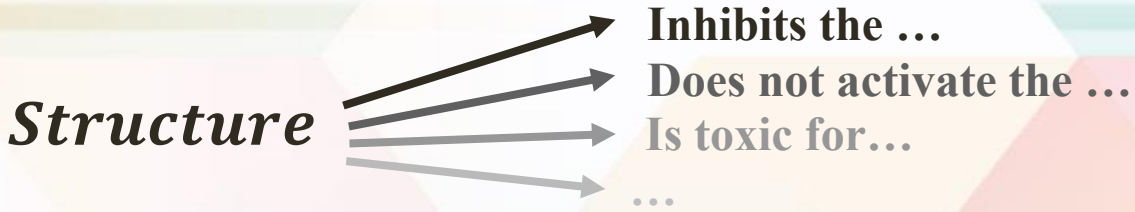
# Self Consistent Classifier SAR Approach

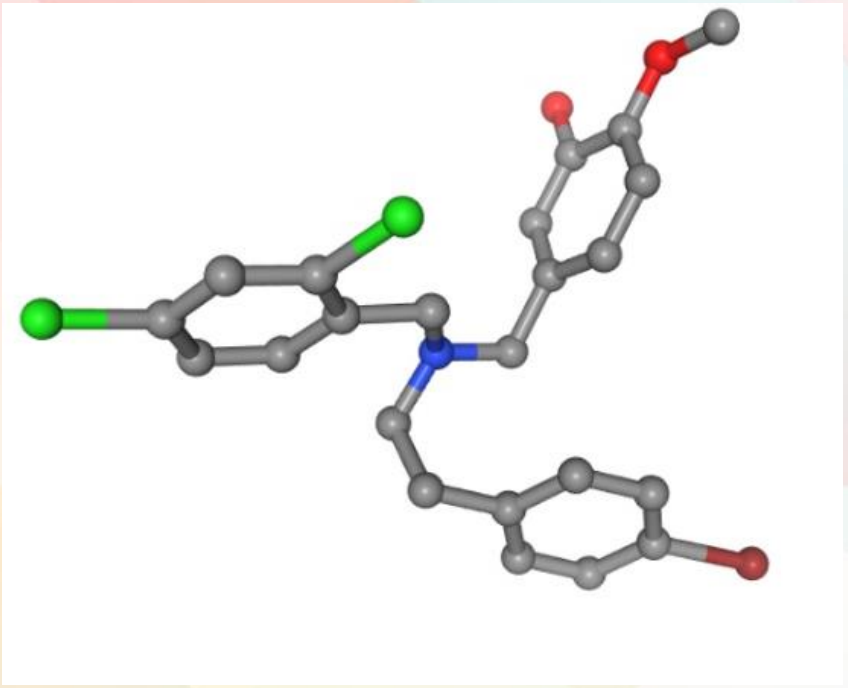
**LEONID STOLBOV, DMITRY FILIMONOV, VLADIMIR POROIKOV**

IBMC, Moscow, 2022

E-mail: stolbovla@yandex.ru

# Analysis of structure-activity relationship - classification



The SAR approach significantly depends on the available experimental data utilized as the training set. To overcome the issues with data quality and diversity, classification methods are used to designate active and inactive structures.

Though a number of algorithms have been developed to satisfy the activity classification needs, the ability of the model to produce a generalized predictions is still challenging.

**In this study we present a new method for classifying chemical compounds by their activity based on the statistical regularization**

# Forerunners

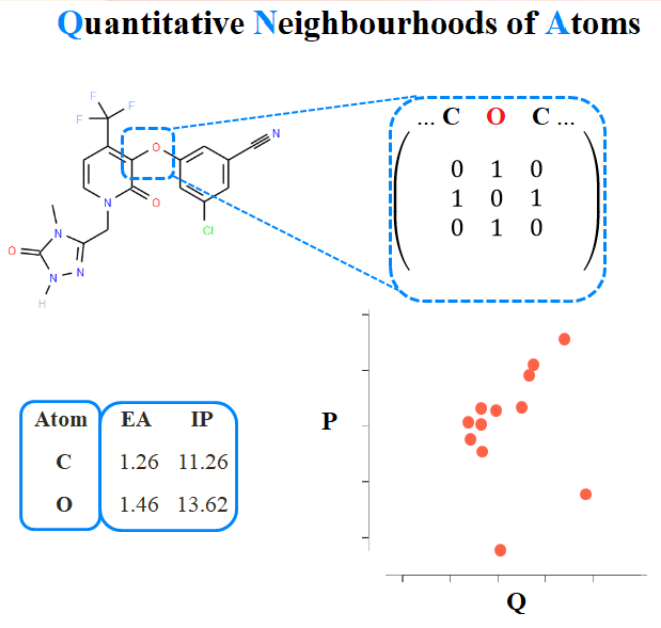
**GUSAR software** was designed to develop (Q)SAR models using **QNA descriptors** and **Self Consistent Regression (SCR)** approach

QNA descriptors are calculated based on connectivity matrix $\boldsymbol{C}$ and the values of ionization potential $IP$ and electron affinity $EA$.

