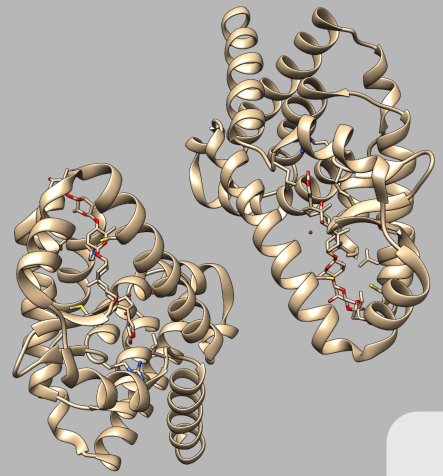
Biological activity of nuclear receptors with RDKit

Rafał A. Bachorz
Institute of Medical Biology of
Polish Academy of Sciences



INSTYTUT
BiolMed

Agenda

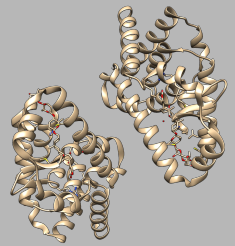
- Introduction
- Nuclear receptors
- How to measure the biological activity?
- The QSAR workflow
- ZINC20 screening
- Summary

Introduction

Nuclear Receptors

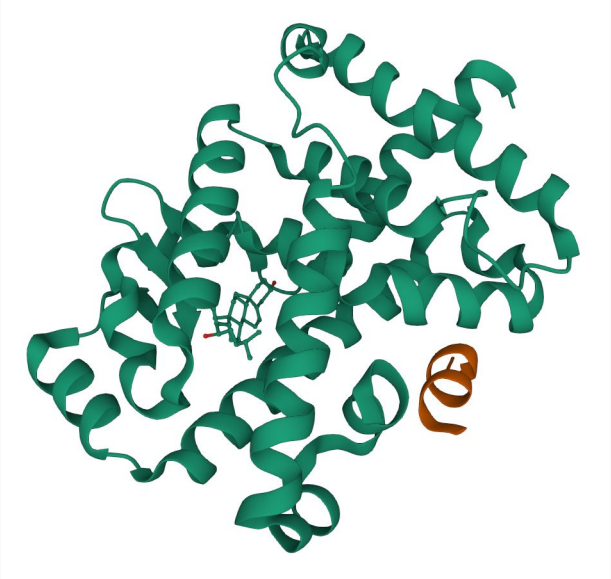
RDKit
Open-Source Cheminformatics and Machine Learning

INSTYTUT BiolMed



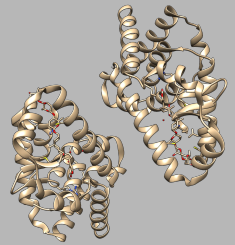
Nuclear Receptors (NR)

- Family of proteins
- RORs: the retinoic acid-related orphan receptors (α, β, γ)
- Many low-weight compounds exert biological activity by binding to NRs
- The activity of NRs depends on the conformation change
- The conformation change can be initiated by binding small molecule to the protein moiety
- Binding of a ligand functions as a switch that induces a conformational switch
- Linked to many human diseases like: atherosclerosis, osteoporosis, autoimmunological disorders, obesity, asthma, and cancer



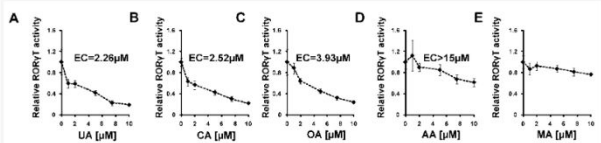
How to measure the biological activity?

Transfection and Luciferase Assay

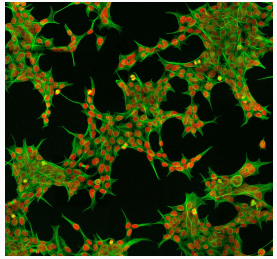


- The reporter vector: it turns into the luminescence protein after transfection and expression
- The GAL4-DBD RORγ fusion construct (containing human RORγ ligand binding domain)
- The HEK293 cell line
- Cotransfection with reporter vector and GAL4-DBD RORγ fusion construct
- Treatment of cotransfected cells with increasing concentrations of compounds
- Luminescence measurement reflects the inhibitory or activatory potential of considered compounds

Figure 5. Effect of ursolic acid analogs on RORγ-dependent transcription in the HEK293 cell line. HEK293 cells were cotransfected with the pGL4.35[luc2P/9XGAL4UAS/Hygro], GAL4-DBD RORγ, and pCMVSEAP vectors. Twenty-four hours later, the cells were treated with increasing concentrations of ursolic (**A**), corosolic (**B**), oleanolic (**C**), asiatic (**D**), and maslinic acids (**E**) for another 48 h. After that time, the cells were lysed, and luciferase activity was measured. Luciferase results are standardized for the transfection efficiency control, which was SEAP. Mean ± SD, *n* = 3. EC50 values were calculated using AAT Bioquest (Sunnyvale, CA, USA) EC50 calculator, (**https://www.aatbio.com/tools/ec50-calculator/**, accessed on 6 December 2021).

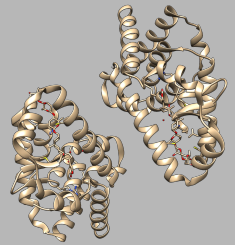


Pastwińska J, Karaś K, Sałkowska A, Karwaciak I, Chałaśkiewicz K, Wojtczak BA, Bachorz RA, Ratajewski M. Identification of Corosolic and Oleanolic Acids as Molecules Antagonizing the Human RORγT Nuclear Receptor Using the Calculated Fingerprints of the Molecular Similarity. Int J Mol Sci. 2022 Feb 8;23(3):1906. doi: 10.3390/ijms23031906. PMID: 35163824; PMCID: PMC8837092.

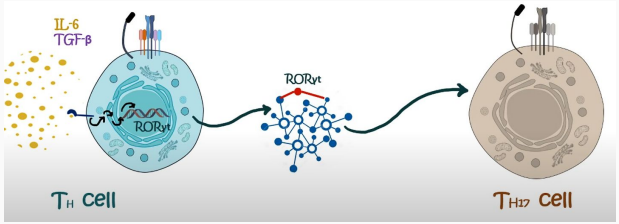


Immunofluorescent Human Embryonic Kidney 293 cells.
Phospho-Histone H3 (yellow), phalloidin (green), nuclei (red)
Author: Iznewton
Source: https://commons.wikimedia.org/wiki/File:HEK_293.jpg

Cytotoxicity towards Th17

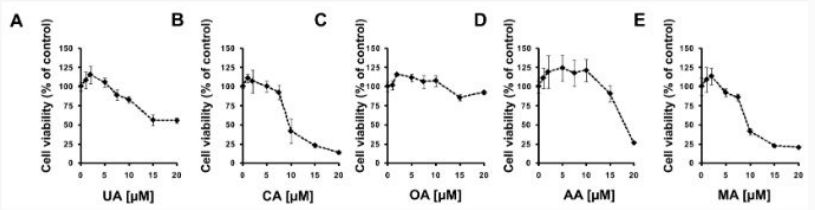


- RORγ are considered as master regulator of Th17 differentiation
- Cytotoxicity of considered compounds towards Th17 cells
- The CD+ cells are subject to Th17 differentiation in the presence of increasing concentrations of selected compounds
- CellTiter-Glo® Luminescent Cell Viability Assay
- No noticeable cytotoxicity up to 7.5 μM

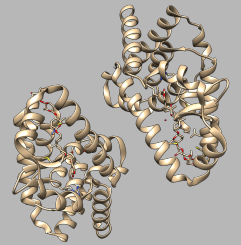


Pastwińska J, Karaś K, Sałkowska A, Karwaciak I, Chałaśkiewicz K, Wojtczak BA, Bachorz RA, Ratajewski M. Identification of Corosolic and Oleanolic Acids as Molecules Antagonizing the Human RORγT Nuclear Receptor Using the Calculated Fingerprints of the Molecular Similarity. Int J Mol Sci. 2022 Feb 8;23(3):1906. doi: 10.3390/ijms23031906. PMID: 35163824; PMCID: PMC8837092.

