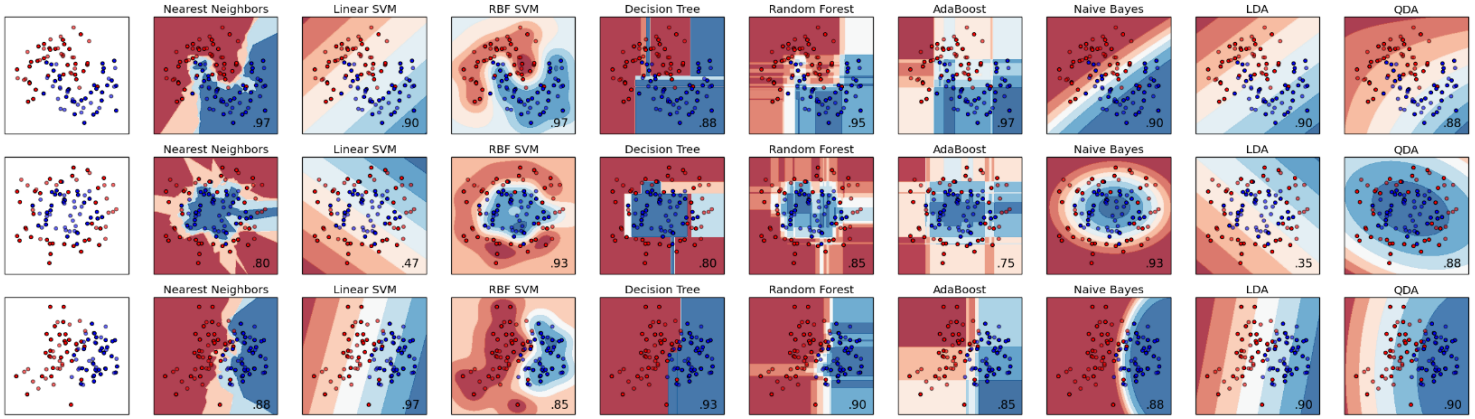




# QSAR Work Flow

by www.RCMD.it




## Regression metrics

$$\texttt{explained\_variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| .$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 .$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$