Each atom is represented by two values:

$$P_i = B_i \sum_k \left( \mathrm{Exp}\left(-\frac{1}{2}\boldsymbol{C}\right) \right)_{ik} \frac{1}{2}(IP_k - EA_k)^{-\frac{1}{2}}$$

$$Q_i = B_i \sum_k \left( \mathrm{Exp}\left(-\frac{1}{2}\boldsymbol{C}\right) \right)_{ik} \frac{1}{2}(IP_k - EA_k)^{-\frac{1}{2}} \frac{1}{2}(IP_k + EA_k)$$




SCR is a Bayes regression with regularization parameters where coefficients could be found:

$$\boldsymbol{a} = \underset{a}{\mathrm{argmax}}\, p(\boldsymbol{a}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{V})$$

$$p(\boldsymbol{a}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{V}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{a})p(\boldsymbol{a}|\,\boldsymbol{V})}{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{V})}$$

$\boldsymbol{X}$ – descriptor matrix

$\boldsymbol{y}$ – activity values

$\boldsymbol{V}$ – regularization parameters


*Filimonov D.A. et al. SAR and QSAR Environ. Res., **2009**, 20: 679-709.*
*Rudik A.V. et al. J. Chem. Inform. Model., **2014**, 54: 498–507.*

# Self consistent regression

- The impact of each feature variable is restricted
- The *a priori* distribution of regression coefficients is assumed
- Parameters of this distribution are introduced to the model development process as regularization parameters

Distribution of regression coefficients $p(\boldsymbol{a}|\mathbf{V})$ with $\mathbf{V}$ parameters. Regression coefficients are found with maximization of *a posteriori* density:

$$\boldsymbol{a} = \underset{\boldsymbol{a}}{\text{ArgMax}}\, p(\boldsymbol{a}|\mathbf{X}, \boldsymbol{y}, \mathbf{V}),$$

Which is expression

$$p(\boldsymbol{a}|\mathbf{X}, \boldsymbol{y}, \mathbf{V}) = \frac{p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{a})p(\boldsymbol{a}|\mathbf{V})}{p(\boldsymbol{y}|\mathbf{X}, \mathbf{V})}$$

composed of likelihood functions

$$p(\boldsymbol{a}|\mathbf{V}) = Exp\left(\frac{1}{2} trLn\left(\frac{1}{2\pi\sigma^2}\mathbf{V}\right) - \frac{1}{2\sigma^2}\boldsymbol{a}'\mathbf{V}\boldsymbol{a}\right).$$

**SCR fits quantitative model development needs but could be applied for classification model development**

# Exponential and Logistic SCR

There are two features of ESCR and LSCR that distinguish them from SCR:

- Misclassification is strongly penalized
- Structures close to separating hyperplane are more important and have increased value of weight in the model

These features make ESCR/LSCR common to support vector machine as well as native SCR and allow to select the most important features in enhanced manner

*The LSCR and ESCR algorithms are implemented from scratch using the C++ programming language and integrated into the R environment with Rcpp mediation for results processing.*

# Probability estimation

Probabilities for positive and negative examples could be expressed with Bernoulli scheme:

$$p(\boldsymbol{y}|\mathbf{X},\boldsymbol{a}) = Exp\left(\sum_{k=1}^{n}\left(y_k LnP_k + (1-y_k)Ln(1-P_k)\right)\right)$$

$P_k = P(\boldsymbol{x}_k,\boldsymbol{a})$ – probability of positive case for $k$ structure with descriptors $\boldsymbol{x}_k$ and regression parameters $\boldsymbol{a}$.

