Re-Thinking Experimentation

(Q)SAR Model Reporting Format (QMRF) for iSafeRat® In Vitro Endocrine Modalities Prediction by SVM for thyroid receptors agonism/antagonism – iVEMPS - THR v1.0

1 QSAR identifier

1.1 QSAR identifier (title)

QSAR Model Reporting Format (QMRF) for iSafeRat® iVEMPS – THR model v1.0.

1.2 Other related models

No other related models.

1.3 Software coding the model

Language: Python (3.12.2)
Main libraries used:

- RDKit (2024.09.2)[1]
- pandas (2.2.2)
- matplotlib (3.9.2)
- scikit-learn (1.5.1)
- PyBioMed (1.0)

2 General information

2.1 Abstract

The iSafeRat® iVEMPS – THR model was developed to identify whether a mono-constituent organic substance can have an agonistic or antagonistic interaction with thyroid receptors alpha and beta.

This model is based on curated and knowledge expert-validated data from the TOX21 dataset available from US EPA Comptox Chemicals Dashboard[2] as well as from the BindingDB database[3]. Active or inactive results are generated from high-throughput *in vitro* screening studies, based on the principle of receptor transactivation with a reporter gene[4].

From the expert-curated data, a dataset consisting of 6,340 substances was compiled. This dataset was then divided into training, validation, and external validation sets while ensuring homogeneity across all three subsets. To account for structural similarity, the dataset was first clustered using the BUTINA method, and subsequently, within each cluster, molecules were assigned to the training, validation, and external validation sets using the Kennard-Stone method. The dataset consists of compounds from many different chemical families. Given this diversity, the structural space is very scattered, meaning that there are many substances in the dataset with no or few close analogues, within both the active and inactive classes. Furthermore, the dataset is highly imbalanced, with over 95% of the substances classified as inactive (i.e., neither agonists nor antagonists up to the cytotoxicity limit), and only 5% classified as active (i.e., either agonists or antagonists).

To address the class imbalance in the dataset, a Support Vector Machine (SVM) model utilizing circular fingerprints (FCFP6) was developed to classify substances into two categories: *active* (agonists or antagonists) and *inactive* (non-agonists and non-antagonists). This approach adapts to the specific challenges posed by the highly imbalanced dataset, as it provides a more effective way to separate the data points despite the disproportionate representation of active and inactive classes.

The model was trained on a dataset of 3,344 substances, with an internal validation set consisting of 1,184 substances and an external validation set of 1,812 substances. After optimizing the model's parameters, the SVM model demonstrated strong performance on the training set, achieving the following metrics for the *active*

class: precision of 95% and sensitivity of 98%. For the *inactive* class, the model achieved 100% precision and 100% specificity, indicating the model's effectiveness in identifying inactive compounds.

Next, an applicability domain was defined based on the structural features of the substances in the training set. Molecules from the validation and external validation sets that fell outside the applicability domain were excluded from further analysis. This ensures that predictions are made only for substances within the defined chemical space, increasing the reliability of the model.

The final evaluation on the external validation set revealed the following performance metrics: for the *active* class, the model achieved 92% precision and 98% sensitivity, while for the *inactive* class, it attained 99% precision and 99% specificity. These results validate the model's ability to predict molecular interactions with thyroid receptors (agonism and/or antagonism), or the absence of such activities, within its applicability domain.

2.2 Date of QMRF

17 January 2025

2.3 Date of QMRF update(s)

Table 1: Dates of QMRF updates.

Date	QMRF update identifier
17 January 2025	KTS/QMRF/ETR/01

2.4 QMRF update(s)

Table 2: Contents of QMRF updates.

(Q)PRF update identifier	Content
KTS/QMRF/ETR/01	First version of QMRF for this model

2.5 Model developer(s) and contact details

Dr. Emel Ay-Albrecht, Dr. Franklin Bauer, Dr. Zlatomir Todorov
KREATiS SAS,
ZAC Saint Hubert
23 rue du Creuzat
38080 L'ISLE D'ABEAU
France
Tel: +33 (0)6 46 46 42 33
Email: contact@kreatis.eu
Website: www.kreatis.eu

2.6 Date of model development and/or publication

The results presented in this QMRF refer to the version of the iSafeRat® iVEMPS – THR model v1.0 generated internally on 21 October 2024.

2.7 Reference(s) to main scientific papers and/or software package

US EPA (**2024**). *CompTox Chemicals Dashboard.* CompTox Chemicals Dashboard v2.5.0. https://comptox.epa.gov/dashboard/.

Gilson, M.K., and Liu, T. (**2023**). *BindingDB: Measured Binding Data for Protein-Ligand and Other Molecular Systems.* (UC San Diego Library Digital Collections). https://doi.org/10.6075/J0HD7VVF.

Freitas, J., Cano, P., Craig-Veit, C., Goodson, M.L., David Furlow, J., and Murk, A.J. (**2011**). *Detection of thyroid hormone receptor disruptors by a novel stable in vitro reporter gene assay. Toxicology in Vitro*, 25, 257–266. https://doi.org/10.1016/j.tiv.2010.08.013.

Landrum, G. (2024). RDKit. Version 2024.09.2. https://www.rdkit.org/

2.8 Availability of information about the model

The model is proprietary but limited information has been made publicly available for the validation/external validation set. Any queries concerning the model, or its validity should be addressed to contact@kreatis.eu. Furthermore, KREATiS undertakes to provide supplementary information to sponsors or regulatory authorities upon request to demonstrate compliance of our QSARs with good practice.

2.9 Availability of another QMRF for exactly the same model

None.

3 Defining the endpoint - OECD Principle 1

3.1 Species

Wistar Rat (Rattus norvegicus)

3.2 Endpoint

Interaction with thyroid receptors (agonism or antagonism), as could be measured in an *in vitro* assay on rat pituitary tumor GH3 cell line transfected with a luciferase reporter gene, as described by Freitas et al., 2011[4].

3.3 Comment on endpoint

Agonist endpoint and antagonist endpoint were regrouped as the unique "active" class for the purpose of this model, since this distinction was not available from BindingDB data, though it was available from TOX21 data.

3.4 Endpoint units

No unit.

3.5 Dependent variable

Presence or absence of interaction (agonistic or antagonistic) with thyroid receptor, which is formalised as "1" (active) or "0" (inactive) in the dataset, respectively.

3.6 Experimental protocol

In the TOX21 dataset, only the following assays have been used for this model:

- TOX21_TR_LUC_GH3_Agonist
- TOX21_TR_LUC_GH3_Antagonist
- TOX21_TR_LUC_GH3_Agonist
- TOX21_TR_LUC_GH3_Antagonist_viability

The testing protocol of these high-throughput assays is as follows:
A stable luciferase reporter gene assay was developed based on the thyroid hormone responsive rat pituitary tumor GH3 cell line that constitutively expresses both thyroid hormone receptor isoforms. Stable transfection of the pGL4CP-SV40-2xtaDR4 construct into the GH3 cells resulted in a highly sensitive cell line (GH3.TRE-Luc). GH3.TRE-Luc cells were incubated for 28h in the presence (for antagonism detection) or absence (for agonism detection) of T3, with or without the indicated test chemical in DMSO. The DMSO concentration was always the same for all exposures within an experiment and always kept 60.5% (v/v) to avoid cytotoxicity[4].
Gene transcription subsequent to thyroid receptor activation was monitored by the activity of the transcripted Luciferase. The bioluminescence signals were detected by CellTiter-Glo Luciferase-coupled ATP quantitation technology, on lysed cells in a microplate luminometer with two injectors (Thermo LabSystems luminoskan Ascent). Changes to bioluminescence signals produced from an enzymatic reaction involving the key substrate [One-Glo] are indicative of changes in transcriptional gene expression due to agonist or antagonist activity regulated by the human thyroid hormone receptor alpha, and thyroid hormone receptor beta [GeneSymbol: THRA & THRB | GeneID:7067 & 7068 | Uniprot_SwissProt_Accession:P10827 & P10828][2].
Cell viability in each well was determined by measuring Luciferase bioluminescence as a function of ATP content in the cell (since ATP is a required cofactor of Luciferase), since dying or dead cells stop producing ATP[2].
All exposures were performed in triplicate.

On the dose–response curves, percentages of maximal luciferase induction for each test compound were calculated by setting luciferase response to solvent control (DMSO) as 0% and the maximum luciferase induction by 10 nM T3 as 100%[4].