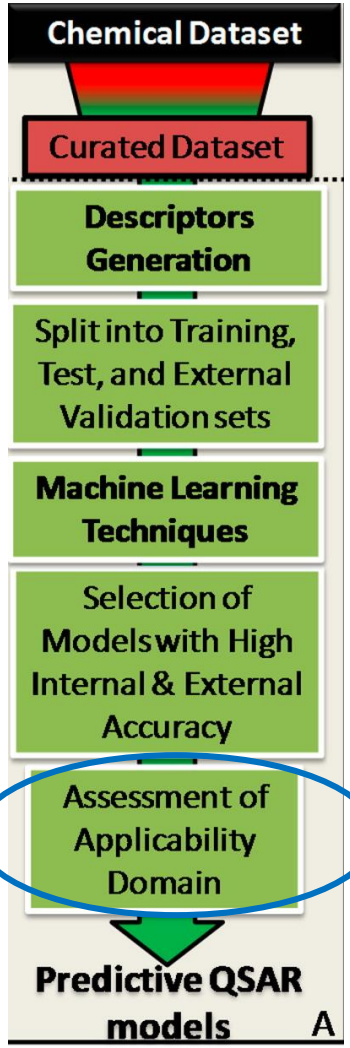
## Classification metrics

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

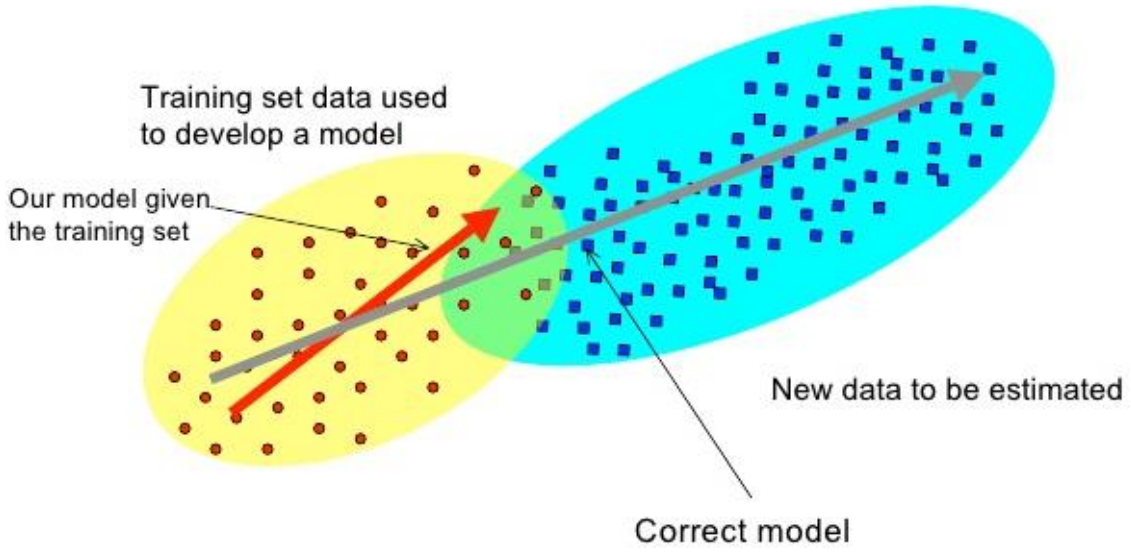
$$\text{precision} = \frac{tp}{tp + fp},$$

