# A decision-analytic QSAR model for planning cannabinoid discovery activities

ICMAT DataLab

ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

# Cannabinoid-based drug discovery

Drugs to target human cannabinoid system:

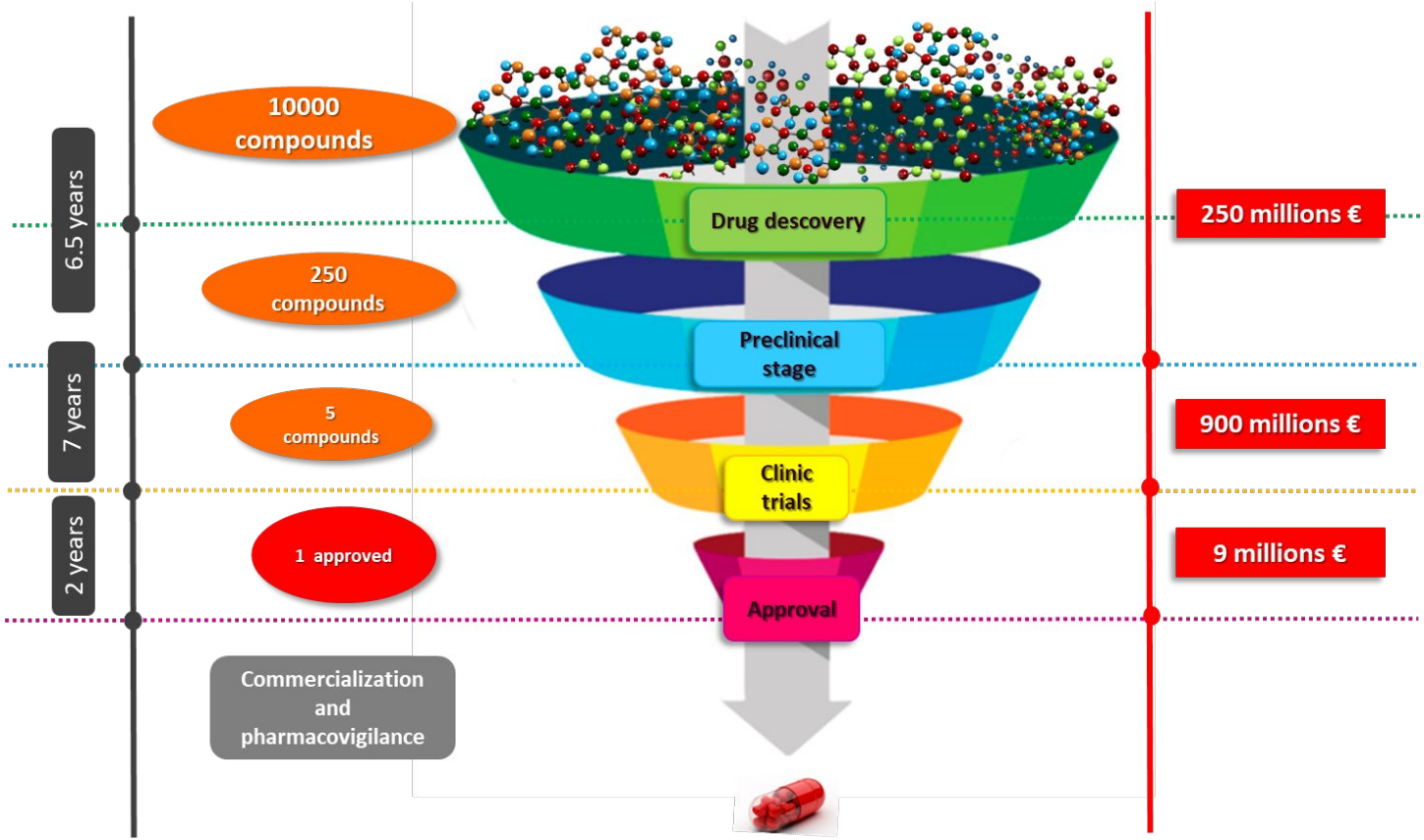
- CB1 receptor:
  - Psychotropic effects

- CB2 receptor (CBR2):
  - Lack of CB1 receptor negative effects
  - Involves interesting biological pathways