BindingDB was used to augment the iVEMPS training dataset with functionally active entries. Data within BindingDB is gathered upon manual curation of articles in selected scientific journals and US patents (no further data quality curation was made internally). It lists quantitative values of the activity and/or the affinity from in vitro experiments involving compounds and target proteins.

3.7 Endpoint data quality and variability

All data from TOX21 dataset have been tested with the same protocol and the resulting data have been processed the same way by US EPA tcpl workflow.
The TOX21 dataset has been automatically curated by KREATiS and then the remaining actives were further validated by KREATiS experts. The automatic curation of TOX21 data consisted of several steps, summarised here:

1. Search of SMILES codes from PubChem using InChI keys when SMILES code was not provided or invalid in the TOX21 database.
2. Standardisation of SMILES codes.
3. Removal of substances without a SMILES code.
4. Removal of substances that are multi-constituents.
5. Comparison of activity curves with cytotoxicity curves. If activity does not start at a significantly lower concentration than cytotoxicity, it is reassigned as “not active up to the cytotoxicity limit” or more simply “inactive”. Confirmed actives were labelled as “active”.

The following curation has been applied to BindingDB data:

1. To be consistent with TOX21 data, only the results expressed as EC50 and IC50 have been kept for this work, which is assumed to come only from *in cellulo* assays.
2. Results reported as censored values (*e.g.* EC50 > 10 µM) have been excluded.
3. Only the active data with an EC50 or IC50 lower than 100 nM have been kept for this work to avoid using so-called active data with low affinity that can easily be confounded with cytotoxicity effects.

Finally, after merging the data from TOX21 and BindingDB, a final step has been carried out:

1. If there are several samples tested for a same substance and which resulted in different outcomes (active/inactive), the substance has been removed from the dataset.

Dataset preparation:

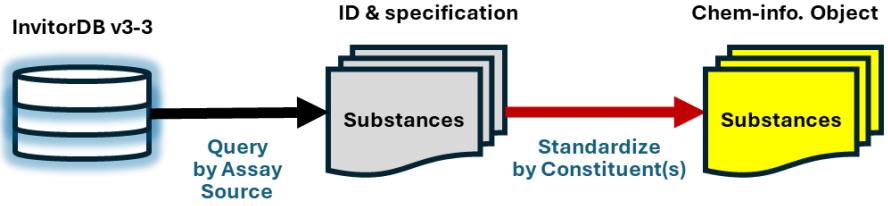
More detailed description of the data preparation procedure, summarised above, follows.

Standardization of chemical structure description:

Within the retrieved database identifiers (DTXSID, n = 9,403) from InvitroDB v3-3, not all of the associated compounds were successfully described *in silico* by unambiguous SMILES code. Available SMILES were standardized by removal of salt, stereochemistry was simplified and selection of a single converging tautomeric state was carried. Hence, alternative SMILES codes for the same compound were removed, resulting in 8,282 standardized SMILES allowing to describe (after salt removal) 7,155 mono-constituent and 1,816 multi-constituent substances.

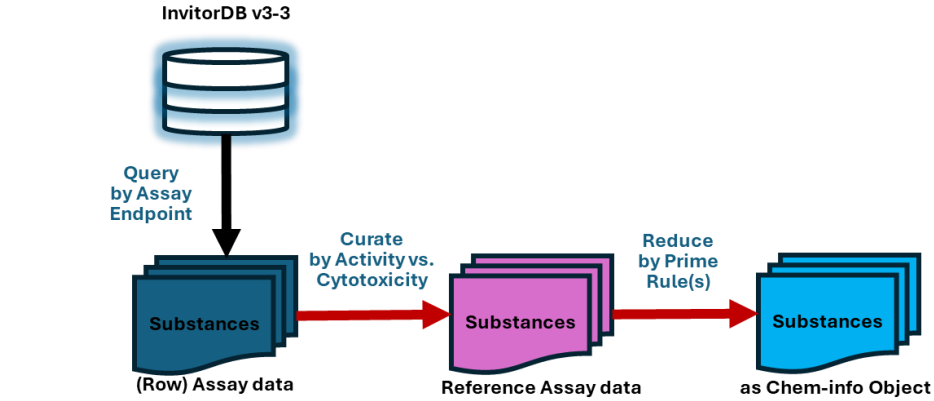
Data category	number
All entries	9403
No SMILES (rejected)	404 (434)
Mono-constituents	7155
Multi-constituents	1816

Object category	number
Unique SMILES (usmi)	8282
1 usmi → Substance	8464
2 usmi → Substance	503
3 usmi → Substance	4



Functional dataset curation:

The functional dataset TOX21_TR_LUC_GH3 from InvitroDB v3-3, was curated based on the methodology described elsewhere (The Thyroid Tox NC3Rs project. Not published yet). In brief, functional data classified as "Active" through binding on Thyroid Hormone Receptor was declassified to "Not active" if the activity threshold is very close to or above the observed cytotoxicity. Furthermore, compounds associated with conflicting functional classes assigned from repeated *in vitro* tests were removed from the reference dataset.



TOX21_TR_LUC_GH3_*	Substances	Samples
Agonist	35	40
Agonist	7947	10096
Antagonist	2318	2842
Antagonist	5846	7294

Substances	Samples
25	30
7954	10106
140	149
7872	9987

Substances	Samples
22	27
7951	10102
104	112
7836	9942

Workflow color code: US EPA, KREATiS
Table color code: Positive, Negative

Prime Rule:
Samples with same substance ID & different functional classes → remove from Ref. dataset

Linking structural descriptions and functional data:

As a matter of fact, not all substances associated with unambiguous functionality from TOX21 were successfully represented *in silico* with a standardized SMILES according to the procedure described in the text above. The product of both procedures "Standardization of chemical structure description" and "Functional dataset curation", resulted in the following counts:

TOX21_TR_LUC_GH3_*	Substances	Samples
Agonist	21	26
Agonist	7924	10067
Antagonist	103	111
Antagonist	7809	9907

Workflow color code: US EPA, KREATiS
Table color code: Positive, Negative

Dataset augmentation:

BindingDB data was used to bring additional "Active" compounds to the reference dataset. Data extraction involved selection for human Thyroid Hormone Receptor binding, and only mono-constituent compounds with reported EC50 or IC50 <0.1 µM (assumed high affinity, without further curation) were selected. Further on, the SMILES standardization procedure as described for the TOX21 dataset was applied. The resulting set of compounds (243 compounds) was merged with the already prepared reference dataset.

4 Defining the algorithm - OECD Principle 2

4.1 Type of model

Support Vector Machine (SVM) Model

4.2 Explicit method

Dataset preparation:

Dataset splitting:

The dataset splitting into training, validation and external validation sets proceeded according to a sophisticated procedure aiming to achieving homogeneity across the 3 subsets given the chemical diversity of the input. The approach ensures balancing the active and inactive classes within the subsets, but also tends to balance how the various chemical families are distributed within the 3 subsets.

Input: Reference dataset (SMILES) ➔ digitized as molecular fingerprints (ref.FPS)

Contains **Positives** and **Negatives for THRα/β activity**

Custom feature counting of PubChem Fingerprint ➔ ref.FPS

Pairwise similarities quantified as **Tanimoto index** (T_c) ➔ $T_c(ref.FPS_i, ref.FPS_j)$

Ensemble $T_c(ref.FPS_i, ref.FPS_j)$ ➔ square symmetric matrix (T_c_mat)

Split level 1: Functional classes

Split into Positive (P) and Negative (N) subsets

Split level 2: Structural similarity clustering

Iterate for ε *form* ε_{high} *to* ε_{low}

$\boldsymbol{\varepsilon_{high}} = 1 - T_c$ ← 99% CDF(T_c_mat); $\boldsymbol{\varepsilon_{low}} = 1 - T_c$ ← 65% CDF(T_c_mat)

Group substances *i* and *j*, if $1 - T_c(i, i) >= \varepsilon$, according to **Butina algorithm**

Compounds without neighbor at $T_c(i, i) >= \boldsymbol{\varepsilon_{low}}$ ➔ Singleton clusters

Split level 3: Selection of Train subset

Non-singleton clusters ($n > 2$) split by Kennard-Stone algorithm[2] ➔ ~80% Train

Non-singleton clusters ($n = 2$) ➔ random select ➔ 50% Train

Ensemble singleton clusters ($n = 1$) ➔ sort by decreasing $sum(T_c\ mat(i))$ ➔

stride along sorted $(T_c\ mat)_{singleton}$ ➔ ~80% Train

Split level 4: Selection of Validation subset

Non-singleton Train clusters ($n > 2$) split by Kennard-Stone algorithm[2] ➔ ~90% Train

Non-singleton Train clusters ($n = 2$) ➔ random select ➔ 50% Train

Ensemble Train singleton clusters ($n = 1$) ➔ sort by decreasing $sum(T_c\ mat(i))$ ➔

stride along sorted $(T_c\ mat)_{Train,\ singleton}$ ➔ ~90% Train

Output: Train and Test subsets

Ensemble substances selected as Train at level 4 ➔ **Train subset**

Ensemble substances selected as Validation at level 4 ➔ **Validation subset**

Ensemble substances selected as Test at level 3 ➔ **Test subset**

Additional filtering of the reference dataset:

Reference compounds associated with mono-constituent SMILES were preserved. All compounds containing atoms outside the following list were removed: allowed atoms **Li, Be, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Se, Br, I**. Moreover, in the final reference dataset the distinction between Agonistic and Antagonistic activity was ignored and replaced by "Active" ("Positive") and "Inactive" ("Negative"). This was due to the low number of active datapoints describing the Agonistic activity, and it was assumed that this could lead to difficulties of training an efficient "Agonist" model.