$$\text{recall} = \frac{tp}{tp + fn},$$

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}.$$

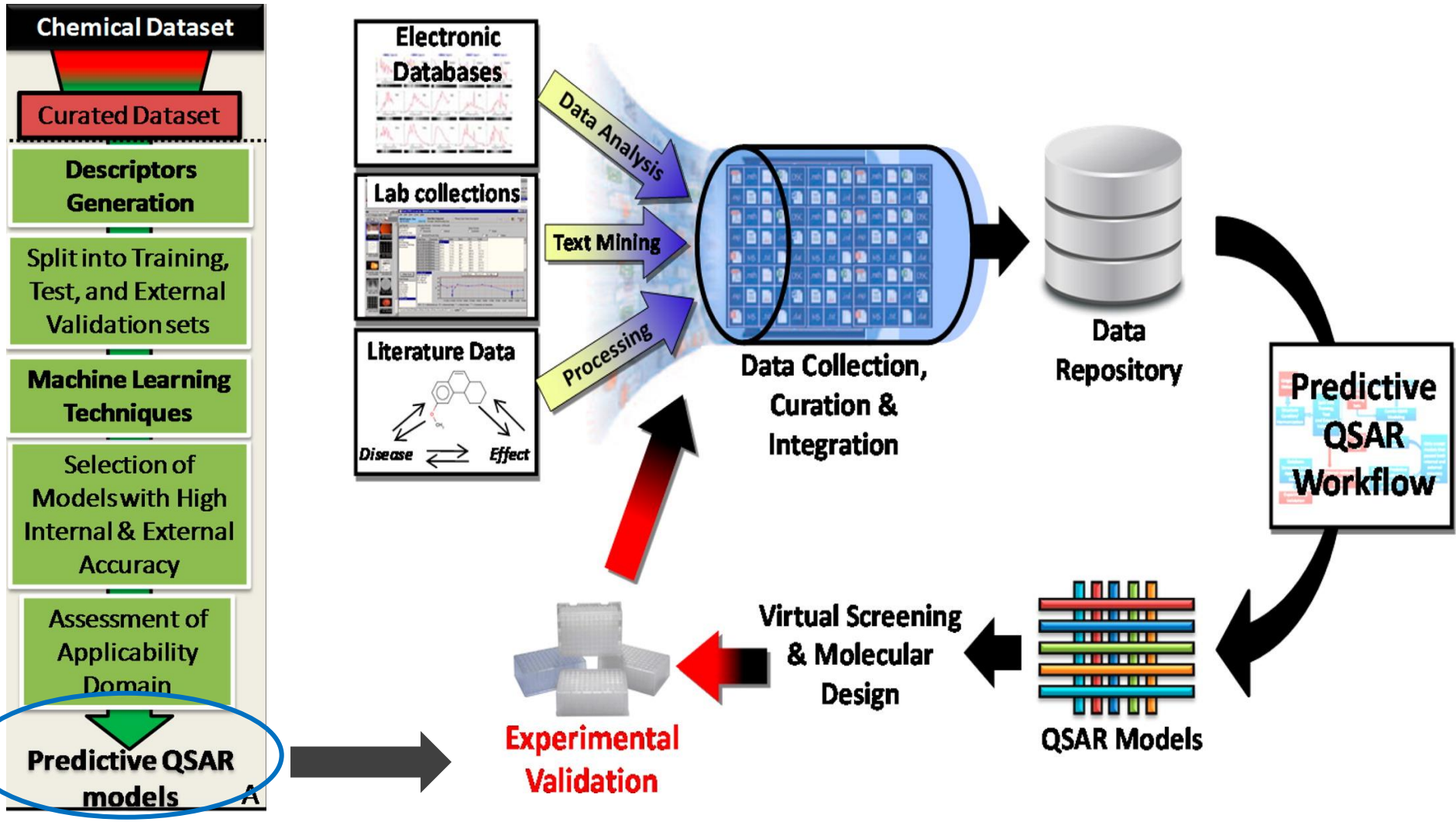


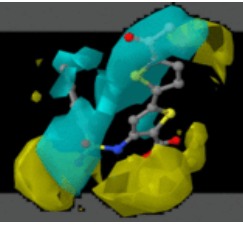

**Models can fail due to chemical diversity of training & test sets (i.e. outside of applicability domain)**



# QSAR in Drug Discovery

by www.RCMD.it





# Molecular Informatics








# Case Study: LSD1



Experimantal Structure Activity Data

Test Set | Training Set

↓

Calculation of Molecular Descriptors

↓

QSAR Model Generation

Model Accepted?

↓

Prediction of Test Set Data

↓

External Validation

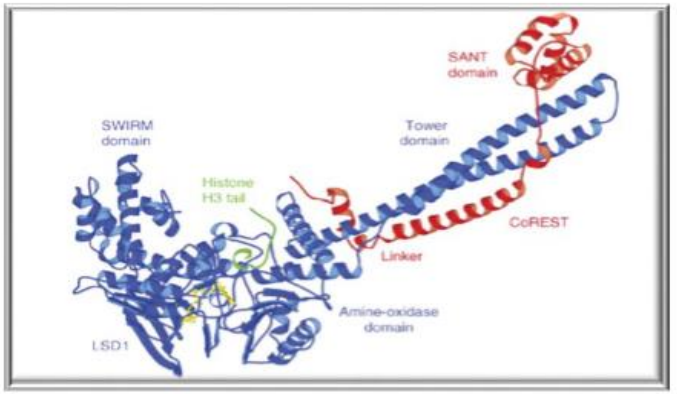
↓

De novo Design and/or Virtual Screening

↓

Experimental Training

# Case Study: LSD1





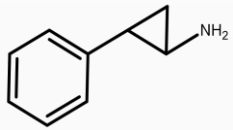


Shi Y et al. Cell 2004, *119*, 941-53. Karytinos A et al. J Biol Chem. 2009, *284*, 17775-82.

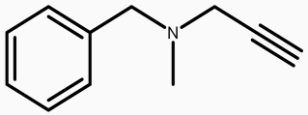
LSD1 functions to demethylate H3K4me1/2. H3K9me1/2. and non-histone proteins including p53. E2F1 and DNMT1. LSD1 is present in different protein complexes. such as the HDAC/CoREST/REST complex and the Mi-2/nucleosome remodeling and deacetylase (NuRD) complex. and displays diverse functions.