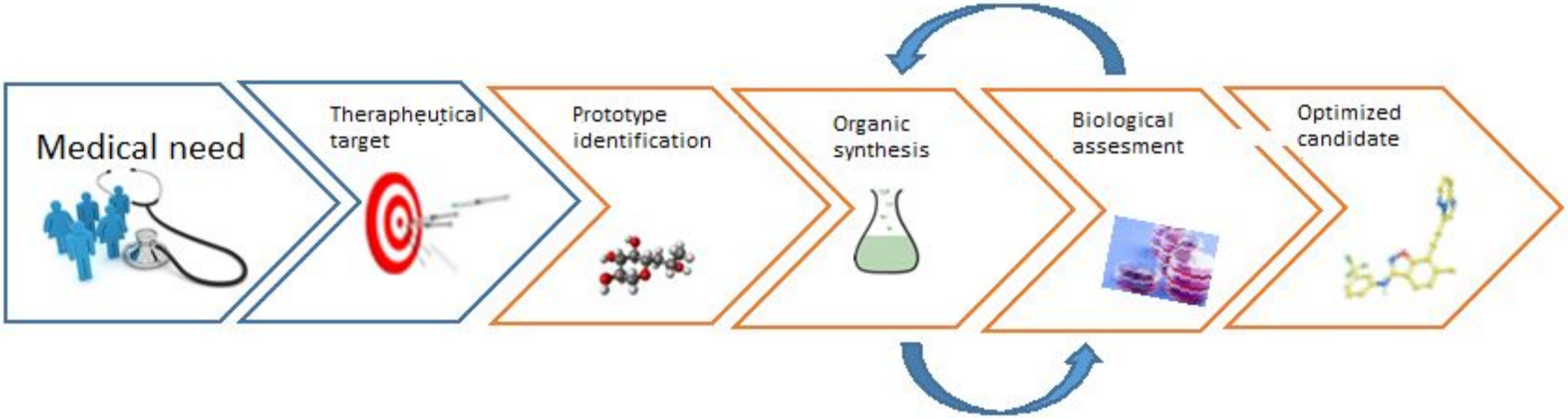
<u>Goal</u>: Ligand discovery with certain properties that target CB2R

→ But entails high research and development costs !

# Drug development



# Traditional drug development



# Objective

QSAR strategy implementation to identify CB2R. It consists of:

- Predictive stage
  - Combines 3 classifiers
  - Forecast behavior/activity properties

- Decision stage
  - Utility model
  - Consider costs/ benefits of design decisions.

# Data collection

- CBR2 ligands from ChEMBL public database:
  - Behaviour: Agonist/ Antagonist
  - $EC_{50}$ bioactivity value
    - Inactive: $EC_{50}$ >= 10 µM
    - Moderately active: 0.01 µM < $EC_{50}$ < 10 µM
    - Active: $EC_{50}$ <= 0.01 µM Active (Active)
- Compound x $\in$ {AgAct, AgMod, AgIn, AntAct, AntMod, AntIn }

# Data description

We split the data:

- $X_{internal}$ (90%) ⇨ Predictive stage
- $X_{external}$ (10%) ⇨ Decision stage

| Class | $X_{internal}$ | $X_{external}$ | Total |
|---|---|---|---|
| AgAct | 360 | 43 | 403 |
| AgMod | 899 | 103 | 1002 |
| AgIn | 93 | 7 | 100 |
| AntAct | 18 | 1 | 19 |
| AntMod | 105 | 7 | 112 |
| AntIn | 41 | 8 | 49 |

Compounds represented by:

- Mordred (997 features)
- BERT (787 features)