The final distribution of the reference database among the Train and External validation subsets is as follows (~10% of the Training subset were used as Validation subset):

Low epsilon (1-Tc)	High epsilon (1-Tc)	Split level 1 Active (entries)	Split level 2 Active (clusters)	Split level 2 Active (singleton)	Split level 3 train / test (entries)	Split level 1 Not active (entries)	Split level 2 Not active (clusters)	Split level 2 Not active (singleton)	Split level 3 train / test (entries)
0.005	0.35	371	173	88	0.744 / 0.256	6012	2178	505	0.712 / 0.288

Model description and optimisation parameters:

The model predicts whether a substance is active (agonist or antagonist to thyroid receptors) or inactive. This prediction is based on molecular fingerprints, which describe the structure of the molecules. It was built using a

training dataset of 3,344 samples and an internal validation set of 1,184 samples. The dataset is highly imbalanced, with 95% of the data representing inactive classes and only 5% representing active classes. Additionally, the structural space is very scattered, meaning that there are many substances in the dataset with no or few close analogues, within both the active and inactive classes.
To optimize the model, **GridSearchCV** from **Scikit-learn**[5], a widely used machine learning library for Python, was applied. This method performs an exhaustive search over a predefined parameter grid using cross-validation to identify the optimal hyperparameters, enhancing model performance and generalization. The optimization process included testing different types of molecular fingerprints, such as **substructure-based fingerprints** (e.g., **PubChem** and **MACCS keys**) and **circular fingerprints** (e.g., Feature-Class Fingerprints **FCFP6** and Extended-Connectivity Fingerprints **ECFP6**)[6]. The best results were achieved using circular fingerprints (FCFP6) as input features.
SVM was chosen for several key reasons: it is well-suited for medium-sized datasets and performs effectively in high-dimensional spaces, such as those generated by molecular fingerprints. SVM also handles class imbalance through techniques like class weighting, and its kernel flexibility allows it to generalize well to both linearly and non-linearly separable data.

A Support Vector Machine (SVM)[7,8] is a machine learning algorithm that finds the best boundary (called a **hyperplane**) to separate data points into different classes. The goal is to maximize the space, or **margin**, between the closest points of each class (called **support vectors**), making the model better at predicting new, unseen data. SVM can handle both simple and complex data by using **kernel functions**. These kernels transform the data into a higher-dimensional space where it becomes easier to separate the classes, even if they're not linearly separable in the original space.

In the present model, the following key parameters were fine-tuned using grid search:

- **C = 0.1**: This parameter controls how much the model allows for errors during training. By setting C to a lower value, the model focuses on creating a wider margin between the classes, even if it means misclassifying some points. This helps avoid overfitting and makes the model more general.
- **Gamma = 0.001**: Gamma determines data driven how much influence a single data point has on the boundary. Gamma is mostly critical for kernels like RBF (used for non-linear data).
- **Kernel = 'linear'**: A linear kernel means the model tries to separate the data with a straight line (or a flat plane in higher dimensions).

By carefully tuning these parameters, the model achieves a good balance between simplicity, accuracy, and generalization to new data.

4.3 Descriptors in the model

In this model, FCFP6 (Feature-Class Fingerprints)[6], represented as 1024-bit binary vectors, were used as descriptors. These fingerprints capture the functional features of molecules and are essential for classification tasks based on molecular similarity.

FCFP6 fingerprints are widely used in cheminformatics because they are data driven. **FCFP6 fingerprints** are generated dynamically. They focus on the functional groups attached to atoms, allowing them to capture a broader range of functional features compared to traditional fingerprints.

These fingerprints encode the chemical environment around each atom, especially emphasizing the functional groups, making them highly effective for understanding biological and chemical interactions. Their adaptability allows them to represent complex and diverse chemical environments, which predefined methods may not cover.

FCFP6 fingerprints are particularly well-suited for tasks like molecular similarity assessment, clustering, and predictive modelling. Their ability to represent molecular data comprehensively ensures accurate classification based on structural and functional similarities.

4.4 Descriptor selection

To optimise the model, grid search method was applied, testing different types of molecular fingerprints such as PubChem, MACCS key, and circular fingerprints. The best results were achieved using circular fingerprints (FCFP6) as input features.

4.5 Algorithm and descriptor generation

The **FCFP6 fingerprints** are generated using **RDKit**[1], specifically through the rdFingerprintGenerator class and the GetMorganGenerator() method, with **atom invariants** enabled to capture the atomic features.

FCFP6 (Functional-Class Extended-Connectivity Fingerprints) is an algorithm that generates molecular fingerprints focusing on the functional groups attached to atoms, in addition to atomic connectivity. Below is a summary of the key steps in the algorithm:

1. **Initialization**: Each atom is assigned an initial fingerprint based on its atomic number, hybridization state, and relevant chemical properties. Functional groups (e.g., hydroxyl or amine groups) are also associated with each atom.

2. **Neighbourhood Expansion**: The algorithm examines the neighbours of each atom within a 3-bond radius (or more) and iteratively expands the environment around each atom, considering both the functional groups and specific chemical properties.

3. **Encoding Functional Groups**: **FCFP6** encodes the presence of functional groups around each atom. This allows the algorithm to represent not only the structural environment but also the chemical functionality of the molecule.

4. **Hashing**: After encoding the functional group and connectivity information, the combined data is hashed into unique integer values that represent both structural and functional features within the local neighbourhood of each atom.

5. **Binary Vector Representation**: The final output is a binary vector, where each bit corresponds to the presence (1) or absence (0) of a specific functional feature or substructure identified through the hashing process.

This functional-class approach allows **FCFP6** to provide a more functional context, capturing both structural and chemical features, which is particularly useful for tasks like predicting biological activity or chemical reactivity.

4.6 Software name and version for descriptor generation

FCFP6 fingerprints are generated by an open-source cheminformatics library: **RDKit** version 2024.09.2 (https://www.rdkit.org)[1]

4.7 Chemicals/Descriptors ratio

- ✓ The ratio of the number of substances (3,344) to the number of descriptors (1,024) is **3.27**.

- ✓ In the current SVM model, the following parameters were fine-tuned using grid search: **C = 0.1**, **gamma = 0.001**, and **kernel = 'linear'**. These hyperparameters define how the model separates the data and manage the trade-off between margin maximization, error tolerance, and complexity of the decision boundary.

5 Defining the applicability domain - OECD Principle 3

5.1 Description of the applicability domain of the model

The process follows a binary classification model, categorizing compounds into two classes: "In Domain" and "Out of Domain." The classification is determined by whether the query compound is sufficiently structurally similar to compounds in the training set, as measured by the Tanimoto similarity index, using the MACCS keys

molecular fingerprints. A K-Nearest Neighbours (KNN) approach is applied to identify analogues. The objective is to identify compounds that are sufficiently similar to the query compound, based on a predefined similarity threshold (e.g., Tanimoto similarity ≥ 0.7), with **K** representing the number of analogues.

For a compound to be classified as "in domain," not only must its structural similarity to the analogues exceed the threshold, but its predicted value must also align with the predicted and experimental values of the analogues. This consistency between the predicted and experimental values across the query compound and its analogues ensures better reliability of the prediction and greater confidence in the model's applicability to the query compound (Figure 1).