LSCR is produced with introducing additional the logistic function:

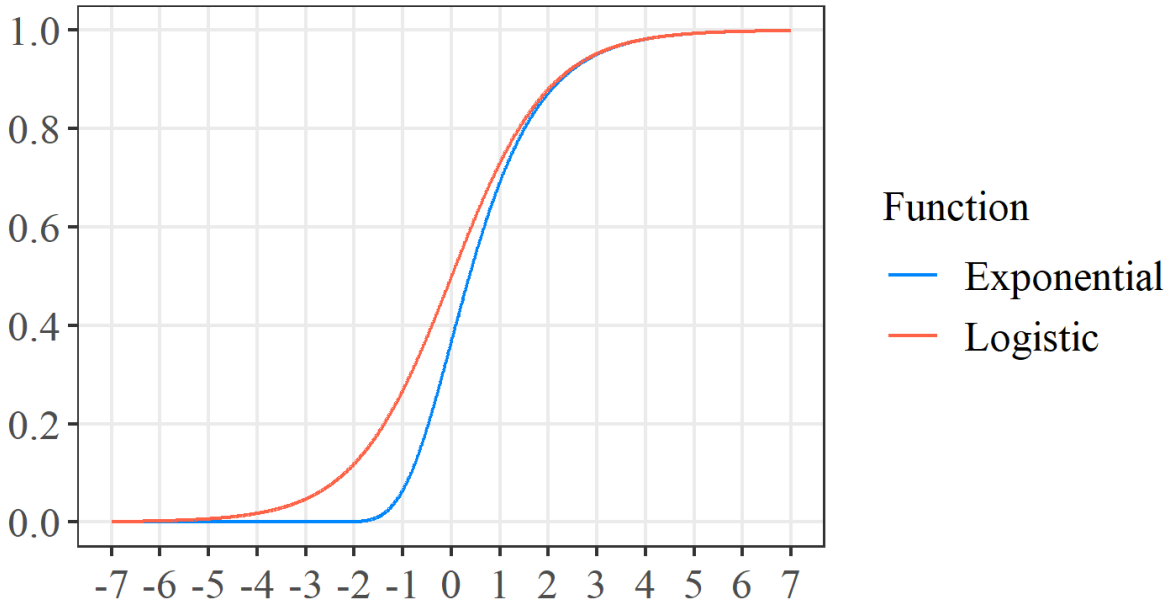
$$P(\boldsymbol{x},\boldsymbol{a}) = \left(1 + Exp(-\boldsymbol{x}'\boldsymbol{a})\right)^{-1} = \frac{1}{1 + Exp(-\boldsymbol{x}'\boldsymbol{a})}$$

Resulting in a likelihood function:

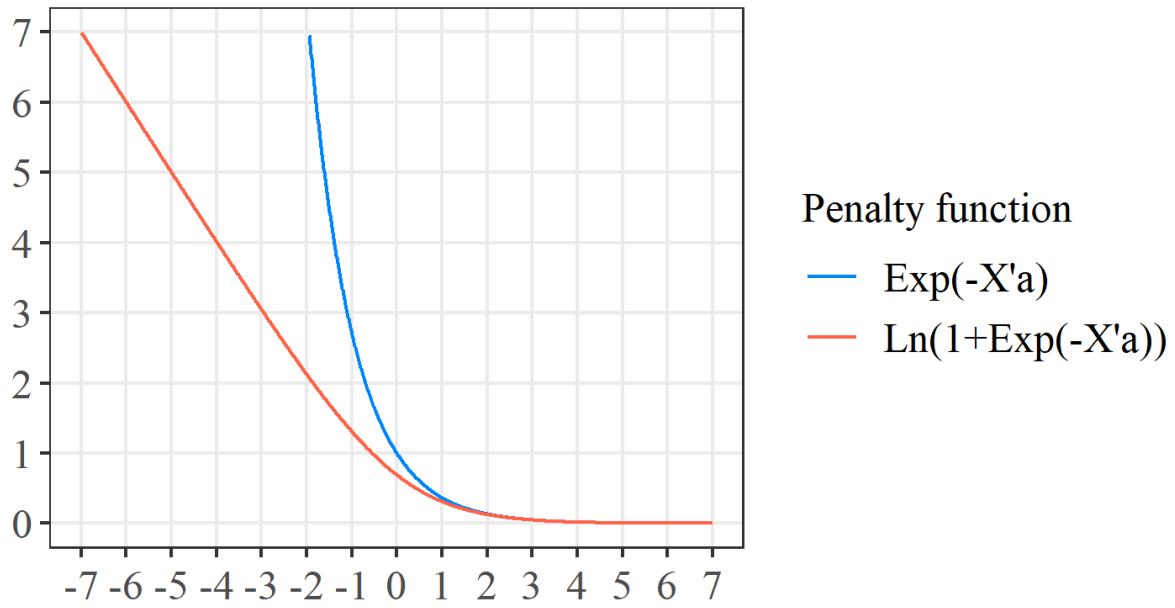
$$Ln\left(p(\boldsymbol{y}|\mathbf{X},\boldsymbol{a})\right) = -\sum_{k=1}^{n} Ln\left(1 + Exp\left(-u_k \boldsymbol{e}_k'\mathbf{X}\boldsymbol{a}\right)\right), \qquad u_k = 2y_k - 1, \qquad u_k = \pm 1, \qquad u_k^2 \equiv 1.$$