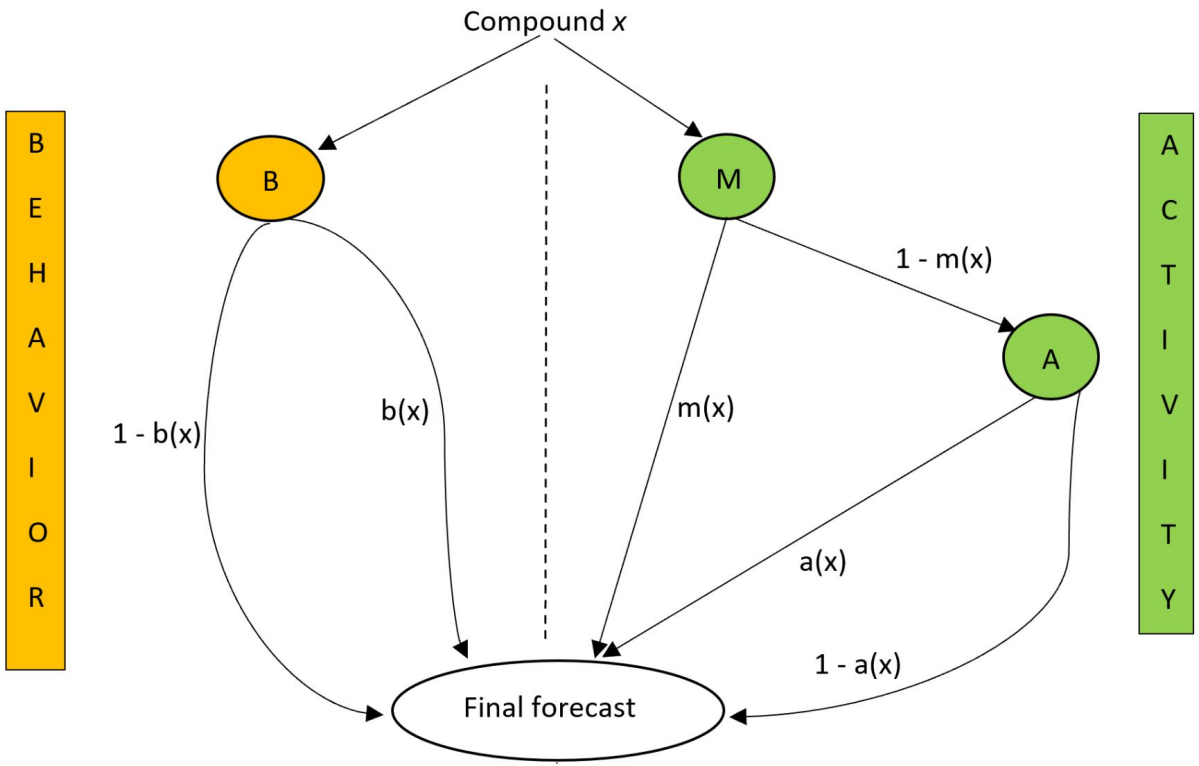
Bigger number of features than #c ompounds !!

# Predictive stage

3 classifiers to forecast compound *x* properties:

- Behaviour:
  - Model B: Agonists ( $y_1$ ) vs Antagonists ( $y_2$)
    - $b(x) = p(y = y_1|x) \Longrightarrow 1 - b(x) = p(y = y_2|x)$
- Activity:
  - Model M: Mod. Active ( $w_1$ ) vs Not mod. Active ( $w_2$ )
    - $m(x) = p(y = w_1|x) \Longrightarrow 1 - m(x) = p(y = w_2|x)$
  - Model A: Active ( $z_1$) vs Inactive( $z_2$)
    - $a(x) = p(y = z_1|x, w_2) \Longrightarrow 1 - a(x) = p(y = z_2|x, w_2)$

# Predictive stage pipeline




Example: $P(x = AgAct) = b(x)(1-m(x))a(x)$

# Calibration task

Focus on probability precision, not predicted label.

- Accuracy, recall, precision, F1- score not appropriate
- Alternatives:
  - Stratified Brier Score(*BS*).

$$BS^{+} = \frac{\sum_{y_i=1} \left(y_i - f(y_i \mid x_i)\right)^2}{N_{pos}} \qquad BS^{-} = \frac{\sum_{y=0} \left(y_i - f(y_i \mid x_i)\right)^2}{N_{neg}}$$

    - Best value: 0 - Worst value: 1
  - ROC- AUC.
    - Best value: 1 - Worst value: 0