Figure 6. (**A**–**E**) Effect of ursolic acid analogs on CD4+ lymphocyte viability. CD4+ cells were isolated from buffy coats of healthy donors and subjected to Th17 polarization in the presence of increasing concentrations of ursolic, corosolic, oleanolic, asiatic, and maslinic acids for 5 days. Then, cell viability was determined using the CellTiter-Glo® Luminescent Cell Viability Assay. Mean ± SD, *n* = 3, compared with control cells.



CD4+ → Th17 differentiation



- The CD4+ cells are isolated from buffy coats
- They are differentiated into Th17 lymphocytes (protocol by Wilson *et al.*) in the presence of the compounds being verified
- Expression of RORγT, IL17A, IL17F, IL21, IL22 and APOD was measured using real time RT-PCR methodology

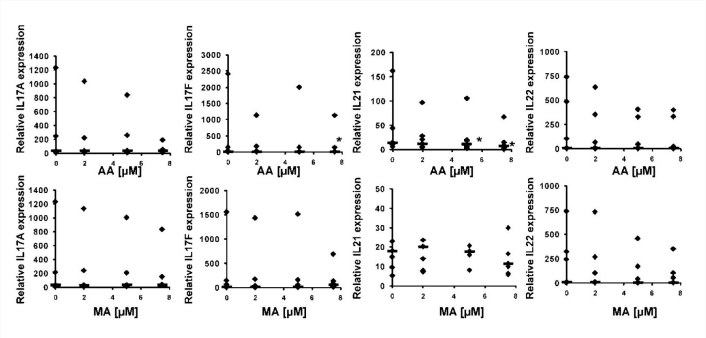
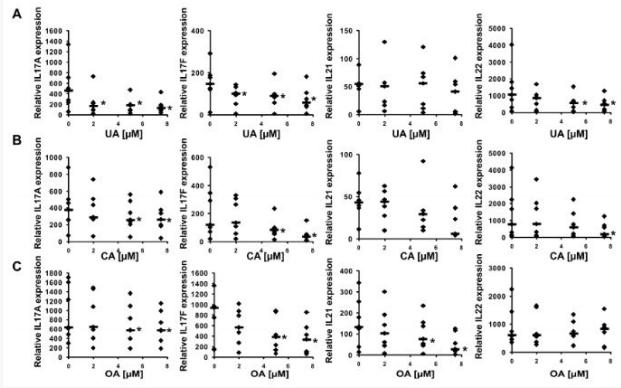


Figure 7. Effect of ursolic acid analogs on the expression of selected genes in human Th17 cells. Human naive CD4+ cells were treated with increasing concentrations of ursolic (**A**), corosolic (**B**), and oleanolic (**C**) acids and cultured under Th17-polarizing conditions for 5 days. Then, cells were collected for RNA extraction. The expression of the *IL17A*, *IL17F*, *IL21*, and *IL22* genes was determined by real-time RT–PCR. The results were normalized to the housekeeping genes *HPRT1*, *HMBS*, and *RPL13A*. An asterisk indicates a statistically significant difference at $p < 0.05$ compared with control cells. The data are presented as statistical dot plots with the median value (bars) from seven independent cultures ($n = 7$).

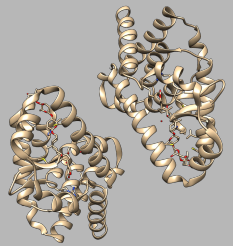


Pastwińska J, Karaś K, Sałkowska A, Karwaciak I, Chałaśkiewicz K, Wojtczak BA, Bachorz RA, Ratajewski M. Identification of Corosolic and Oleanolic Acids as Molecules Antagonizing the Human RORγT Nuclear Receptor Using the Calculated Fingerprints of the Molecular Similarity. Int J Mol Sci. 2022 Feb 8;23(3):1906. doi: 10.3390/ijms23031906. PMID: 35163824; PMCID: PMC8837092.

The QSAR workflow

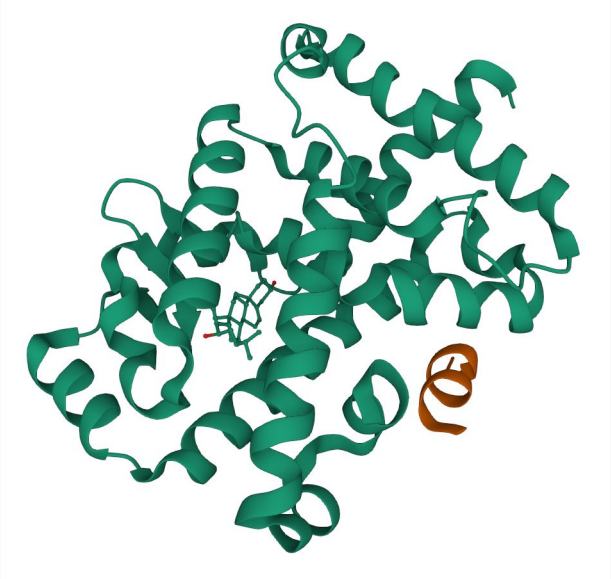
RDKit
Open-Source Cheminformatics and Machine Learning

INSTYTUT BiolMed



Modeling of biological activity

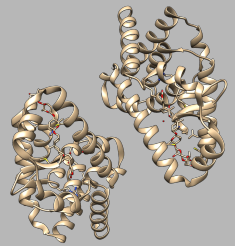
- Generic framework for QSAR studies
- QSAR approach
- The 2D (potentially also 3D) molecular descriptors and molecular fingerprints as features
- Fully configurable data processing workflow
- Implemented in the Python ecosystem
- Only open source libraries (RDKit and Mordred on the Cheminformatics site)
- Web application written with Jinja templates
- The source code will be published