Key Features of MACCS Keys

- **Fixed Dictionary of Features:**
- MACCS Keys are based on a predefined set of 166 structural fragments, such as aromatic rings, halogens, and specific functional groups. Each key corresponds to the presence or absence of a particular feature in the molecule.
- **Binary Representation:**
- The molecule is represented as a binary vector of 166 bits:
- A value of 1 indicates the presence of a specific structural feature.
- A value of 0 indicates the absence of that feature.

Optimized Parameters for Applicability Domain

The applicability domain of the model is assessed based on two key parameters:

1. **Threshold (t):** A Tanimoto similarity threshold of 0.7 is used to determine the structural similarity between the query compound and compounds in the training set. Compounds with a Tanimoto similarity equal to or greater than this threshold are considered structurally similar.
2. **K (Number of Neighbors):** The model evaluates the query compound's analogues using a K-Nearest Neighbours (KNN) approach, where **K** represents the number of neighbours ($1 \leq K \leq 3$) used in the analysis. This helps identify the most similar compounds in the training set for comparison.

By using these fixed parameters, the model ensures that only compounds with sufficient similarity to the training set are considered for predictions, providing a reliable assessment of the applicability domain.

Binary Classification: "In Domain" vs. "Out of Domain"

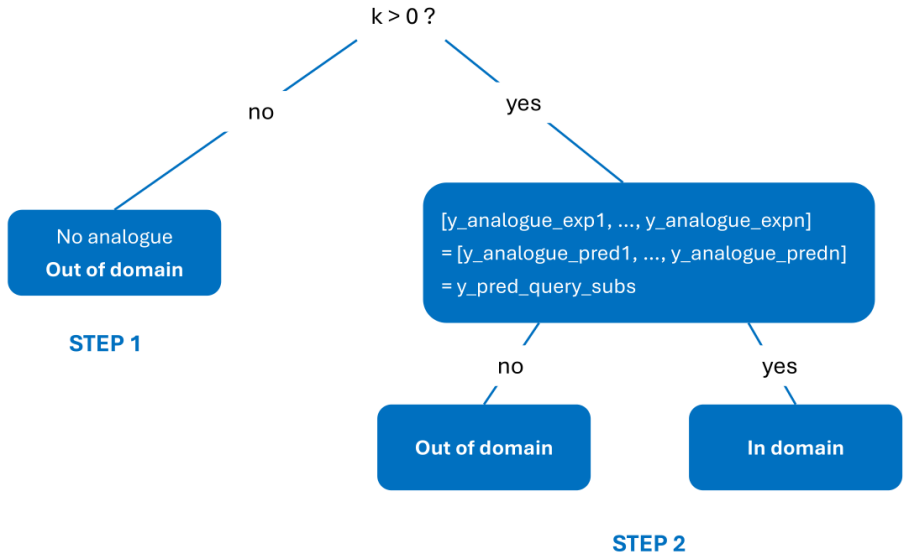


Figure 1: Applicability Domain method

First, the algorithm calculates and identifies the number of analogues (K) from the query substance using Tanimoto similarity with a threshold ≥ 0.7, by comparing the structure of the query compound with those in the training set.

Step 1:

- ✓ **If K = 0:** No analogues are found.

The query compound is classified as *out of domain* (i.e., outside the applicability domain of the model).

Step 2:

- ✓ **If K > 0:** Analogues are found. For each identified analogue, the model predicts values (e.g., y_analogue_pred1, y_analogue_pred2, ..., y_analogue_predn).

    - If the predicted values of the query compound are consistent with both the experimental and predicted values of the analogues, the query compound is classified as *in domain*. Specifically, this means:
        a. All analogues have the same experimental class.
        b. All analogues have the same predicted class.
        c. The predicted class of the query compound matches both the experimental and predicted classes of the analogues.
    - Otherwise, the query compound is classified as out of domain

In this two-step process, the KNN approach with a threshold of ≥ 0.7 ensures that only compounds with sufficient structural similarity to those in the training set are considered for prediction. This not only improves the reliability of the model by ensuring the query compound is well-represented within its local neighbourhood but also takes into account the predicted values and experimental results of its analogues, further enhancing the confidence in the prediction.

a) Fixed or probabilistic boundaries

The model uses **fixed boundaries** to assess applicability:

1. **Threshold (t)**: Tanimoto similarity threshold of **0.7**.
2. **K**: The number of neighbours (1=<K>=3) used to evaluate query compound analogues.

b) Response domain

In the developed model, the **response domain** is binary, with predictions classified as:

- **1 (active)**: Indicates that the compound or query is associated with the predicted property or activity of interest, i.e. agonist and/or antagonist at the thyroid receptor.
- **0 (inactive)**: Indicates the absence of the property or activity, i.e. not agonist and not antagonist up to the level of cytotoxicity.

c) Descriptor domain

As the descriptors are the FCFP6 fingerprints, which, as circular fingerprints, can be generated for any structure as far as the SMILES can be read by RDKit, the descriptor domain would only be limited by what types of SMILES can be read and interpreted correctly by RDKit.

d) Structural fragment domain

Following the method detailed in section 5.2, if a substance does not have any close analogues (Tanimoto less than the threshold of 0.7) in the training set, it is outside the structural fragment domain of the model. If the substance has close analogues, a maximum of 3 of the closest analogues are considered. If the predictions of the test substance, the predictions of the closest analogues and the experimental classes of the closest analogues are consistent with each other, then the substance is inside the structural fragment domain, otherwise it is outside of the structural fragment domain.

The choice of FCFP6 for the SVM model and MACCS keys for the applicability domain (AD) is based on their respective strengths. FCFP6 provides fine granularity, ideal for classification, while MACCS keys, being more generic, are better suited for similarity calculations required for AD. MACCS keys yield more stable results for AD evaluation due to their simplified representation, reducing biases linked to specific molecular details. This compromise maximizes model performance while ensuring robust AD assessment. Each fingerprint is utilized according to its relevance to the targeted task.

e) Mechanistic domain

All molecular mechanisms leading to an agonistic or antagonistic response that is distinct from cytotoxicity signals are detected, and with a proper Hill curve or gain-loss curve would be classified identically by *in cellulo* assays that serve as basis for the datasets of this model. In most cases, the molecular mechanisms supposed to happen are the competitive binding of the molecule to the thyroid receptor hormone binding site, and then the bound ligand either activates the receptor thus engaging the cascade of events leading to the transcription of the reporter gene (agonism), or the bound ligand does not activate the receptor (thus blocking the receptor in its inactive state) and thus not engaging the cascade (antagonism). However, similar reporter gene signals can be obtained if the molecule binds to an allosteric binding pocket on the receptor, inducing or blocking a change of conformation of the protein. Other unspecific molecular interactions could be interpreted as agonism or antagonism, but most of these should have been ruled out by the cytotoxicity signal. Examples of unspecific interactions that would give both an antagonist signal and a cytotoxicity signal in the same range of concentrations are:

- Absorption of light at the same wavelengths as the light produced by the luciferase activity[9],
- Inhibition of some step of the cascade of events leading to gene transcription or inhibition of gene transcription,
- Unspecific adducts formation (covalent bonds) with proteins, which would denaturate both thyroid receptors and luciferase[10],
- Fast degradation of the luciferase substrate (One-Glo),
- Inhibition of luciferase[9].

Since these mechanisms would be ruled out by the cytotoxicity signal[11], they are out of the mechanistic domain.

f) Metabolic domain, if relevant

As this model predicts what would happen in an assay on a rat pituitary cell line, metabolic activity of these cells is intrinsically included in the model even if the exact metabolic cascade for each substance is not known. It is indeed probable that some substances have been tested active in these *in vitro* tests due to metabolic transformation and receptor binding of the metabolite, or conversely tested inactive because the metabolites do not bind to the receptors while the parent substance may have the ability to bind.

g) Possible defined (graphical) expression of how the descriptor values of the chemicals in the training set are distributed in relation to the endpoint values predicted by the model.

Not relevant.

5.2 Method used to assess the applicability domain

The evaluation of a model's applicability domain (AD) aims to verify whether the AD method is sufficiently effective in detecting misclassified predictions by the SVM model. The confusion matrix provides insight into the performance of the AD method.