# LSD1 Inhibitors

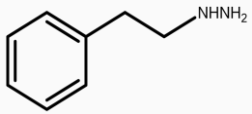




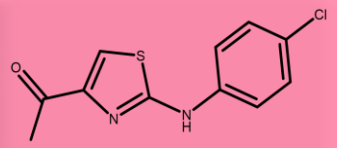

TRANILCIPROMINE



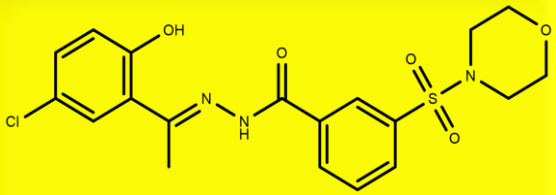

PARGILINE



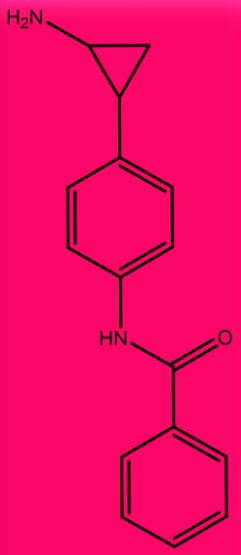

FENELZINE



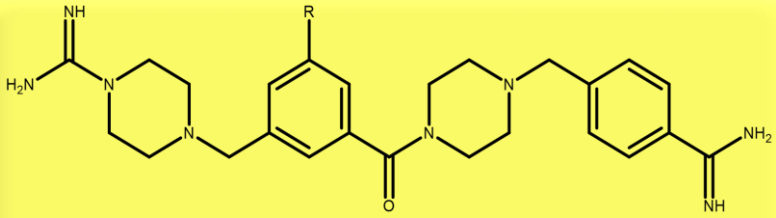

HITCHIN et al.



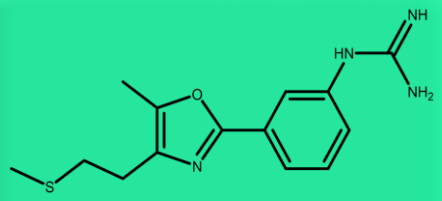

SORNA et al.




MAI et al.