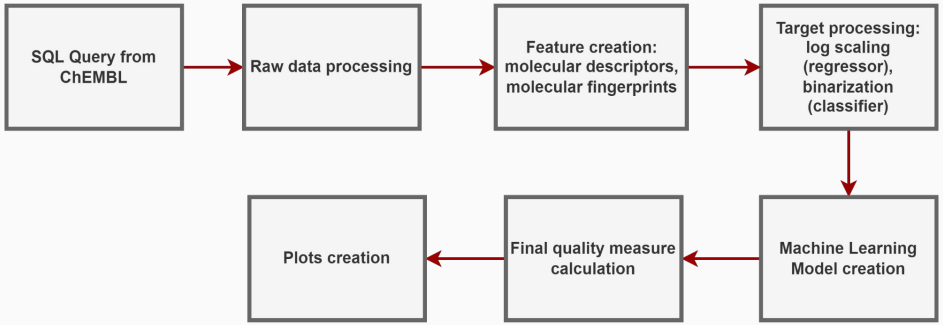


RDKit
Open-Source Cheminformatics and Machine Learning

INSTYTUT BioIMed

Modeling of biological activity: workflow

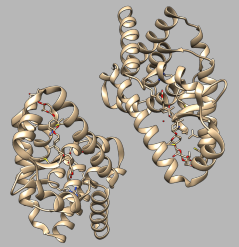




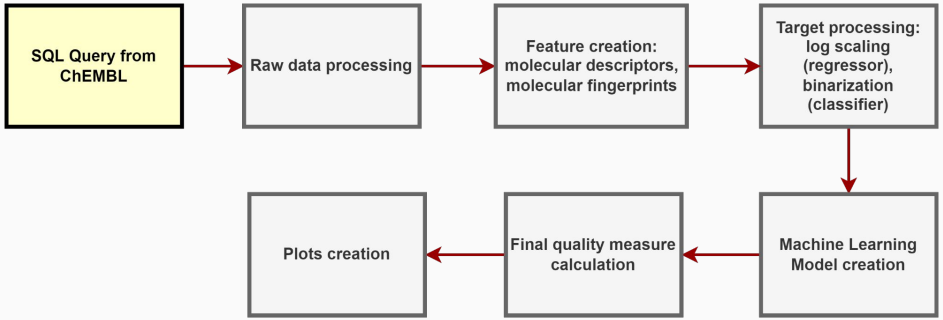
RDKit
Open-Source Cheminformatics and Machine Learning

INSTYTUT BioIMed

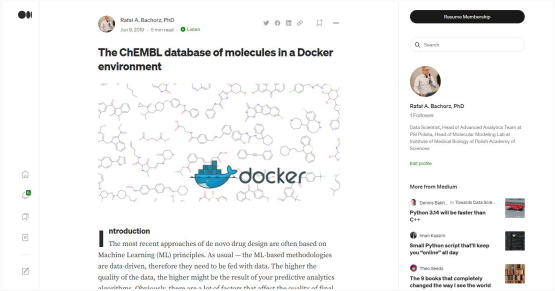
Modeling of biological activity: workflow



- Literally the SQL query
- Local instance of the ChEMBL database (dockerized setup)
- The query brings the raw data
- Can be carried out for any target (ChEMBL ID needed)



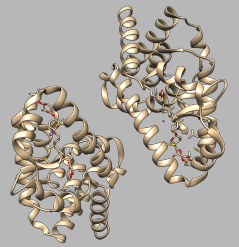

```python
def create_query(receptor):
    query = "select * from ( \
    select * from activities as act \
    inner join assays as assays on act.assay_id=assays.assay_id \
    left join docs as docs on act.doc_id=docs.doc_id \
    join target_dictionary as tardi on assays.tid=tardi.tid \
    inner join compound_structures as cs on act.molregno=cs.molregno \
    inner join compound_properties as cp on act.molregno=cp.molregno \
    left join molecule_dictionary as md on act.molregno=md.molregno \
    left join ligand_eff as le on act.activity_id=le.activity_id \
    where tardi.chembl_id in ('"+receptor+"')) as combined"
    return query
```



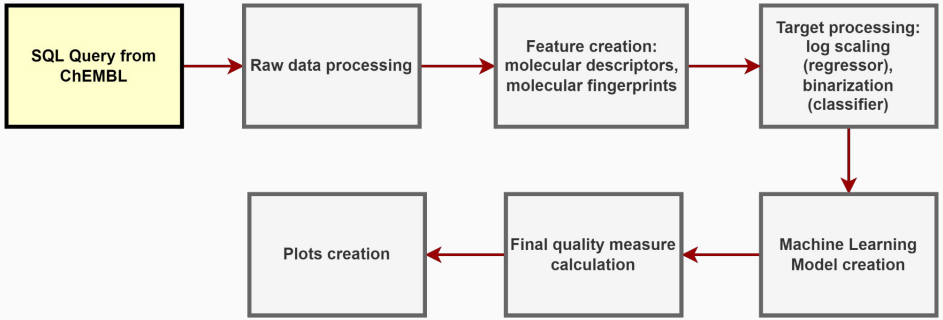
RDKit
Open-Source Cheminformatics and Machine Learning

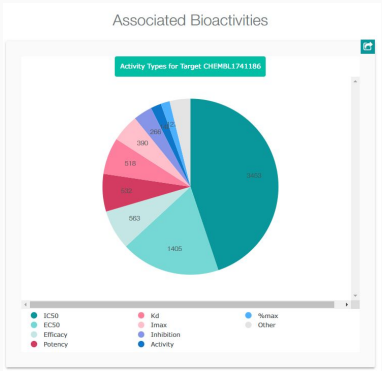
INSTYTUT BioIMed

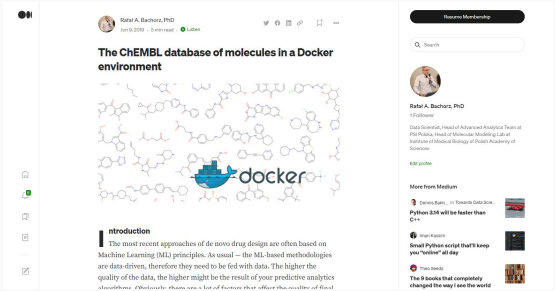
Modeling of biological activity: workflow



- Literally the SQL query
- Local instance of the ChEMBL database (dockerized setup)
- The query brings the raw data
- Can be carried out for any target (ChEMBL ID needed)

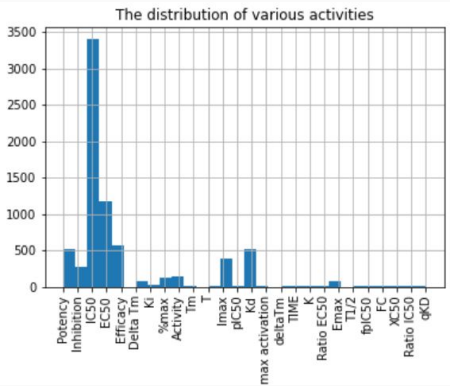


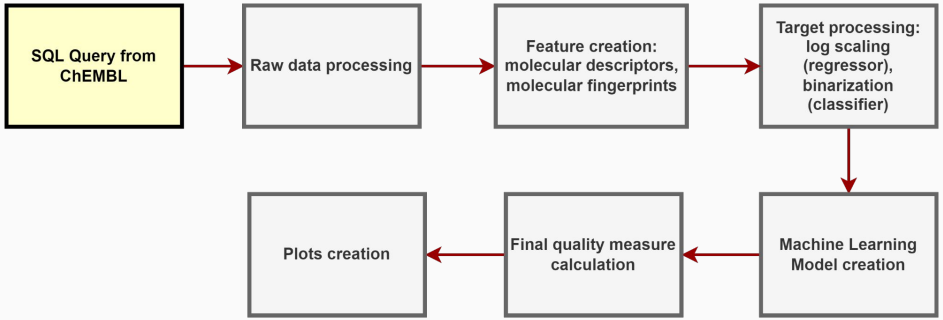


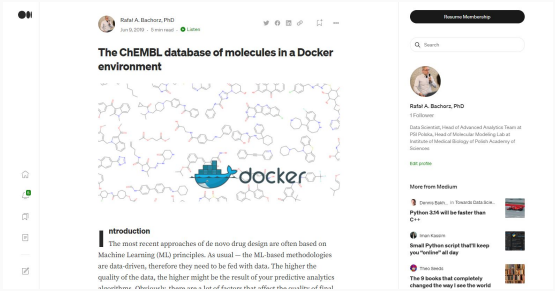


Modeling of biological activity: workflow

- Literally the SQL query
- Local instance of the ChEMBL database (dockerized setup)
- The query brings the raw data
- Can be carried out for any target (ChEMBL ID needed)