# Penalties

## Probability as a function




## Penalty functions



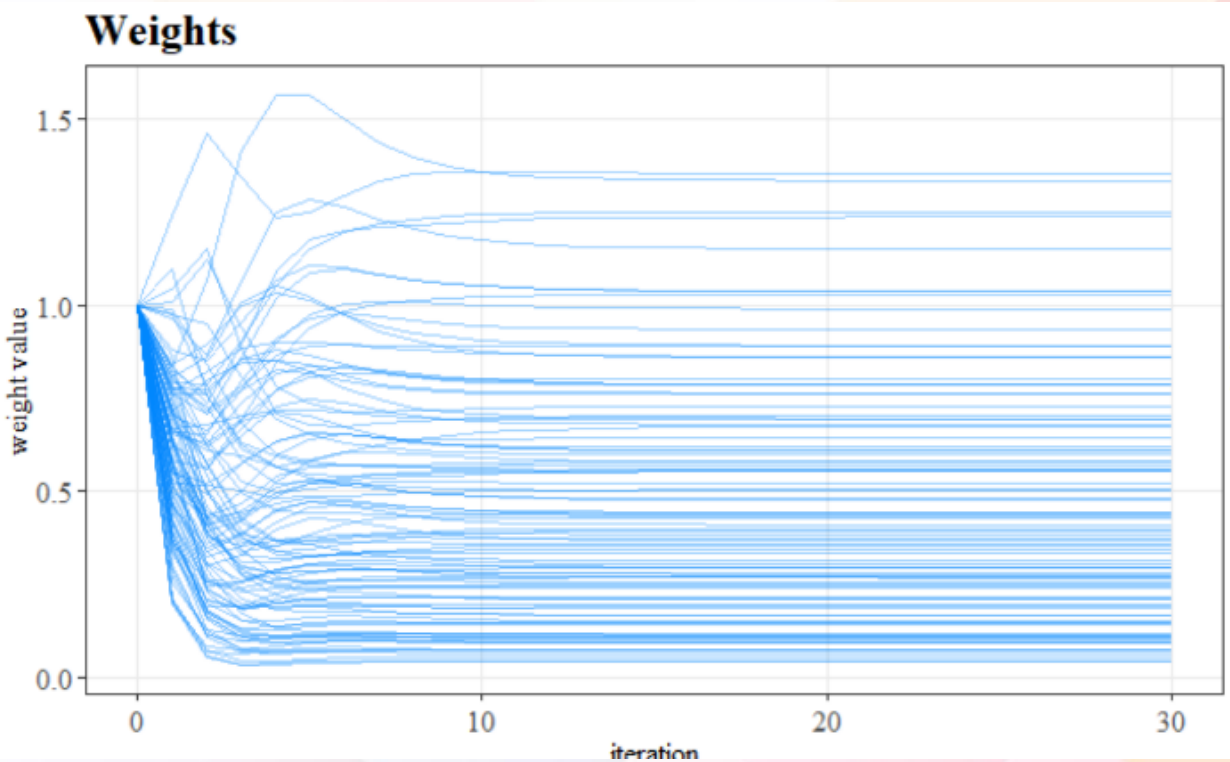

$$Ln\left(1 + Exp\left(-u_k \boldsymbol{e}_k' \mathbf{X}\boldsymbol{a}\right)\right)$$