The AD Confusion matrix is structured as follows:

Table 3: Confusion matrix of AD

AD model / SVM model	Predicted X-Outliers (X-O)	Predicted X-Inliers (X-I)
True Y-Outliers (Y-O)	True Outliers (TO)	False Inliers (FI)
True Y-Inliers (Y-I)	False Outliers (FO)	True Inliers (TI)

Predicted X-Outliers (X-O): Instances predicted as outliers by the AD model, **Predicted X-Inliers (X-I)**: Instances predicted as inliers by the AD model, **True Y-Outliers (Y-O)**: Instances that are incorrectly predicted by the SVM model, **True Y-Inliers (Y-I)**: Instances that are correctly predicted by the SVM model, **True Inliers (TI)**: Instances that are correctly predicted by the SVM model and lie inside the applicability domain, **True Outliers (TO)**:

Instances that are incorrectly predicted by the SVM model and lie outside the applicability domain, **False Outliers (FO)**: Instances that are correctly predicted by the SVM model but incorrectly predicted as outliers by the AD model, **False Inliers (FI)**: Instances that are incorrectly predicted by SVM but incorrectly predicted as inliers by the AD model.

To assess the AD method's performance, the **balanced accuracy** metric has been used. The balanced accuracy formula is as follows:

$$Balanced\ Accuracy = \frac{TPR + TNR}{2}$$

- **TPR (True Positive Rate)**: This is the proportion of actual outliers (Y-O) that are correctly identified as outliers by the AD model. It measures the AD model's ability to detect outliers.
- **TNR (True Negative Rate)**: This is the proportion of actual inliers (Y-I) that are correctly identified as inliers by the AD model. It measures the AD model's ability to identify inliers.

It is also essential to take into account the rate of molecules excluded by the algorithm when evaluating the applicability domain (AD). However, it is crucial to maintain a balance between two important objectives:

1. **A low exclusion rate**: This means that the model includes the majority of examples within its applicability domain, allowing the model to make predictions on a wide range of data, without overly restricting the set of usable examples.
2. **Excluding misclassified examples**: At the same time, it is important for the model to appropriately exclude examples that are truly outside its applicability domain, meaning those that are misclassified by the SVM model.

$$Exclusion\ rate = \frac{Number\ of\ excluded\ examples}{Total\ number\ of\ examples} \cdot 100$$

5.3 Software name and version for applicability domain assessment

The applicability domain assessment was conducted using **Python** with libraries such as **RDKit** (2024.09.2) for MACCS key fingerprints.

5.4 Limits of applicability

A substance is fully within the applicability domain of the model if it fits within the descriptor and structural fragments domains of the model. The following situations can occur:

1. Descriptor domain
   a. A substance is inside descriptor domain if the SMILES can be handled by RDKit.
   b. A substance is outside descriptor domain if the SMILES cannot be handled by RDKit.
2. Structural fragment domain
   a. A substance is inside structural fragment domain if it has close analogues with fully consistent outcomes, as detailed in section 5.1.d and 5.2.
   b. A substance is outside structural fragment domain if it does not have any close analogues or if it does have close analogues, but with inconsistent outcomes, as detailed in section 5.1.d and 5.2.

If the substance is outside descriptor or structural fragment domain, it is thus outside global applicability domain of the model. If it is inside descriptor domain and inside structural fragment domain, it is thus inside global applicability domain of the model.

6 Defining goodness-of-fit and robustness (internal validation) - OECD principle 4

6.1 Availability of the training set

The training set of the model is proprietary and has not been made publicly available. The training set of the model may be shared with regulatory authorities upon their request.

6.2 Available information for the training set

CAS RN: Yes (confidential business information)
Chemical Name: Yes (confidential business information)
SMILES: Yes (confidential business information)
Formula: No
INChI: No
MOL file: No
For substances from TOX21 dataset: DTXSID (identifier from comptox dashboard): Yes (confidential business information)
For substances from BindingDB dataset: PubChem CID (identifier from PubChem database): Yes (confidential business information)

6.3 Data for each descriptor variable for the training set

Descriptors are molecular fingerprints, which are stored in a file internally at KREATiS, but not made publicly available since the dataset is proprietary. However, KREATiS undertakes to provide supplementary information to sponsors or regulatory authorities upon request to demonstrate compliance of our QSARs with good practice.

6.4 Data for the dependent variable for the training set

The training set of the model is proprietary and has not been made publicly available yet.
Data available for substances from TOX21 dataset: outcome of the agonist assay, outcome of the antagonist assay, outcome of the cytotoxicity assay, “1” (active) or “0” (inactive) assignment for thyroid receptor interaction.
Data available for substances from BindingDB dataset: “1” (active) assignment for thyroid receptor interaction (only actives available from this source).
Data can be provided to regulatory authorities upon request.

6.5 Other information about the training set

None.

6.6 Pre-processing of data before modelling

Agonist endpoint and antagonist endpoint were regrouped as the unique “active” class for the purpose of this model, since this distinction was not available from BindingDB data, though it was available from TOX21 data. “Inactive” class was assigned if there was absence of both agonist and antagonist activity up to the cytotoxicity limit.
In case of conflicting evidence of positivity between TOX21 data and BindingDB data for a same substance, the substance was removed from the dataset.

6.7 Statistics for goodness-of-fit

A confusion matrix is plotted below, mapping the number of true and false positives and negatives from the training set of the model:

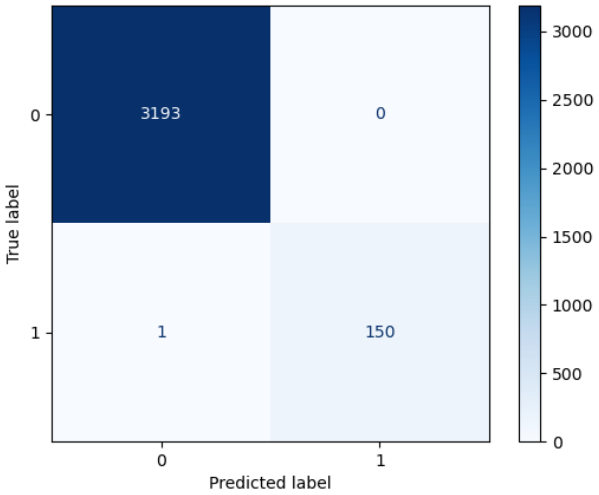


Figure 2: Confusion matrix for the training set

The confusion matrix (**Figure 2**) shows that the model is extremely effective. It correctly predicts all examples of the majority class (3,193/3,193) while also efficiently capturing the minority class, with only one false negative out of 151 instances (150 true positives and 1 false negative). This indicates an exceptional performance, particularly in the context of a highly imbalanced dataset. The statistics (Table 4) for the minority class reinforce this observation. With a **sensitivity of 99.34%** and **precision of 100%**, the model identifies nearly all examples of the minority class (class 1) with high precision, despite its low proportion in the dataset (4.52%). Additionally, with a **specificity of 100%**, it perfectly classifies all examples of the majority class (class 0), avoiding any false positives. These results demonstrate that the model maintains high performance on both classes, even in a highly imbalanced data context, by balancing the needs of identifying minority-class instances while maintaining error-free classification of the majority class.

In comparison, a naïve model that always predicts the majority class (class 0) would achieve an accuracy of **95.48%**, corresponding to the proportion of this class in the dataset. The present model, with an accuracy of **99.97%**, far exceeds this threshold. This proves that it effectively captures examples of the minority class, a task that is typically challenging in such scenarios due to the inherent imbalance.

The **balanced accuracy of 99.67%** further emphasizes the model's ability to handle class imbalance. Unlike standard accuracy, which may be misleading in imbalanced datasets, balanced accuracy gives equal importance to the performance on both classes. The high balanced accuracy confirms that the model is robust and fair in its treatment of the minority class while maintaining its effectiveness for the majority class.

Moreover, the **AUC-ROC (99.97%)** (**Figure 3a**) and **AUC-PR (99.69%)** (**Figure 3b**) metrics underline the model's outstanding performance. The AUC-ROC score highlights the model's ability to distinguish between classes across all decision thresholds, with a near-perfect value indicating excellent discriminative power. Meanwhile, the AUC-PR score, which is particularly informative in imbalanced datasets, shows the precision-recall trade-off, confirming that the model excels in identifying the minority class without compromising precision.

In summary, these results highlight several key points:

- The model effectively addresses the challenges posed by class imbalance, achieving a very high sensitivity (99.34%) for the minority class and perfect specificity (100%) for the majority class.
- It outperforms a naïve model by a significant margin, with a high overall accuracy (99.97%) and balanced accuracy (99.67%).
- The exceptional values for AUC-ROC and AUC-PR demonstrate the model's reliability in distinguishing and correctly classifying both classes.

These findings demonstrate that the present model is both **well-balanced** and **highly effective** across all classes, which is a notable achievement in problems involving imbalanced datasets.