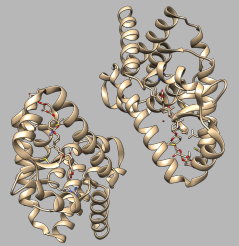




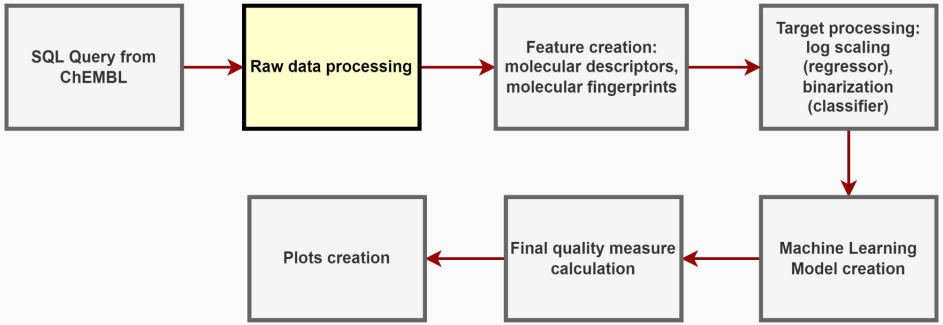
RDKit
Open-Source Cheminformatics and Machine Learning

INSTYTUT BioIMed

Modeling of biological activity: workflow

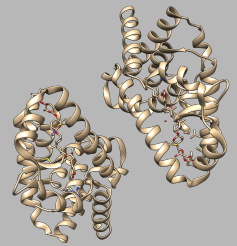


- Raw data sometimes confusing
- Multiple entries for the same molecule
- Sometimes contradictory
- The processing strategy
  - depending on std value the observation can be removed
  - for the remaining the aggregation strategy (mean, median)

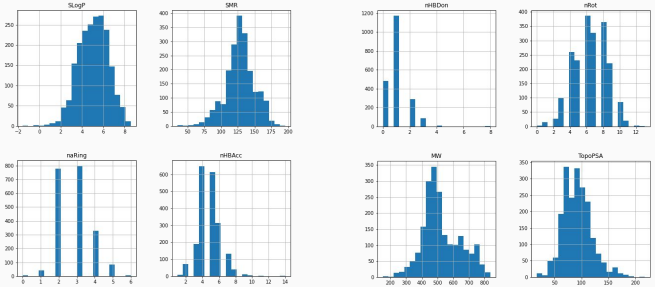


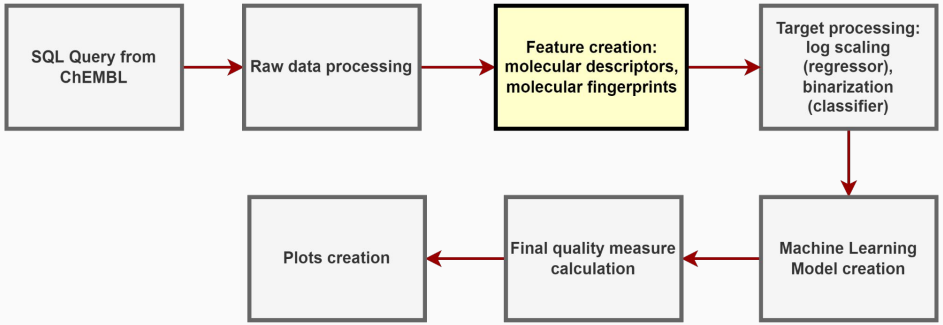
	molregno	standard_value_get_mean	standard_units_get_levels	standard_type_get_levels	chembl_id_compound_get_levels	canonical_smiles_get_levels	act_0	act_1	act_2	act_3	act_4	act_5	act_6	act_7	act_8	act_9	act_10	act_11	act_12	act_13	act_14
0	2208	199.530000	nM	IC50	CHEMBL38	CC1=C(/C=C/C(C)=C/C=C/C(C)=C/C(=O)O)C(C)(C)CCC1	199.53	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	40459	31900.000000	nM	IC50	CHEMBL28721	O=C(O)C(S)C(S)C(=O)O	31900.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	92776	1770.094667	nM	IC50	CHEMBL62136	O=S(=O)(c1ccccc1)N(CC(F)(F)F)c1ccc(C(O)(C(F)(F...	60.00	60.0	204.0	316.23	330.0	501.19	540.0	1700.0	1700.0	1730.0	2010.0	3000.0	3500.0	4400.0	6500.0
3	155592	10000.000000	nM	IC50	CHEMBL418971	CC(C)(c1ccc(O)cc1)c1ccc(O)cc1	10000.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	201200	2820.000000	nM	IC50	CHEMBL169	C[C@@H]1[C@H]2C3=CC[C@@H]4[C@@]5(C)CC[C@H](O)C...	130.00	680.0	1060.0	9410.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2467	2536694	239.000000	nM	IC50	CHEMBL4800103	O=C(O)Cc1ccc(CNC(=O)Nc2cc(Cl)c(-c3ccccc3OC(F)(...	239.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2468	2536937	419.000000	nM	IC50	CHEMBL4800346	CCS(=O)(=O)c1ccc(CNc2nc3cnc(-c4c(OC)ncnc4C4CC4...	98.00	740.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2469	2536979	44.500000	nM	IC50	CHEMBL4800388	CCCS(=O)(=O)c1ccc(CNc2nc3cnc(-c4c(C)ncnc4C4CC4...	15.00	74.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2470	2537316	340.000000	mix	IC50	CHEMBL4802041	COc1nc2ccc([C@@](O)(c3ccc(Cl)cc3)c3cncn3C)cc2c...	10.00	670.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2471	2537317	181.000000	nM	IC50	CHEMBL4802042	COc1nc2ccc([C@@](O)(c3ccccc3)C3CCN(C(C)=O)CC3)...	30.00	332.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Modeling of biological activity: workflow

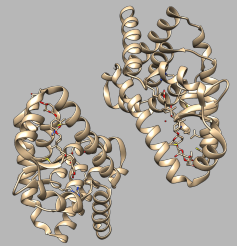


- The molecular descriptors with Mordred/RDKit
- The molecular fingerprints with RDKit
- The features can be transformed with PCA
    - Important for fingerprints taken by neural networks (sparsity)

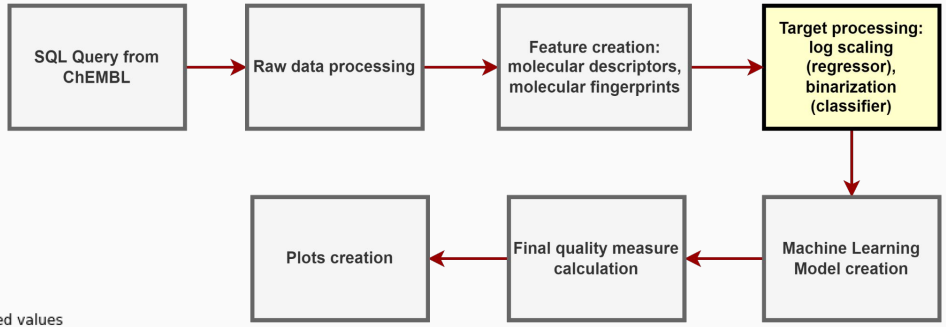


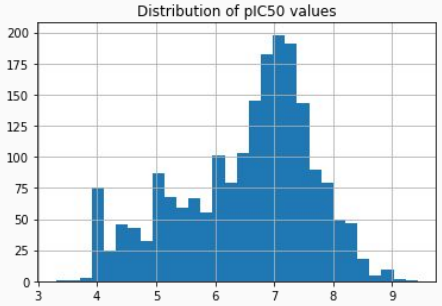


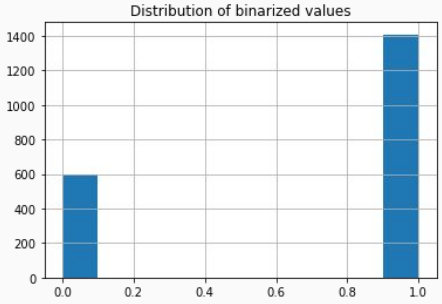
Modeling of biological activity: workflow