# Imbalanced problem

- Split data:
  - $X_{train}$ (80%): Hyperparametrization + training
  - $X_{test}$ (20%): Performance assessment

| Stage | $X_{train}$ | | $X_{test}$ | | Total | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 0 | 1 | 0 | 1 |
| B | 1085 | 127 | 267 | 37 | 1352 | 164 |
| M | 803 | 409 | 201 | 103 | 1004 | 512 |
| A | 304 | 105 | 74 | 29 | 378 | 134 |

B: Agonists (1) vs Antagonists (0)
M: Mod. active (1) vs No Mod. active (0)
A: Inactive (1) vs Inactive (0)

- Imbalance data for each stage. To handle it:
  - Metric
  - Undersampling + Bag classifier

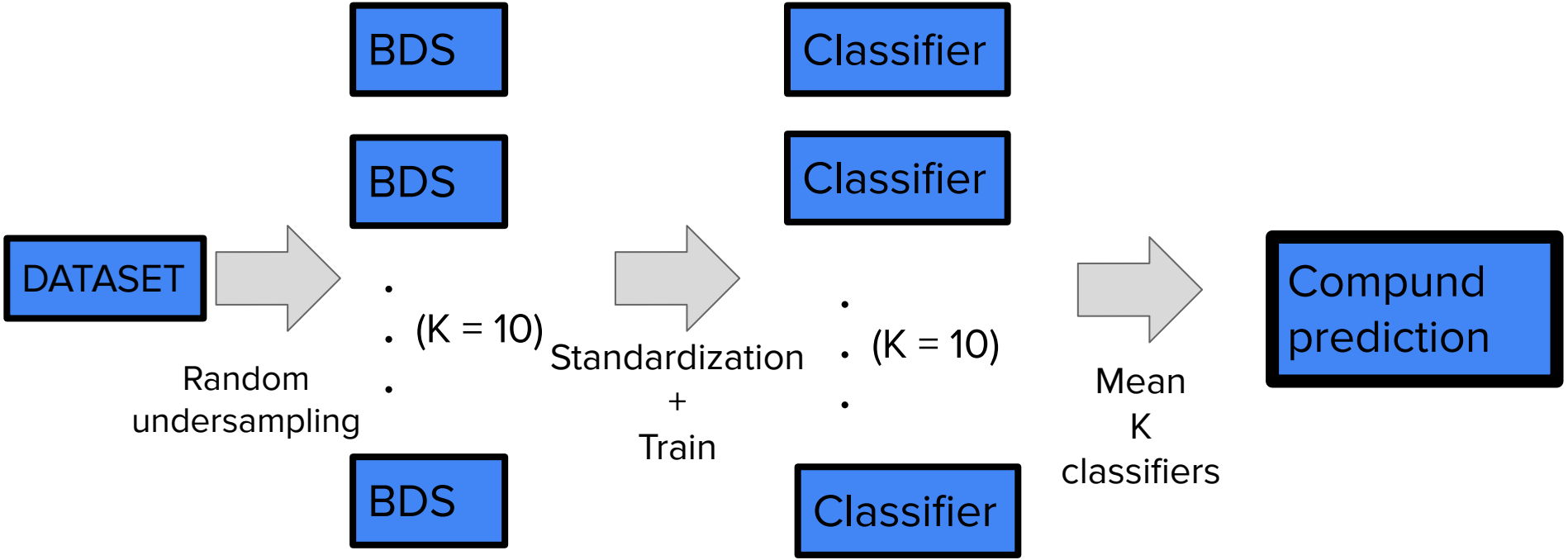
# Undersampling

M compounds in majority class; N in minority

- Random undersampling
  - Consider all compounds from minority class and randomly select N from majority class
  - Way to obtain a balanced dataset

# Bag classifier

BDS = Balanced data subset



# Model performance

- Classifiers: *KNN, Naive Bayes, AdaBoost, Gradient Boosting, Random Forest, Logistic Regression, SVM*

| Model | Feature | Classifier | BS + | | BS - | | ROC-AUC | |
|---|---|---|---|---|---|---|---|---|
| | | | Train | Test | Train | Test | Train | Test |
| B | BERT | Logistic Regression (l1 reg.) | 0.03 | 0.08 | 0.08 | 0.09 | 0.99 | 0.96 |
| M | Mordred | Random Forest | 0.19 | 0.23 | 0.21 | 0.22 | 0.8 | 0.69 |
| A | BERT | Logistic Regression (l1 reg.) | 0.09 | 0.09 | 0.15 | 0.17 | 0.93 | 0.92 |