Table 4: Statistics for goodness-of-fit

Statistics	Values
Concordance or Accuracy (Non-error Rate)	99.97%
Error Rate	0.03%
NO-MODEL Error Rate, NOMER%	4.52%
Prior probability of a class	95.48%
Prior proportional probability of a class	4.52%
Sensitivity of a class 1 (Active)	99.34%
Precision of a class 1 (Active)	100.00%
F1-score of a class 1 (Active)	99.67%
Precision of a class 0 (Inactive)	99.97%
Specificity of a class 0 (Inactive)	100%
F1-score of a class 0 (Inactive)	99.98%
Misclassification risk	0.03%
Balanced accuracy	99.67%
auc_roc	99.97% (**Figure 3a)**
auc_pr	99.69% (**Figure 3b)**

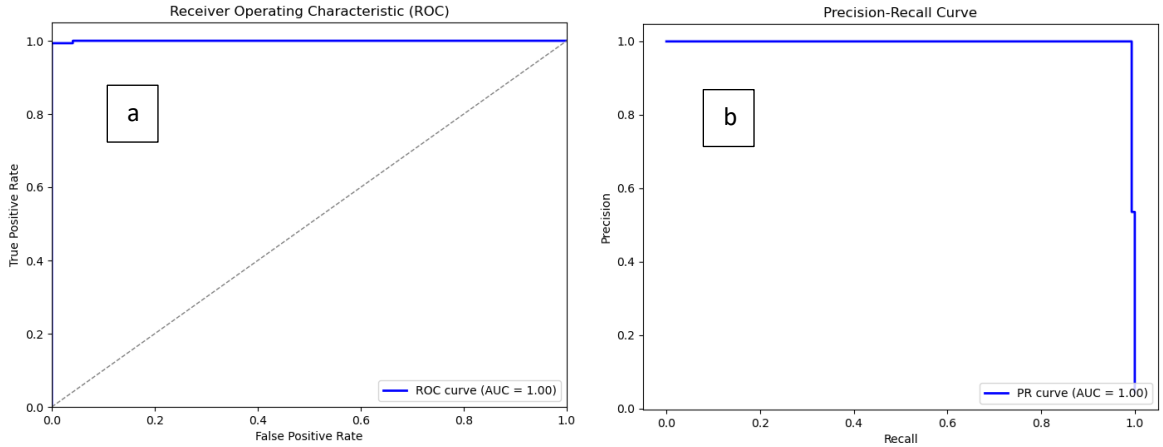


Figure 3: AUC-ROC (a) and AUC-PR (b) for the training set

6.8 Robustness - Statistics obtained by leave-one-out cross-validation

This is not applicable.

6.9 Robustness - Statistics obtained by leave-many-out cross-validation

A leave-many-out cross-validation of the model has been performed, by dividing the training set into 5 folds (k-folds cross validation, with k = 5). This dataset splitting has been done the same way as for the splitting of training, validation and test sets, ensuring homogeneity across the 5 folds.
The statistics obtained for each fold and the mean of all folds are presented in the tables below (**Tables 5a-g**).

Tables 5a-g: Cross validation statistics

a. Fold 1

Class	Precision	Sensitivity	F1-score	Number of substances
0 (inactive)	0.9905	0.9905	0.9905	629
1 (active)	0.7778	0.7778	0.7778	27
Accuracy	0.9817			

b. Fold 2

Class	Precision	Sensitivity	F1-score	Number of substances
0 (inactive)	0.9829	0.9953	0.9891	637
1 (active)	0.8800	0.6667	0.7586	33
Accuracy	0.9791			

c. Fold 3

Class	Precision	Sensitivity	F1-score	Number of substances
0 (inactive)	0.9889	0.9952	0.9920	625
1 (active)	0.8929	0.7813	0.8333	32
Accuracy	0.9848			

d. Fold 4

Class	Precision	Sensitivity	F1-score	Number of substances
0 (inactive)	0.9985	0.9863	0.9923	656
1 (active)	0.7500	0.9643	0.8438	28
Accuracy	0.9854			

e. Fold 5

Class	Precision	Sensitivity	F1-score	Number of substances
0 (inactive)	0.9938	0.9985	0.9961	646
1 (active)	0.9643	0.8710	0.9153	31
Accuracy	0.9926			

f. Mean of the 5 folds

Class	Precision	Sensitivity	F1-score
0 (inactive)	0.9909	0.9931	0.9920
1 (active)	0.8530	0.8122	0.8257
Accuracy	0.9847		

g. Standard deviation of each statistical parameter along the 5 folds

Class	Precision	Sensitivity	F1-score
0 (inactive)	0.0052	0.0043	0.0024
1 (active)	0.0787	0.0999	0.0551
Accuracy	0.0051		

Class 0 (inactive) shows stable performance, with high precision, sensitivity, and F1 scores, and a low standard deviation, indicating good consistency across the different folds.

However, class 1 (active) shows weaker performance, with lower precision, sensitivity, and F1-score compared to class 0. The standard deviation is higher for class 1 compared to class 0, which suggests that the classification of this class is more variable from one fold to another. This could indicate difficulty in correctly identifying this class in certain folds. This phenomenon could be due both to the class imbalance and the presence of a large number of singletons, which may have affected the data splitting method.

6.10 Robustness - Statistics obtained by Y-scrambling

This is not applicable.

6.11 Robustness - Statistics obtained by bootstrap

This is not applicable.

6.12 Robustness - Statistics obtained by other methods

None.

7 Defining predictivity (External validation) - OECD Principle 4

7.1 Availability of the external validation set

The external validation set of the model is proprietary and has not been made publicly available. The external validation set of the model may be shared with regulatory authorities upon their request.

7.2 Available information for the external validation set

CAS RN: Yes (confidential business information)
Chemical Name: Yes (confidential business information)
SMILES: Yes (confidential business information)
Formula: No
INChI: No
MOL file: No
For substances from TOX21 dataset: DTXSID (identifier from comptox dashboard): Yes (confidential business information)
For substances from BindingDB dataset: PubChem CID (identifier from PubChem database): Yes (confidential business information)

7.3 Data for each descriptor variable for the external validation set

Descriptors are molecular fingerprints, which are stored in a file internally at KREATiS, but not made publicly available since the dataset is proprietary. However, KREATiS undertakes to provide supplementary information to sponsors or regulatory authorities upon request to demonstrate compliance of our QSARs with good practice.

7.4 Data for the dependent variable for the external validation set

The external validation set of the model is proprietary and has not been made publicly available yet.
Data available for substances from TOX21 dataset: outcome of the agonist assay, outcome of the antagonist assay, outcome of the cytotoxicity assay, “1” (active) or “0” (inactive) assignment for thyroid receptor interaction.
Data available for substances from BindingDB dataset: “1” (active) assignment for thyroid receptor interaction (only actives available from this source).
Data can be provided to regulatory authorities upon request.

7.5 Other information about the external validation set

None.

7.6 Experimental design of test set

As explained in section 4.2, the reference dataset has been split into a training validation and external validation set (test set = external validation set), so that structural diversity and active/inactive proportions are homogenous between these different subsets. As such, all data in the external validation set are not present in the training set of the model.

7.7 Predictivity - Statistics obtained by external validation

1. **Statistics obtained before applying Applicability Domain algorithm**

Before the application of the applicability domain (AD) algorithm, the model demonstrated excellent prediction capability for the majority class (Inactive), with 1,724 true positives and only 7 false positives (**Figure 4**). This reflects a high level of reliability for this class. However, the minority class (Active) showed more nuanced results. While the sensitivity was relatively high at 83.95%, with 68 true positives out of 81, the model still had 13 false negatives, indicating difficulties in correctly identifying certain instances of this class (**Table 6**). These results highlight the model's limitations, particularly due to class imbalance, which is a common challenge in this type of problem.