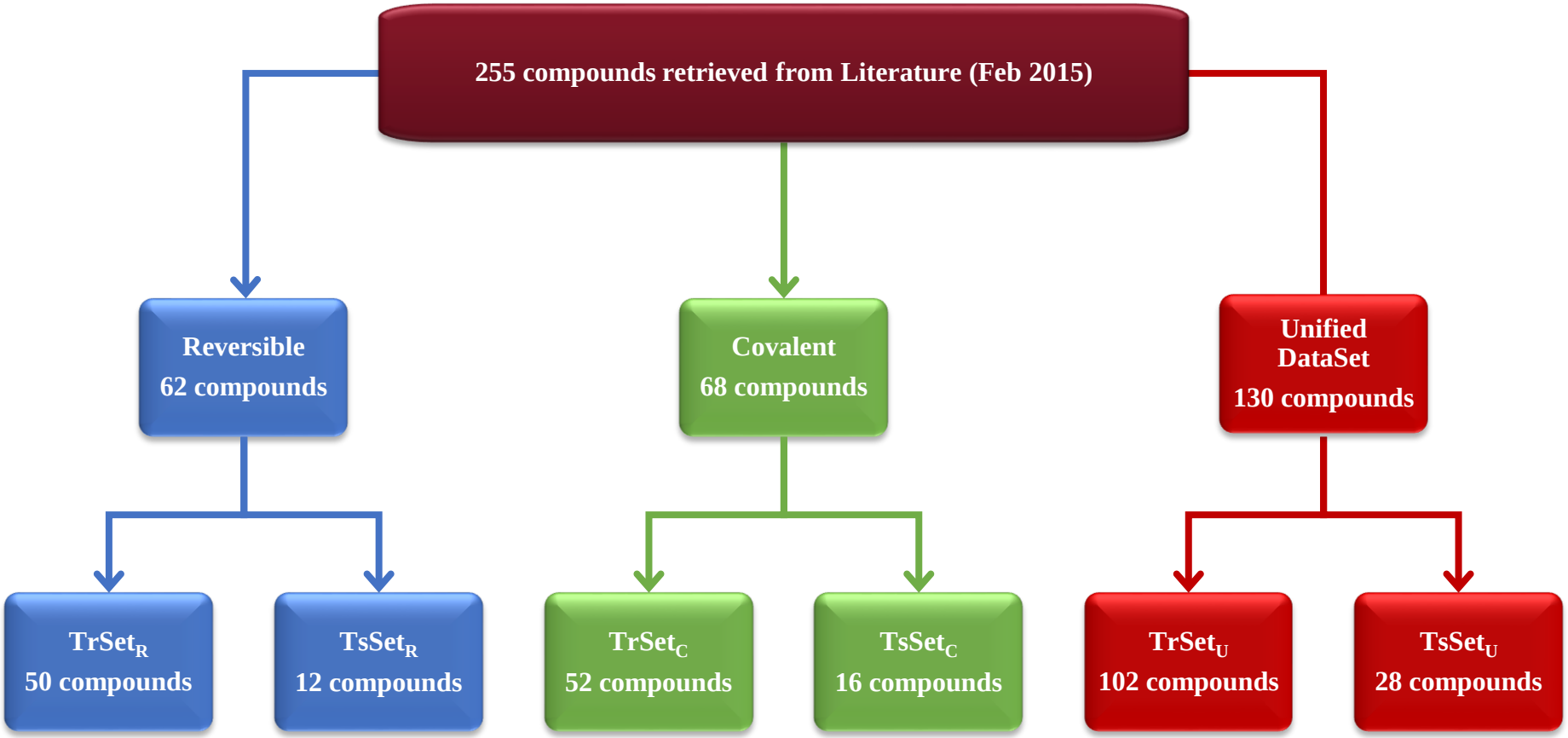

WANG et al.