# Decision stage

Given a compound x, select one of the actions: {synthesize, keep in portfolio,reject}.

| | Active Agonist | Moderate Agonist | Inactive Agonist | Active Antagonist | Moderate Antagonist | Inactive Antagonist |
|---|---|---|---|---|---|---|
| Synthesize | Best | Moderately good | Worst | Very good | Good | Worst |
| | User synthesizes compound. Great opportunity to identify hit and start med. chem. program. Hit to lead compound. | User starts a med. chem. program to improve the compound and possibly find a new family. | User synthesizes or study an inactive compound. Loss of time and money. | User starts new research line, quickly obtaining results to continue drug development. | User possibly starts new research line. | User synthesizes or study an inactive compound. Loss of time and money. |

Depending on the compound type, the action is more or less appropriate.

# Utility assessments

We identify ten different situations and codify them from best to worse:

$$1 > u_1 > u_2 > u_3 > u_4 > u_5 > u_6 > u_7 > u_8 > 0$$

| | | Active Agonist | Moderate Agonist | Inactive Agonist | Active Antagonist | Moderate Antagonist | Inactive Antagonist |
|---|---|---|---|---|---|---|---|
| | Synthesize | 1 | $u_3$ | 0 | $u_1$ | $u_2$ | 0 |
| **Decision** | Keep in portfolio | $u_5$ | $u_4$ | $u_7$ | $u_4$ | $u_4$ | $u_7$ |
| | Reject | $u_8$ | $u_7$ | 1 | $u_6$ | $u_5$ | 1 |

# Represent different preferences

| | | Active Agonist | Moderate Agonist | Inactive Agonist | Active Antagonist | Moderate Antagonist | Inactive Antagonist |
|---|---|---|---|---|---|---|---|
| | Synthesize | 1 | $u_3$ | 0 | $u_1$ | $u_2$ | 0 |
| **Decision** | Keep in portfolio | $u_5$ | $u_4$ | $u_7$ | $u_4$ | $u_4$ | $u_7$ |
| | Reject | $u_8$ | $u_7$ | 1 | $u_6$ | $u_5$ | 1 |

| Utility | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ | $u_8$ |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{U}_1$ | 0.1 | 0.09 | 0.08 | 0.07 | 0.06 | 0.05 | 0.02 | 0.01 |
| $\mathcal{U}_3$ | 0.99 | 0.98 | 0.97 | 0.09 | 0.06 | 0.05 | 0.02 | 0.01 |

$\mathcal{U}_1$ : Focus on determining active agonists

$\mathcal{U}_3$ : Active and moderately active agonists are sought for

# Expected utility

Compute each expected utility for each action $i \in$ {synthesize,
keep in portfolio,
reject}.

$$\psi(i|x) = \sum_{j=1}^{6} u_{ij} p(j|x)$$

j $\in$ {AgAct, AgMod, AgIn,
AntAct, AntMod, AntIn }

Choose action with the maximum expected utility:

$$i^*(x) = \arg\max_i \psi(i|x)$$

# Results

$\mathcal{U}_1$

| Decision | Compound type | | | | | |
|---|---|---|---|---|---|---|
| | AgAct | AgMod | AgIn | AntAct | AntMod | AntIn |
| Synthesize | 37.2±(0.75) | 53.1±(2.02) | 2.0±(0.63) | 0.0±(0.0) | 1.6±(0.49) | 0.7±(0.78) |
| Keep in portfolio | 0.0±(0.0) | 0.0±(0.0) | 0.0±(0.0) | 0.0±(0.0) | 0.0±(0.0) | 0.0±(0.0) |
| Reject | 5.8±(0.75) | 49.9±(2.02) | 5.0±(0.63) | 1.0±(0.0) | 5.4±(0.49) | 7.3±(0.78) |
| Total compounds | 43 | 103 | 7 | 1 | 7 | 8 |

# Discussion

- QSAR model to speed cannabinoid-based drug discovery.
- Consider compounds properties and costs/ benefits of design decisions.
- Significant profits in terms of time and money.