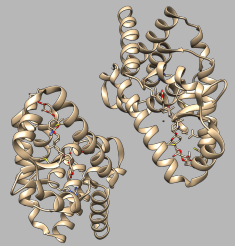
- The logarithmic transformation (IC50 -> pIC50), usually done for regressors
- The binarization with certain threshold (here 1000 nM), classifiers



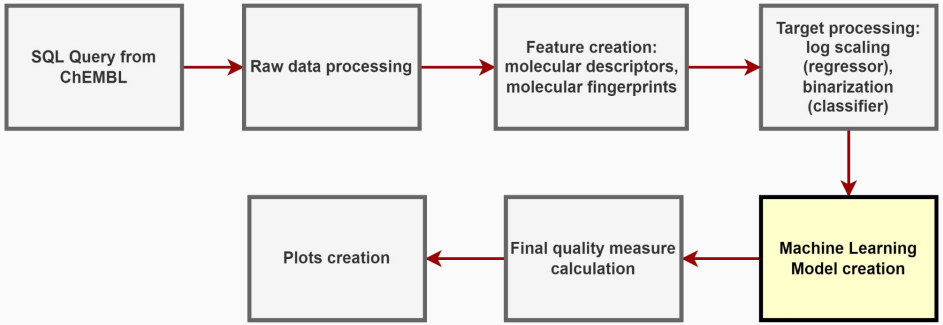




Modeling of biological activity: workflow



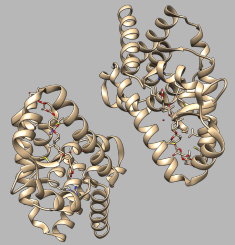
- The methodologies:
  - XGBoost
  - Multilayer perceptron
  - Random forest
  - Bagging
  - Support vector machines
  - Ridge
- Hyperparameters optimization with hyperopt
- Resulting models stored in a serialized form
- The serialized model is ready-to-deployment



Open-Source Cheminformatics and Machine Learning

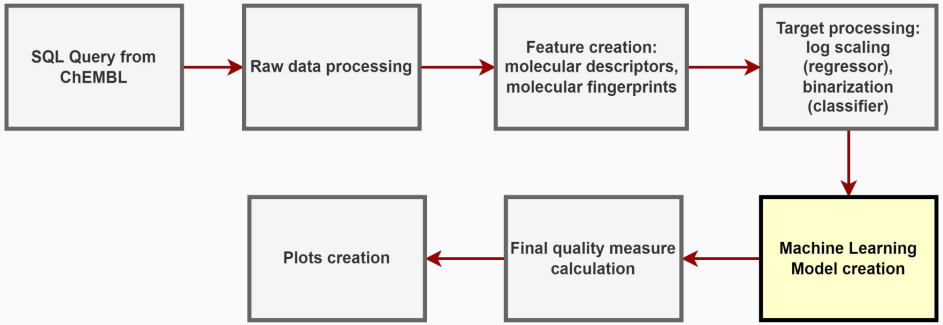
INSTYTUT BioIMed

Modeling of biological activity: workflow



pipelines > xgboost > {} pipeline_configuration_xgboost_classification_Morgan_1024_pca_256_QED.json > ...

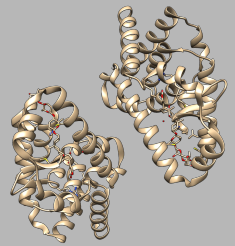

```
{
    "feature_transform":
        {
        "Calculate_molecular_features":
        {
            "proceed": "yes",
            "calculate_fps": "yes",
            "calculate_descriptors": "yes",
            "fingerprint_size": 1024,
            "fp_type": "morgan",
            "descriptors_types": ["QED", "Lipinski"],
            "meta_descriptors_types": ["QED"],
            "label": "Molecular features"
        },
        "Scaling":
        {
            "proceed": "no",
            "scaling_features": ["SLogP", "SMR", "naRing", "nHBAcc", "nHBDon", "nRot", "MW", "TopoPSA", "QED"],
            "label": "Scaling"
        },
        "PCA":
        {
            "proceed": "yes",
            "apply_to_md": "no",
            "apply_to_fp": "yes",
            "n_components": 256,
            "label": "PCA"
        }
    },
    "target_transform":
    {
        "Target_transform":
        {
            "proceed": "no"
        },
        "Target_binarization":
        {
            "proceed": "yes",
            "threshold": 1000,
            "label": "Target Binarization"
        },
        "O_removal":
        {
            "proceed": "no",
            "factor": 1.5,
            "label": "Outlier removal"
        }
```

RDKit
Open-Source Cheminformatics and Machine Learning

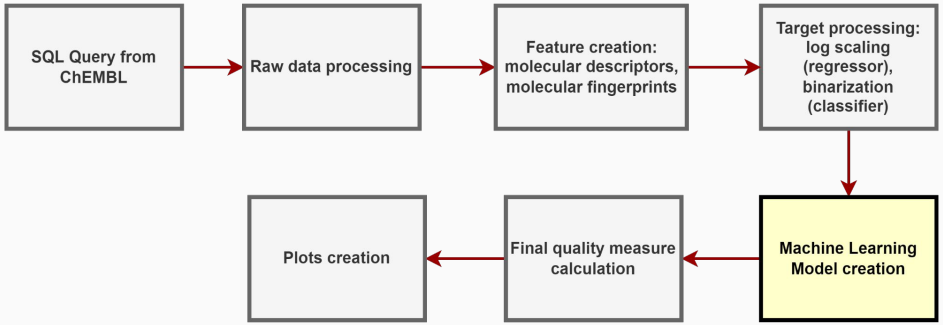
INSTYTUT BioIMed

Modeling of biological activity: workflow

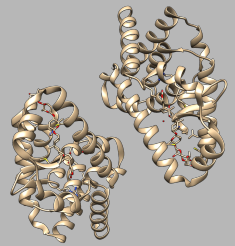


ng_configurations > mean_100 > {} bagging_classification_Morgan_1024_pca_512_QED.json > ...