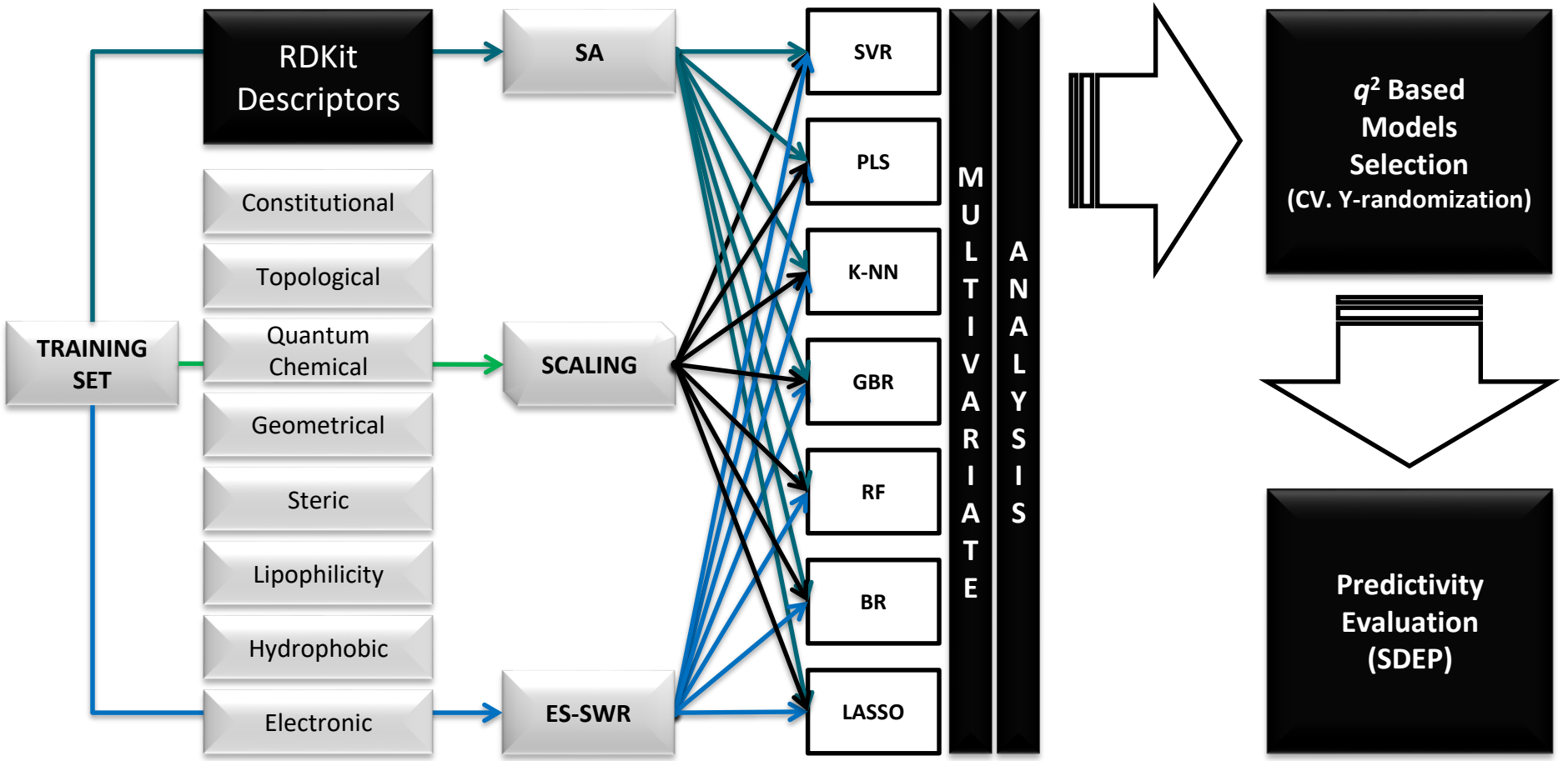

DULLA et al.