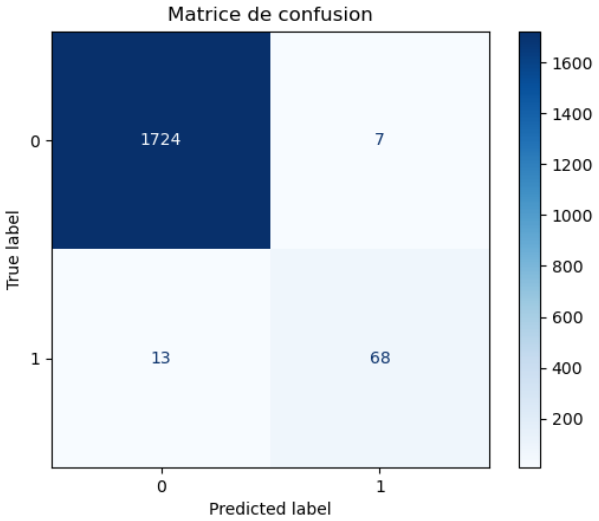


Figure 4: Confusion Matrix for the external validation set before applying AD

Table 6: Statistics for external validation before applying AD

Statistics	Values
Concordance or Accuracy (Non-error Rate)	98.90%
Error Rate	1.10%
Sensitivity of a class 1 (Active)	83.95%
Precision of a class 1 (Active)	90.67%
F1-score of a class 1 (Active)	87.12%
Precision of a class 0 (Inactive)	99.25%
Specificity of a class 0 (Inactive)	99.60%
F1-score of a class 0 (Inactive)	99.42%
Misclassification risk	1.10%
Balanced accuracy	91.77%
auc_roc	93.84% (**Figure 5a**)
auc_pr	88.49% (**Figure 5b**)

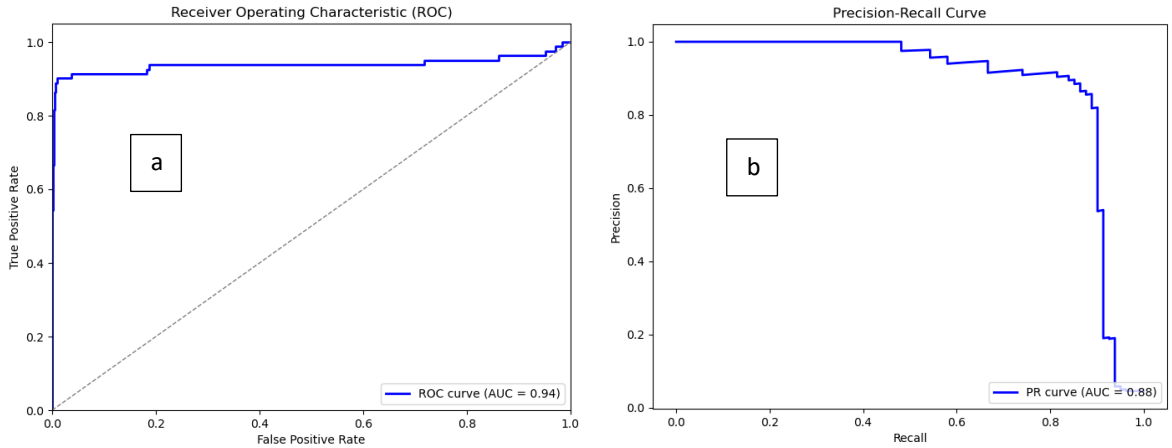


Figure 5: AUC-ROC (a) and AUC-PR (b) for the external validation set before applying AD

2. Statistics obtained after applying Applicability Domain algorithm

The application of the AD improves these results by filtering out misclassified examples. The AD identifies certain cases as outliers, categorized into a distinct group, "Predicted X-Outliers (X-O):" (see section 5.1 and 5.2). Among these cases, 238 molecules are deemed out of domain. However, these molecules had been correctly classified by the initial model, suggesting that, although they are seen as uncertain by the AD, their initial classification remains reliable. This illustrates the conservative role of the AD, which excludes examples with higher uncertainty to ensure better overall robustness of the model (**Figure 6**).

The rate of molecules excluded by the algorithm is approximately 13.13%. This proportion of excluded examples highlights the importance of the AD in contexts where prediction reliability is crucial. However, this exclusion raises the question of the balance to maintain between reducing prediction errors and preserving a sufficient number of examples for meaningful classification. This exclusion phenomenon can be broken down into two distinct steps:

First step: **191 molecules excluded**: These molecules are singletons, meaning they don't have any close analogue in the training set. They do not share enough characteristics with other molecules and are therefore considered out of domain. **Second step: 61 molecules excluded:** These molecules are poorly represented locally, which leads to uncertain predictions.

The algorithm acts conservatively by excluding these examples to enhance the reliability and robustness of the predictions.

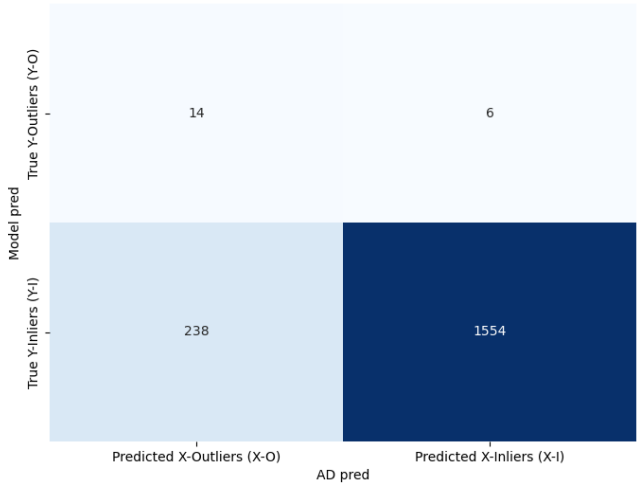


Figure 6: Confusion Matrix for the assessment of AD algorithm

After the full application of the AD, the model's performance improves significantly. The compounds that were out of domain were dropped, and the statistics were recalculated. The majority class (Inactive) maintains an exceptional specificity of 99.93%, with 1,496 true negatives and only 1 false positive (**Figure 7**). The minority class (Active), on the other hand, shows a notable improvement, with a sensitivity of 92.06% (58 true positives) and a reduction of false negatives to just 5. These results demonstrate that the model has been improved after the application of the AD and is now better able to correctly identify the minority class while maintaining very high precision.

The balanced accuracy, reflecting the balance between the two classes, also improves, rising from 91.77% to 96%, confirming an overall performance boost (**Table 7**).

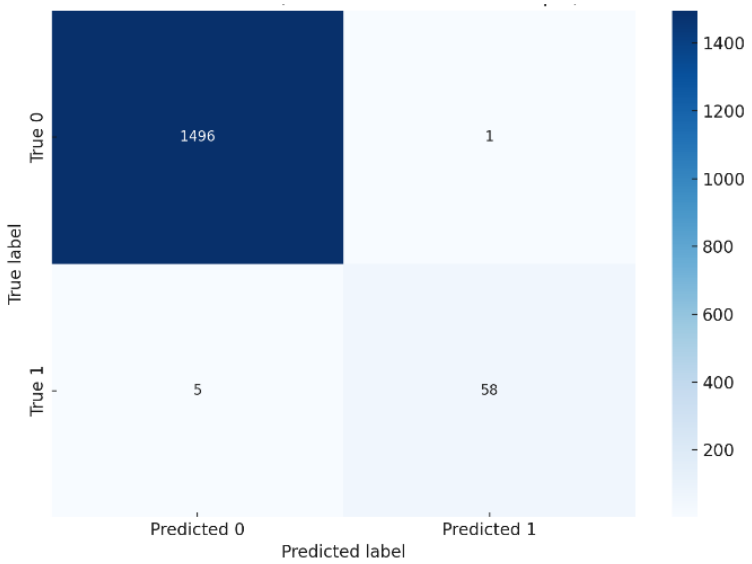


Figure 7: Confusion Matrix for the external validation set after applying AD

Table 7: Statistics for External Validation after applying AD

Statistics	Values
Concordance or Accuracy (Non-error Rate)	99.62%
Error Rate	0.38%
Sensitivity of a class 1 (Active)	92.06%
Precision of a class 1 (Active)	98.31%
F1-score of a class 1 (Active)	95%
Precision of a class 0 (Inactive)	99.67%
Specificity of a class 0 (Inactive)	99.93%
F1-score of a class 0 (Inactive)	100%
Misclassification risk	0.38%
Balanced accuracy	95.6%

In summary, the application of the AD has an undeniable positive impact on the model's performance. It enhances the handling of class imbalance while effectively filtering out out-of-domain examples, thereby strengthening the model's robustness and overall efficiency. This improvement is particularly noticeable in the classification of the minority class.

7.8 Predictivity - Assessment of the external validation set

To ensure the representativity of the external validation set and its compatibility with the training set, multiple statistical and visual analyses were conducted. These tests aim to verify that the two datasets share similar descriptor distributions, minimizing potential biases and ensuring robust model evaluation. The following sections summarize the key findings.