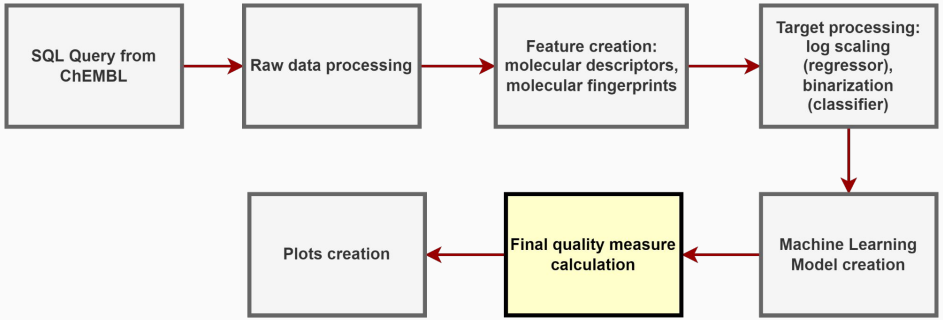

```
{
    "prediction_methodology": "M_Bagging",
    "prediction_type": "classification",
    "GPU_support": true,
    "training_aux_data":
    {
        "run_name": "test",
        "n_outer": 1,
        "n_cv": 5,
        "development": false,
        "goal_function": "precision",
        "goal_function_multiplier": -1.0,
        "threshold": 0.5,
        "track_model": false,
        "experiment": null,
        "comment": null,
        "max_evals": 200,
        "pipeline_file": "bagging/pipeline_configuration_bagging_classification_Morgan_1024_pca_512_QED.json"
    },
    "data_preparation": {
        "data_file": "ror_gamma_homo_sapiens_chembl_3.0_IC50.csv",
        "max_level_activity": 15,
        "std_threshold": 100,
        "strategy": "mean"
    },
    "model_storage": {
        "resulting_model": "model_bagging_classifier_Morgan_1024_pca_512_QED_mean_100.model"
    }
}
```

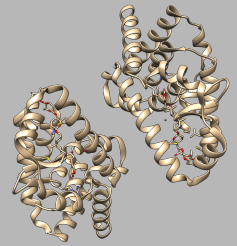
Modeling of biological activity: workflow

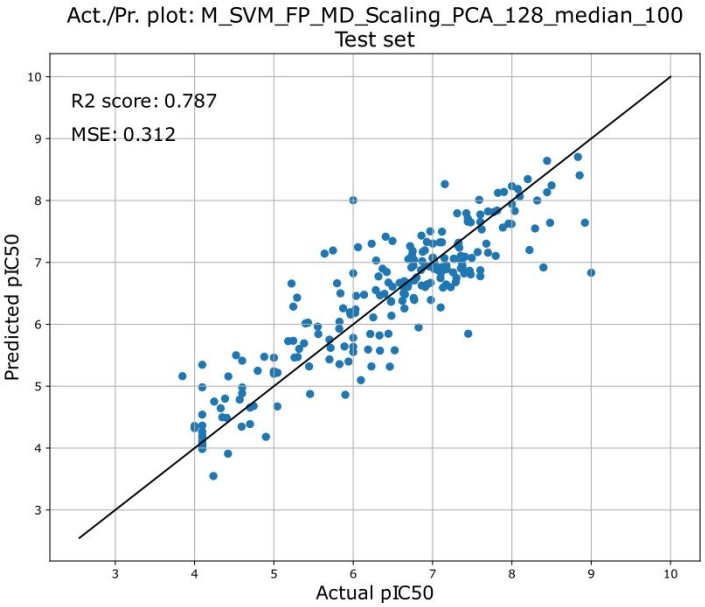


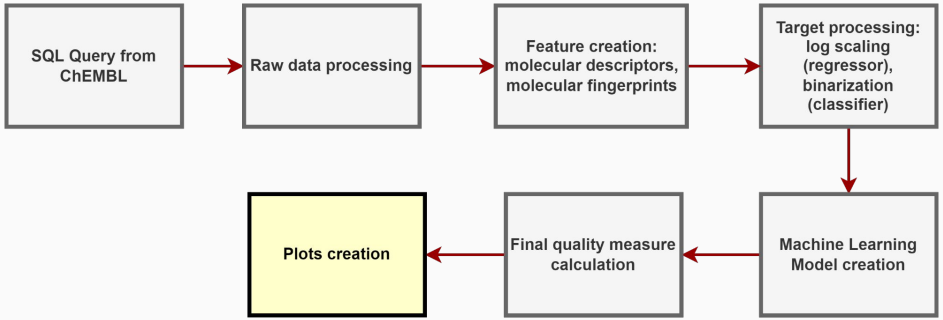
- The regressors
  - R2 score
  - Mean squared (log) error
  - Mean absolute error
  - Mean absolute percentage error
- The classifiers
  - Precision
  - Recall
  - Accuracy
  - F1-score
  - ROC-AUC
  - Average precision
  - Matthews coefficient



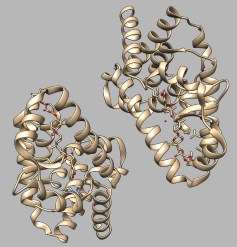
Modeling of biological activity: workflow

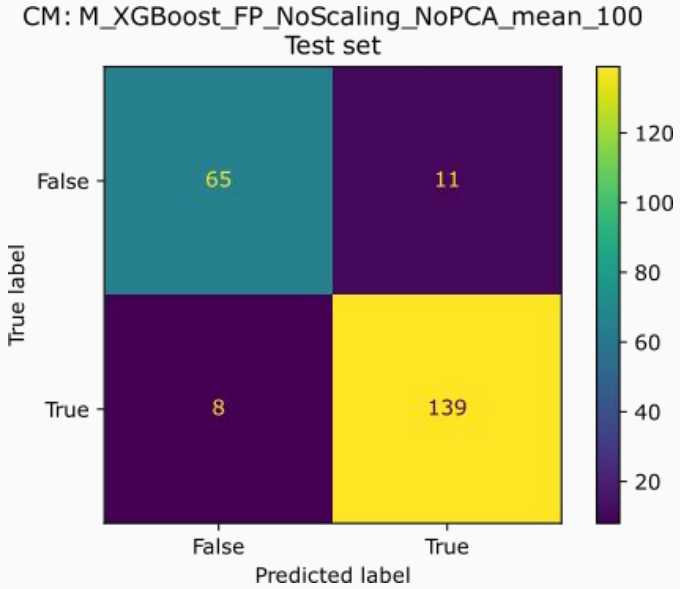


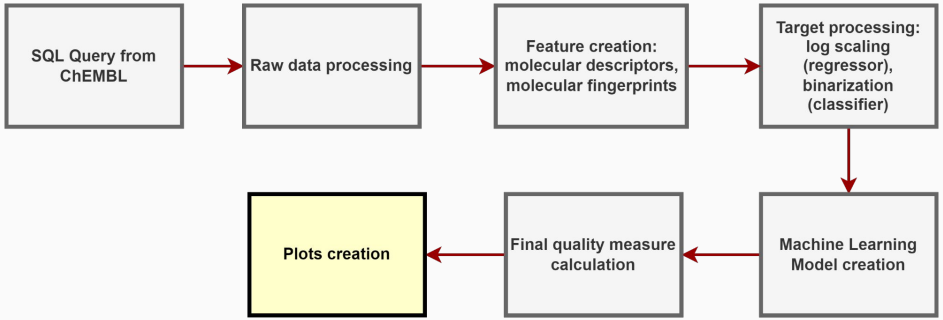




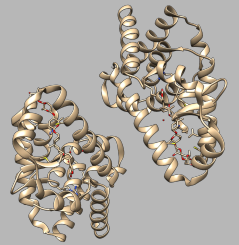
Modeling of biological activity: workflow



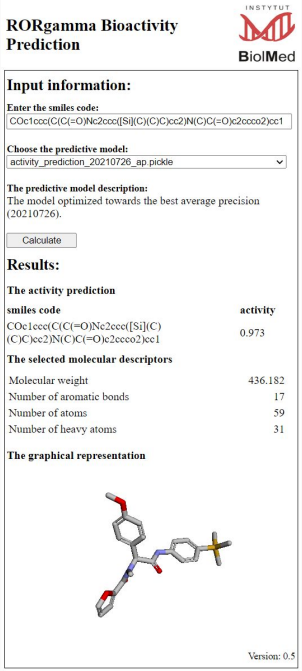




Modeling of biological activity: webapp



- Web application
- Jinja templates
- Dockerized setup
- With load balancing (NGINX)
- Used by experimental scientists



RORgamma Bioactivity Prediction

INSTYTUT BioIMed

Input information:

Enter the smiles code:

COc1ccc(C(C(=O)Nc2ccc([Si](C)(C)C)cc2)N(C)C(=O)c2ccco2)cc1

Choose the predictive model:

activity_prediction_20210726_ap.pickle

The predictive model description:

The model optimized towards the best average precision (20210726).