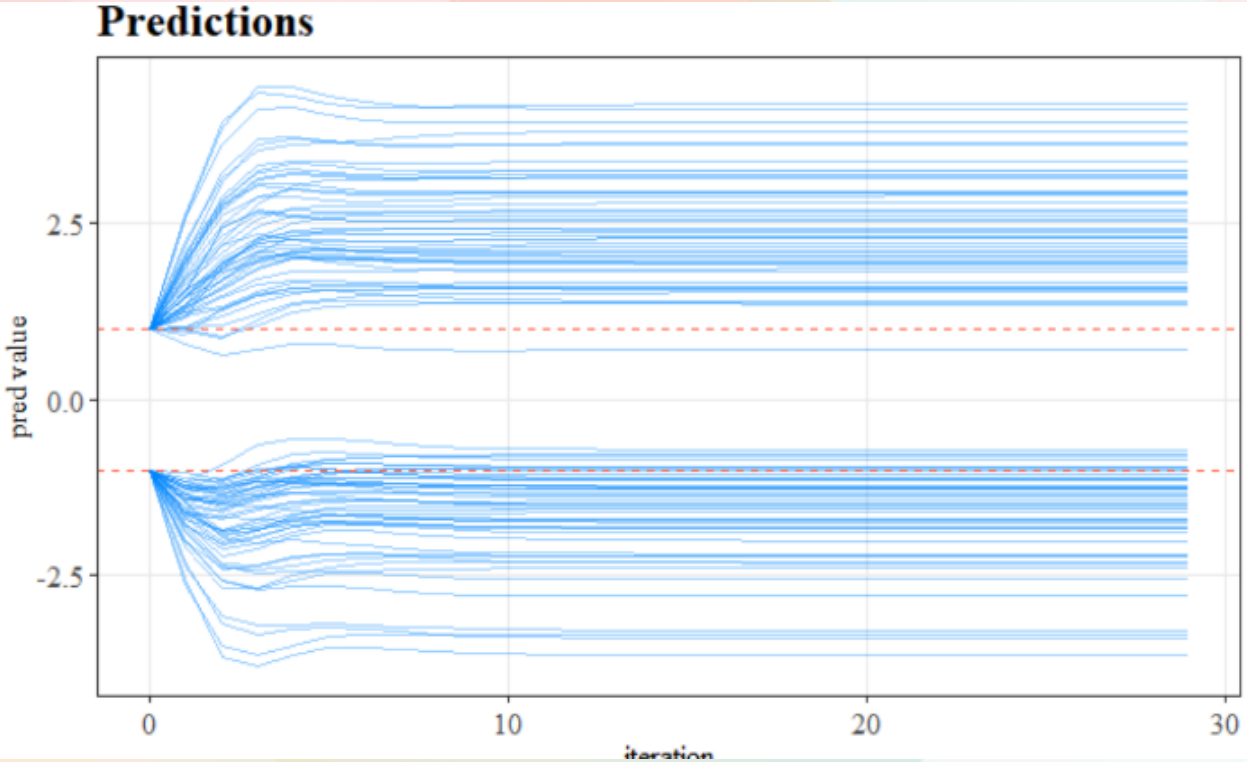
function penalizes in almost linear manner

The penalties of misclassification could be enhanced using

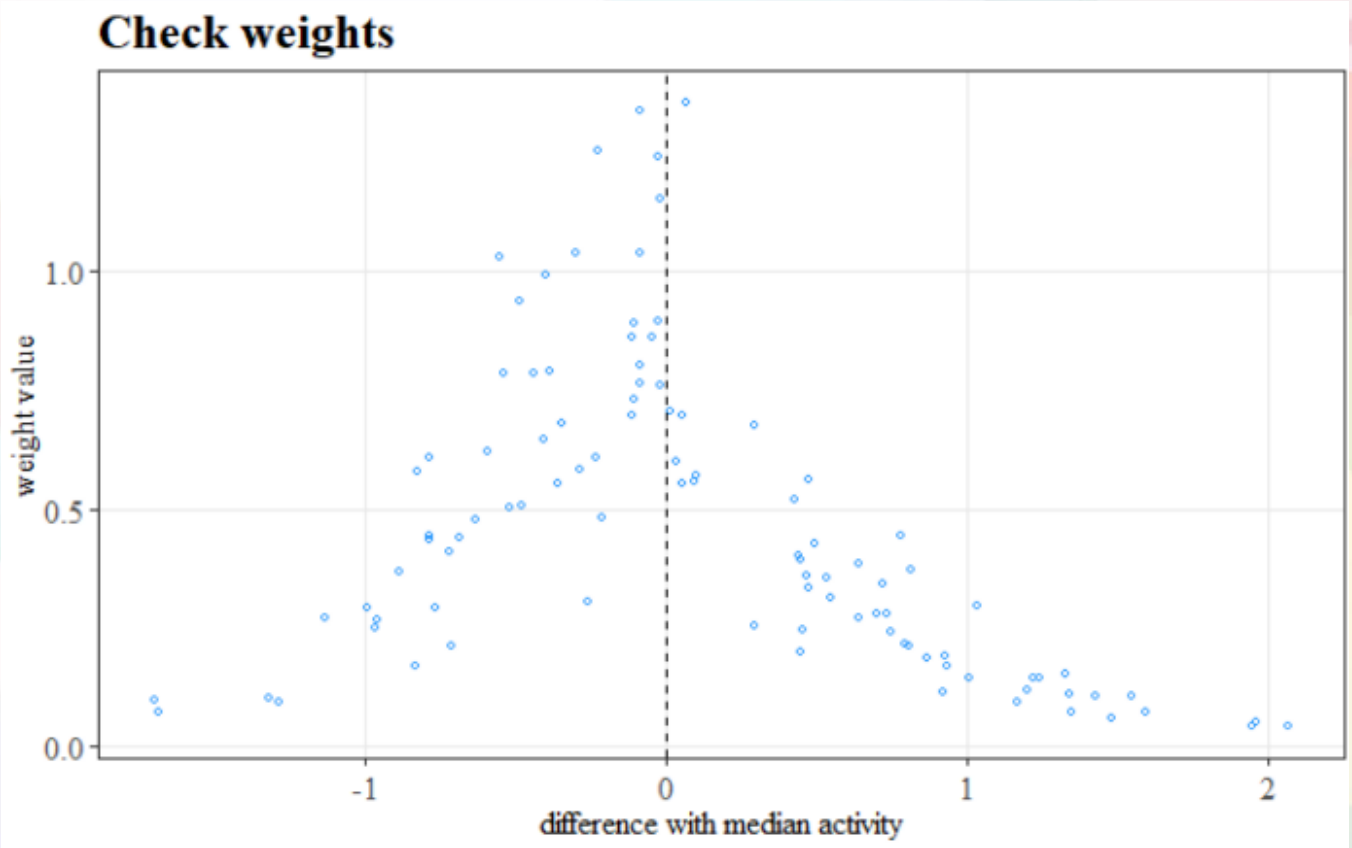
$Exp\left(-u_k \boldsymbol{e}_k' \mathbf{X}\boldsymbol{a}\right)$ instead