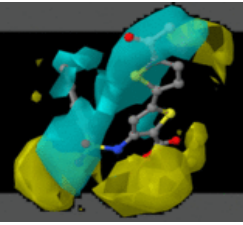
# Data Set Definition





# Models Generation

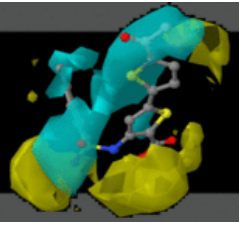




# Results

by www.RCMD.it

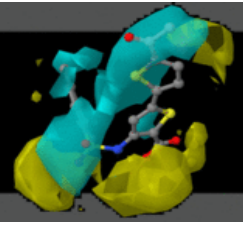
| Data Set | Method | K-fold | | | | Y-scrambling | | | | | | External Validation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $q^2$ | $r^2$ | SDEP | SDEC | Average $q^2$ | Average $r^2$ | Average SDEP | Average SDEC | Maximum $q^2$ | Number of positive $q^2$ | SDEP | AAEP | $r^2_{pred}$ |
| **Reversible** | PLS | 0.73 | 0.90 | 0.92 | 0.44 | -0.65 | 0.48 | 1.60 | 0.91 | 0.02 | 1 | 0.45 | 0.36 | 0.90 |
| | KNN | 0.89 | 0.95 | 0.80 | 0.35 | -0.60 | 0.48 | 1.57 | 0.90 | 0.02 | 1 | 0.42 | 0.37 | 0.92 |
| | BR | 0.89 | 0.96 | 0.93 | 0.28 | -0.08 | 0.06 | 1.30 | 1.22 | 0.16 | 3 | 1.04 | 0.68 | 0.50 |
| | SVR | 0.69 | 0.99 | 0.69 | 0.11 | -0.45 | 0.78 | 1.50 | 0.57 | 0.14 | 2 | 0.78 | 0.63 | 0.70 |
| | GBR | 0.74 | 0.98 | 0.79 | 0.14 | -0.64 | 0.96 | 1.59 | 0.23 | -0.03 | 0 | 0.37 | 0.35 | 0.94 |
| | RF | 0.79 | 0.98 | 0.92 | 0.48 | -0.34 | 0.77 | 1.44 | 0.60 | 0.16 | 8 | 0.63 | 0.52 | 0.90 |
| **Covalent** | PLS | 0.82 | 0.88 | 0.75 | 0.46 | -0.78 | 0.24 | 1.50 | 0.99 | 0.04 | 1 | 0.62 | 0.44 | 0.72 |
| | KNN | 0.46 | 0.84 | 0.83 | 0.46 | -0.53 | 0.50 | 1.40 | 0.81 | 0.02 | 0 | 0.87 | 0.53 | 0.49 |
| | BR | 0.83 | 0.92 | 0.71 | 0.45 | -0.07 | 0.06 | 1.17 | 0.11 | 0.11 | 4 | 1.16 | 0.87 | 0.23 |
| | SVR | 0.79 | 0.99 | 0.78 | 0.11 | -0.95 | 0.90 | 1.57 | 0.35 | -0.11 | 0 | 1.17 | 0.93 | 0.19 |
| | GBR | 0.67 | 0.99 | 0.82 | 0.05 | -0.60 | 0.98 | 1.43 | 0.16 | 0.1 | 3 | 0.77 | 0.45 | 0.58 |
| | RF | 0.56 | 0.88 | 0.91 | 0.37 | -0.35 | 0.78 | 1.31 | 0.54 | 0.37 | 3 | 0.53 | 0.46 | 0.79 |



# Results

by www.RCMD.it

| Unified Data Set | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | K-fold | | | | Y-scrambling | | | | | | External Validation | | |
| | $q^2$ | $r^2$ | SDEP | SDEC | Average $q^2$ | Average $r^2$ | Average SDEP | Average SDEC | Maximum $q^2$ | Number of positive $q^2$ | SDEP | AAEP | $r^2_{pred}$ |
| **PLS** | 0.69 | 0.79 | 0.82 | 0.61 | -0.36 | 0.29 | 1.42 | 1.03 | 0.01 | 1 | 0.89 | 0.67 | 0.69 |
| **KNN** | 0.69 | 0.82 | 0.88 | 0.56 | -0.36 | 0.30 | 1.43 | 1.01 | -0.10 | 0 | 0.97 | 0.66 | 0.42 |
| **BR** | 0.75 | 0.85 | 0.99 | 0.69 | -0.03 | 0.04 | 1.24 | 1.20 | 0.07 | 5 | 1.58 | 0.85 | 0.17 |
| **SVR** | 0.77 | 0.99 | 0.74 | 0.11 | -0.92 | 0.90 | 1.69 | 0.38 | -0.11 | 0 | 1.06 | 0.95 | 0.34 |
| **GBR** | 0.77 | 0.99 | 0.77 | 0.10 | -0.43 | 0.46 | 0.93 | 0.31 | -0.06 | 0 | 0.74 | 0.59 | 0.65 |
| **RF** | 0.63 | 0.92 | 0.84 | 0.39 | -0.04 | 0.03 | 1.29 | 1.21 | 0.02 | 4 | 0.69 | 0.52 | 0.73 |



# Conclusions and Perspectives

by www.RCMD.it

- QSAR Models for Covalent and Reversible LSD1 Inhibitors have been built
- The Unified Model allows a fast selection of potential inhibitors regardless their mechanism of action.
- Selected molecules inspected for their chemical reactivity could be split into likely covalent and reversible inhibitor and their predicted activity confirmed with the «ad hoc» models
- A virtual screening on the ZINC database is ongoing and soon a set of molecules will be sent for biological assay