Calculate

Results:

The activity prediction

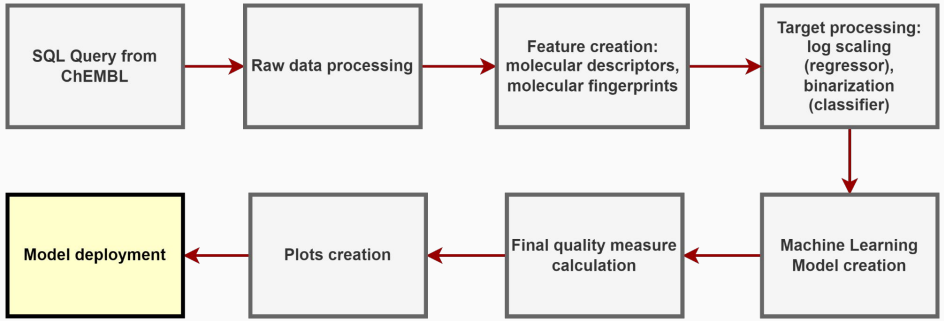
smiles code	activity
COc1ccc(C(C(=O)Nc2ccc([Si](C)(C)C)cc2)N(C)C(=O)c2ccco2)cc1	0.973

The selected molecular descriptors

Molecular weight	436.182
Number of aromatic bonds	17
Number of atoms	59
Number of heavy atoms	31

The graphical representation

Version: 0.5



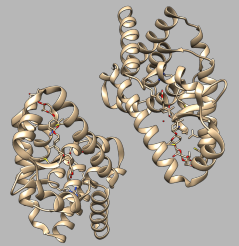
Manuscript in preparation...

The ZINC20 screening

Open-Source Cheminformatics and Machine Learning

INSTYTUT BiolMed

The ZINC20 search

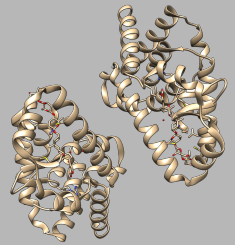


- The “active” RORγ ligands were taken as the reference species
- “active”: IC50 < 1000 nM
- 1704 molecules
- The ZINC20 library was considered as a target
- ca. 884 Ms of species
- ca. 750 Ms of them are purchasable
- The **chemfp** library for similarity search

Molecular Weight (up to, Daltons)

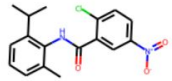
LogP (up to)	200	250	300	325	350	375	400	425	450	500	>500	Totals, by LogP
-1	27,791	172,563	710,795	1,072,978	2,241,498	786,738	276,834	116,066	92,417	77,790	7,310	5,582,780
0	139,434	934,776	3,655,384	5,126,157	10,608,025	3,498,214	1,663,579	708,919	570,546	507,344	4,734	27,417,112
1	362,437	2,884,636	12,030,074	16,154,544	33,650,249	11,885,957	6,807,876	3,178,487	2,648,581	2,412,998	9,940	92,025,779
2	467,220	4,584,223	22,941,208	30,908,513	65,047,385	26,752,849	17,839,254	9,349,272	8,099,970	7,686,687	24,554	193,701,135
2.5	167,513	2,136,113	12,849,121	17,977,157	38,682,058	18,584,223	13,812,274	8,111,104	7,197,414	6,979,014	24,126	126,520,117
3	90,548	1,570,772	11,037,383	16,282,627	34,831,558	19,940,391	16,037,132	10,339,743	9,362,233	9,118,717	37,422	128,648,526
3.5	36,748	929,872	7,920,574	12,490,662	27,380,104	18,703,024	16,485,194	11,784,160	10,774,472	10,693,411	58,791	117,257,012
4	9,017	369,565	4,332,131	6,472,808	10,487,856	13,034,155	14,329,253	11,683,208	10,891,465	11,003,975	86,262	82,699,695
4.5	993	86,613	1,814,492	3,457,942	6,367,225	8,853,064	10,320,054	9,945,353	9,486,869	9,825,079	117,980	60,275,664
5	150	13,393	536,018	1,405,708	3,168,584	4,995,850	6,471,525	7,025,034	6,976,742	7,325,833	144,297	38,063,134
>5	39	1,097	22,854	103,521	376,905	927,395	1,670,856	2,195,160	2,588,702	3,052,048	767,762	11,706,339
, by Weight	1,301,890	13,683,623	77,850,034	111,452,617	232,841,447	127,961,860	105,713,831	74,436,506	68,689,411	68,682,896	1,283,178	**883,897,293** Substances **1.9K** Tranches

The ZINC20 search

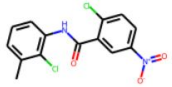


- The **chemfp** library was used to calculate the Morgan fingerprints
- The potential of entire content of the ZINC20 library was used
- The Tanimoto similarity coefficients were calculated between ZINC20 and “active” RORγ antagonists
- The similarity threshold: 0.7
- After removal of training set duplicates: 1673 potential RORγ antagonists
- By considering the commercial issues: three of them were initially selected to the experimental verification

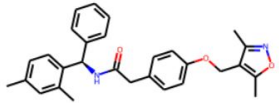
Found species:



Classifier: 0.95
Regressor: 922 nM

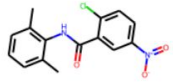


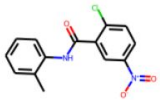
Classifier: 0.93
Regressor: 889 nM

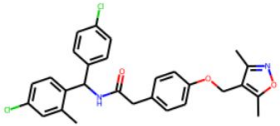


Classifier: 0.91
Regressor: 37 nM

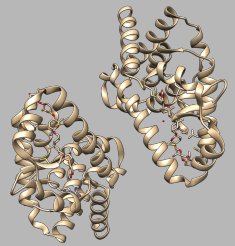
The most similar analogues (from ChEMBL)





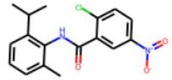


The ZINC20 search

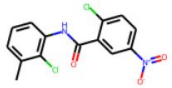


- The **chemfp** library was used to calculate the Morgan fingerprints
- The potential of entire content of the ZINC20 library was used
- The Tanimoto similarity coefficients were calculated between ZINC20 and “active” RORγ antagonists
- The similarity threshold: 0.7
- After removal of training set duplicates: 1673 potential RORγ antagonists
- By considering the commercial issues: three of them were initially selected to the experimental verification

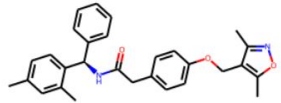
Found species:



Classifier: 0.95
Regressor: 922 nM

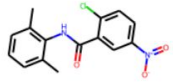


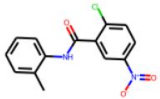
Classifier: 0.93
Regressor: 889 nM

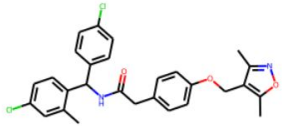


Classifier: 0.91
Regressor: 37 nM

The most similar analogues (from ChEMBL)

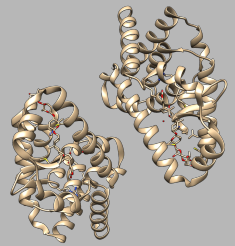






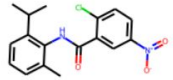
Already known potent RORγ antagonist not included in the ChEMBL database
Measured IC50: 30 nM

The ZINC20 search

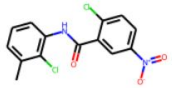


- The **chemfp** library was used to calculate the Morgan fingerprints
- The potential of entire content of the ZINC20 library was used
- The Tanimoto similarity coefficients were calculated between ZINC20 and “active” RORγ antagonists
- The similarity threshold: 0.7
- After removal of training set duplicates: 1673 potential RORγ antagonists
- By considering the commercial issues: three of them were initially selected to the experimental verification

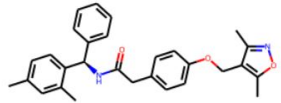
Found species:



Classifier: 0.95
Regressor: 922 nM

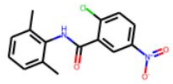


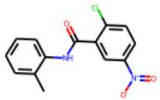
Classifier: 0.93
Regressor: 889 nM

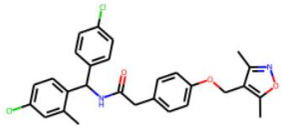


Classifier: 0.91
Regressor: 37 nM

The most similar analogues (from ChEMBL)

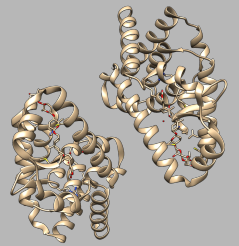




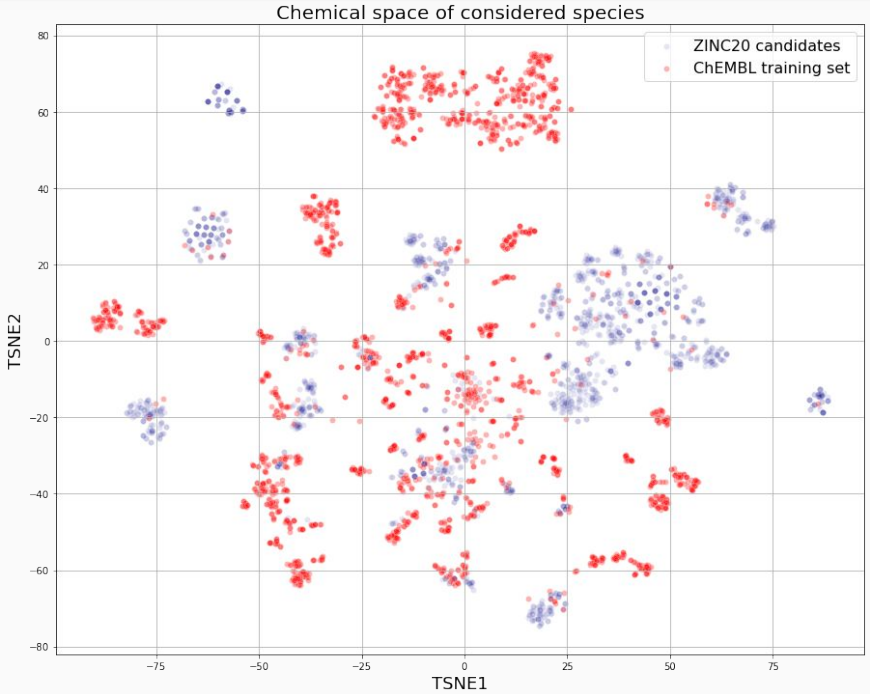


Already known potent RORγ antagonist not included in the ChEMBL database
Measured IC50: 30 nM

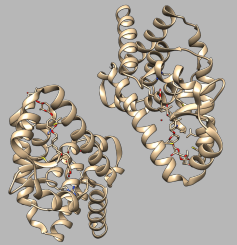
The chemical space analysis



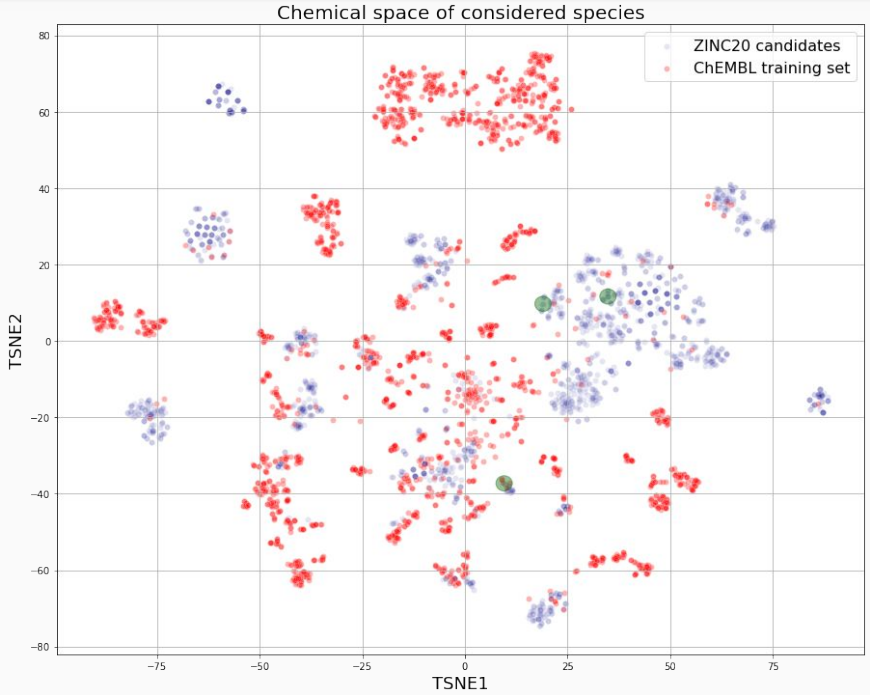
- The 2D representation of the chemical space
- The transformation:
  - 1024 Morgan fingerprint
  - PCA transformation into 1024 space of principal components
  - t-SNE into 2D space
- The area of applicability of the model
- Applied for the training set as well as for the set of potential inhibitors (ZINC20)
- Are the predictions trustworthy?
- The ML has the regression rather than extrapolation potential



The chemical space analysis



- The 2D representation of the chemical space
- The transformation:
  - 1024 Morgan fingerprint
  - PCA transformation into 1024 space of principal components
  - t-SNE into 2D space
- The area of applicability of the model
- Applied for the training set as well as for the set of potential inhibitors (ZINC20)
- Are the predictions trustworthy?
- The ML has the regression rather than extrapolation potential



Summary

- Introduction
- Nuclear receptors
- How to measure the biological activity?
- The QSAR workflow
- ZINC20 screening
- Summary

Acknowledgements

- **Thank you for your attention!**

- Marcin Ratajewski (for experiments and funds)
- Damian Nowak (for support with coding)
- Greg Landrum (for RDKit - a great library)
- Andrew Dalke (for chemfp)
- Polish National Science Centre Grant 2019/33/B/NZ7/00795