# Simulations with generated data



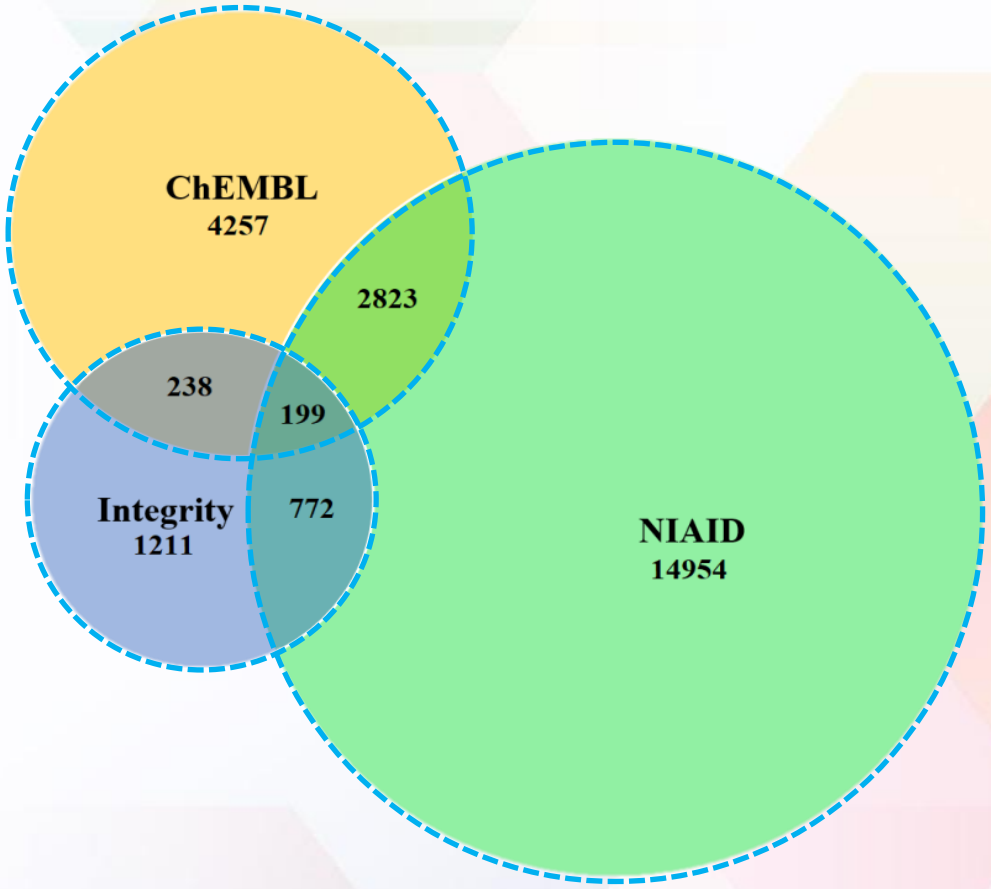



# Generated data weights




**The structures close to classes bounds have the highest weights**

# Data sources: HIV datasets




**Number of compounds extracted from each database**

HIV inhibitors structural and activity data was extracted from three sources:

- **NIAID ChemDB HIV** is a freely available database of HIV inhibitors;
- **ChEMBL** is a freely available database with the data on drug-like compounds;
- **Integrity** is a commercial database with pharmaceutical development data.

All extracted data was curated in accordance with modern requirements.

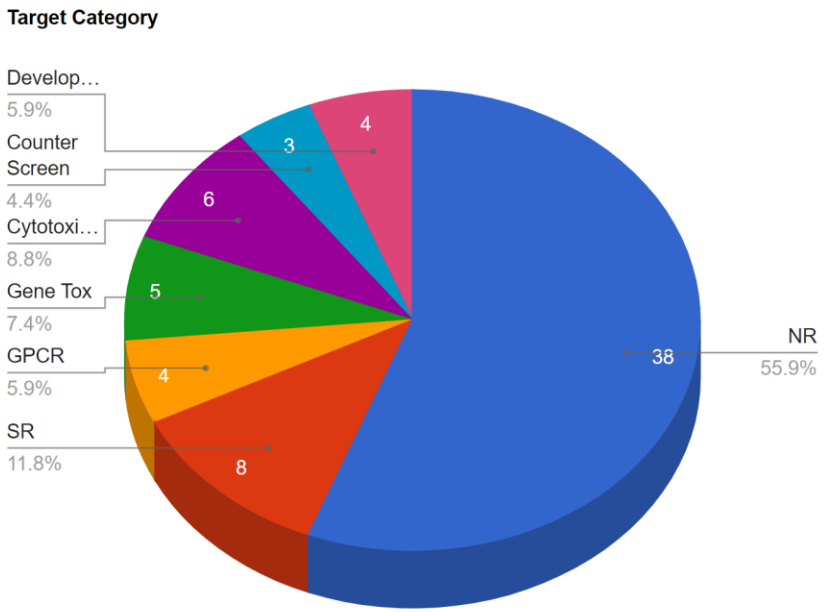
| | IN | PR | RT |
|---|---|---|---|
| **NIAID** | 10377/**3459** | 7604/**5972** | 8936/**5675** |
| **ChEMBL** | 2283/**1430** | 2387/**1542** | 2149/**1390** |
| **Integrity** | 563/**328** | 316/**268** | 731/**615** |

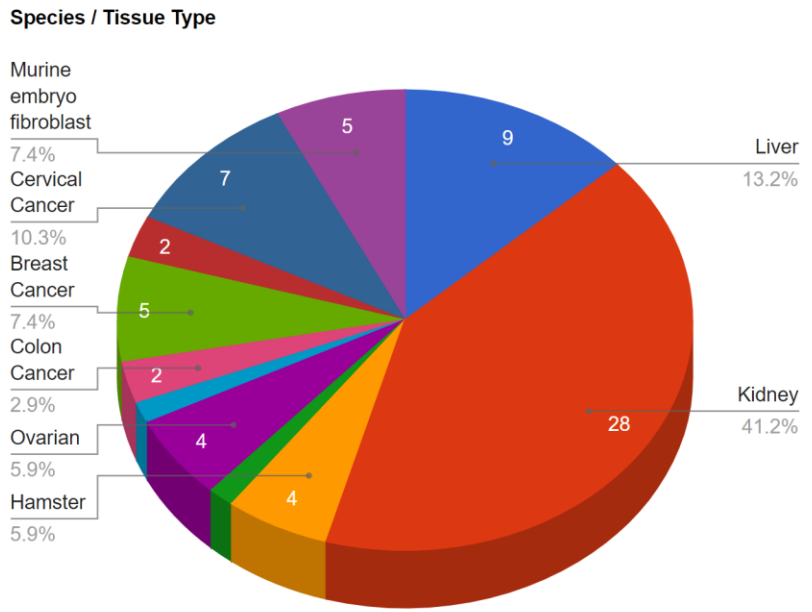
**Number of entries before / after curation**

# Data sources: Tox21 datasets

**Tox21 – 57 datasets**

data generated from nuclear receptor signaling and stress pathway assays




**HIV data was classified based on the cutoff equal to 1 μM**
**Tox21 data taken initially classified**

# HIV models

**Integrase**

| Модель | N | BA | V | CV |
|---|---|---|---|---|
| SVM | 4091 | 0.826 | 111* | 0.698 |
| ESCR | 4091 | 0.840 | 111 | 0.818 |
| LSCR | 4091 | 0.839 | 107 | 0.819 |

**Protease**

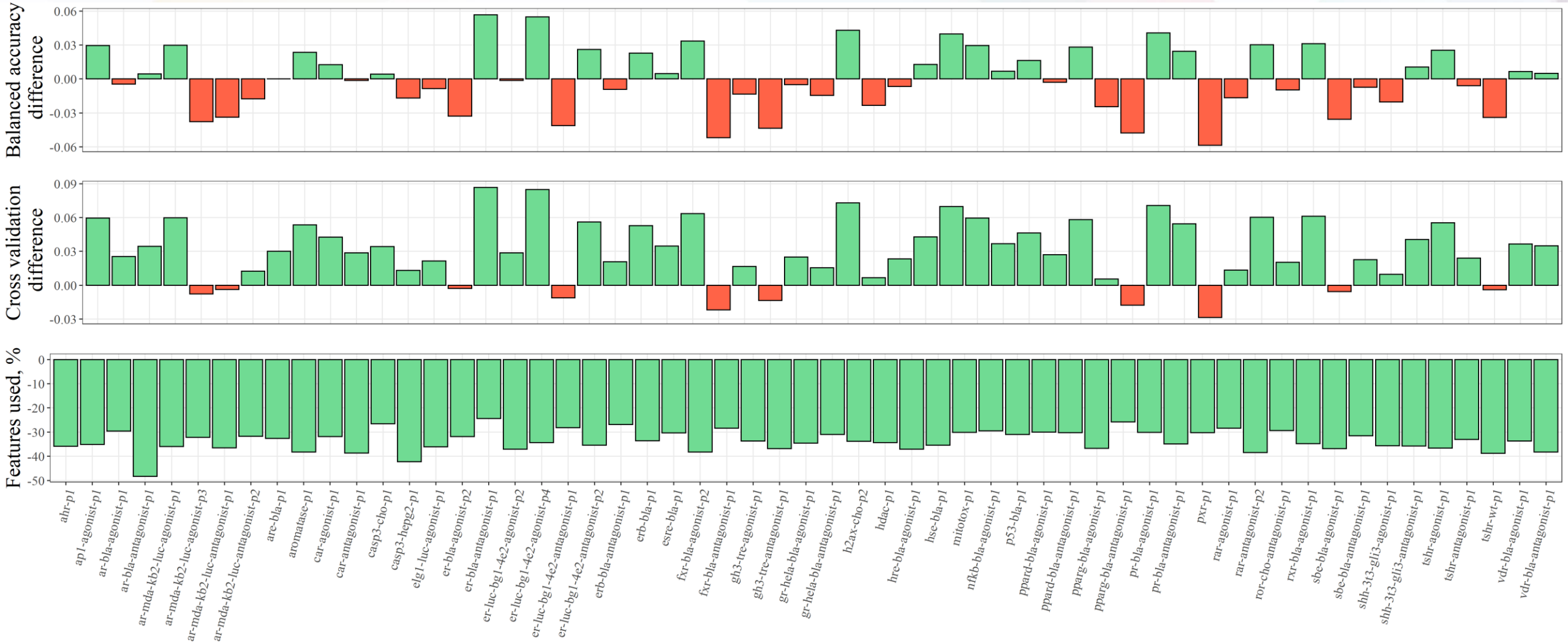
| Модель | N | BA | V | CV |
|---|---|---|---|---|
| SVM | 6552 | 0.785 | 140* | 0.715 |
| ESCR | 6552 | 0.843 | 140 | 0.816 |
| LSCR | 6552 | 0.828 | 134 | 0.808 |

**Reverse transcriptase**

| Модель | N | BA | V | CV |
|---|---|---|---|---|
| SVM | 6309 | 0.672 | 166* | 0.636 |
| ESCR | 6309 | 0.777 | 166 | 0.758 |
| LSCR | 6309 | 0.761 | 128 | 0.740 |

*N – number of compounds in the dataset*
*BA – balanced accuracy*
*V – number of selected features*
*CV – 5-fold cross validation mean balanced accuracy*
** - SVM was applied with descriptors taken from ESCR*

# Tox21 models



# Conclusions

- A new approach to the classification models development was implemented from scratch.
- HIV inhibiting data and Tox21 data were used to build classification models.
- SCR, SVM, LSCR, ESCR approaches were used.
- Validation and comparison of models was carried out.
- LSCR and ESCR models showed advances in balanced for both training and cross validation test sets in comparison with SVM.
- LSCR and ESCR showed advances in dimensionality reduction problem in comparison with native SCR.

# Acknowledgements

Author thanks the following contributors for the assistance:
Filimonov Dmitry
Vladimir Poroikov

This study was supported by the Russian Foundation for Basic Research grant No. 20-04-60285

# Thank you for your kind attention!

# Back up

$$Ln\left(p(\boldsymbol{y}|\mathbf{X},\boldsymbol{a})\right) = -\sum_{k=1}^{n} Exp\left(-u_k \boldsymbol{e}_k' \mathbf{X}\boldsymbol{a}\right),$$

$$u_k = 2y_k - 1, \qquad u_k = \pm 1, \qquad u_k^2 \equiv 1.$$