- ✓ **Kolmogorov-Smirnov Test (KS-Test):**
  The KS-Test was applied to compare the distribution of each descriptor between the **training** and **external validation** sets:
  - **Total number of descriptors:** N=1024
  - **Descriptors with statistically significant differences (ρ < 0.05):** 1

  This result suggests that almost all descriptors share similar distributions between the two datasets, indicating overall representativity.

✓ **Descriptor Exclusivity:**

- **Descriptors present only in the training set: 0**
- **Descriptors present only in the external validation set: 0**

These findings confirm that both datasets share a common descriptor space.

✓ **Average Wasserstein Distance:**

The mean Wasserstein distance between the descriptor distributions of the **training** and **external validation** sets is **0.0034**. This low value highlights a high degree of similarity between the two datasets.

✓ **Pearson Correlation:**

Overall correlation between descriptor proportions: 0.998. The Pearson correlation of 0.998 between the descriptor proportions indicates an almost perfect similarity between the training and external validation sets, ensuring good model generalization.

To further explore the representativity of the two datasets, additional visual analyses were conducted:

- **Descriptor Proportions:**
- The proportions of the most represented descriptors in the training and external validation sets are visualized using a histogram. This graphical representation allows for a clear comparison of how these descriptors are distributed across the two datasets (**Figure 8**).
- The histograms of descriptor distributions reveal comparable proportions between the training and external validation sets, indicating a high level of consistency. Such alignment is crucial for ensuring that models trained on the training set generalize well to the external validation set, minimizing potential biases caused by distributional differences.

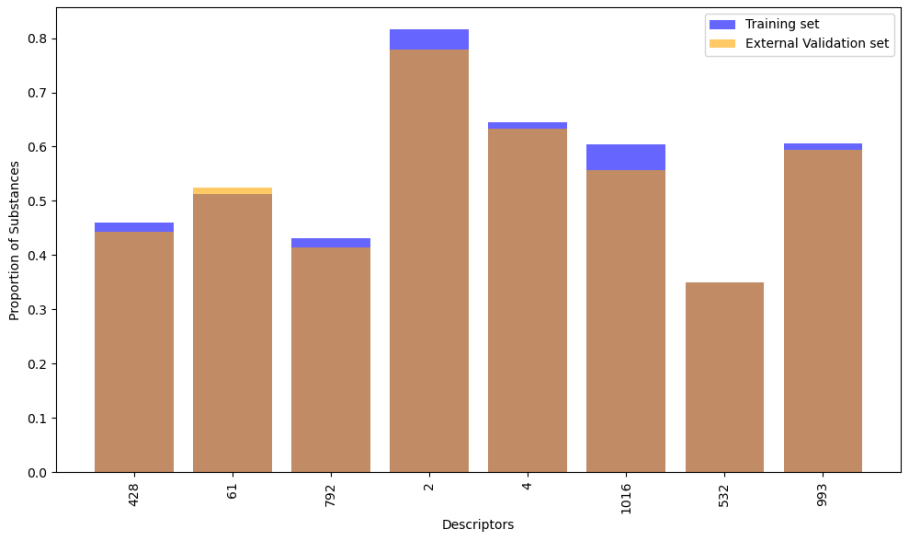


Figure 8: Proportions of the most represented descriptors (Training vs External validation)

- **t-SNE Projection:**
  A **t-Distributed Stochastic Neighbour Embedding** (t-SNE)[12] projection was used to visualize the clustering of molecules in the **training** and **external validation** sets. The results show significant overlap between the two datasets, supporting their mutual representativity (**Figure 9**).

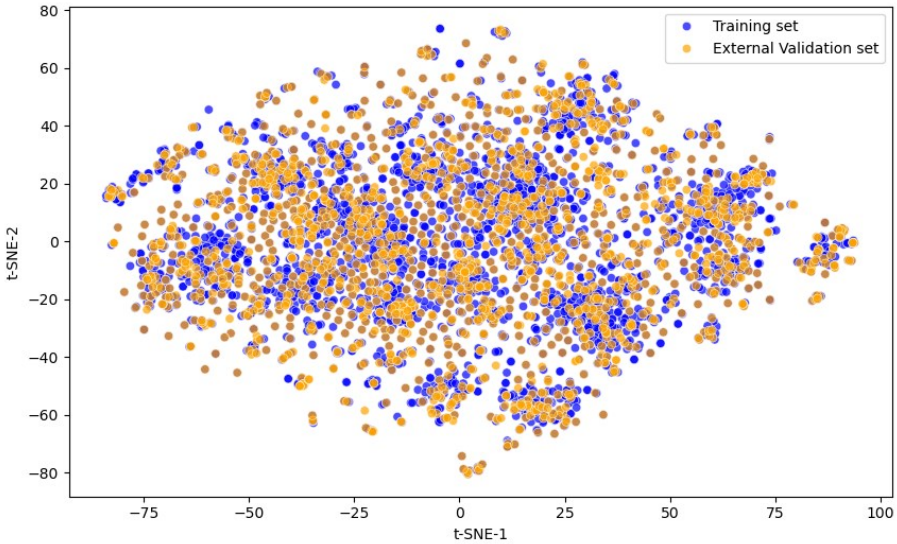


Figure 9: Visualisation of Training and External validation set by t-SNE projection

The statistical and visual analyses confirm that the training and external validation sets are well-aligned in terms of descriptor distributions. This minimizes the risk of systemic bias and ensures that model performance on the external validation set is reflective of its behaviour on the training set.

7.9 Comments on the external validation of the approach

None.

8 Providing a mechanistic interpretation - OECD Principle 5

8.1 Mechanistic basis of the model

No mechanistic interpretation of the descriptors (structural features) and algorithm were investigated.

8.2 A priori or a posteriori mechanistic interpretation

No mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

None.

9 Miscellaneous information

9.1 Comments

None.

9.2 Bibliography

1. Landrum, G. (2024). RDKit. Version 2024.09.2.

2. US EPA (2024). CompTox Chemicals Dashboard. CompTox Chemicals Dashboard v2.5.0. https://comptox.epa.gov/dashboard/.

3. Gilson, M.K., and Liu, T. (2023). BindingDB: Measured Binding Data for Protein-Ligand and Other Molecular Systems. (UC San Diego Library Digital Collections). https://doi.org/10.6075/J0HD7VVF https://doi.org/10.6075/J0HD7VVF.

4. Freitas, J., Cano, P., Craig-Veit, C., Goodson, M.L., David Furlow, J., and Murk, A.J. (2011). Detection of thyroid hormone receptor disruptors by a novel stable *in vitro* reporter gene assay. Toxicology in Vitro *25*, 257–266. https://doi.org/10.1016/j.tiv.2010.08.013.

5. scikit-learn (2024). 3.2. Tuning the hyper-parameters of an estimator. scikit-learn. https://scikit-learn/stable/modules/grid_search.html.

6. Rogers, D., and Hahn, M. (2010). Extended-Connectivity Fingerprints. J. Chem. Inf. Model. *50*, 742–754. https://doi.org/10.1021/ci100050t.

7. Crammer, K., and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. J. Mach. Learn. Res. *2*, 265–292.

8. Smola, A.J., and Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing *14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

9. Nikolov, N.G., Nissen, A.C.V.E., and Wedebye, E.B. (2023). A method for *in vitro* data and structure curation to optimize for QSAR modelling of minimum absolute potency levels and a comparative use case. Environmental Toxicology and Pharmacology *98*, 104069. https://doi.org/10.1016/j.etap.2023.104069.

10. Rishton, G.M. (1997). Reactive compounds and *in vitro* false positives in HTS. Drug Discovery Today *2*, 382–384. https://doi.org/10.1016/S1359-6446(97)01083-0.

11. Hornung, M.W., Tapper, M.A., Denny, J.S., Sheedy, B.R., Erickson, R., Sulerud, T.J., Kolanczyk, R.C., and Schmieder, P.K. (2017). Avoiding False Positives and Optimizing Identification of True Negatives in Estrogen Receptor Binding and Agonist/Antagonist Assays. Appl In Vitro Toxicol *3*, 163–181. https://doi.org/10.1089/aivt.2016.0021.

12. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579–2605.

9.3 Supporting information

Training set(s):
Proprietary. All the queries must be directly addressed to KREATiS SAS.

External validation set(s):
Proprietary. All the queries must be directly addressed to KREATiS SAS.

Supporting information:
None.

10 Summary (KREATiS QMRF Database)

10.1 QMRF number (For KREATiS internal records only)

KTS/QMRF/ETR/01

10.2 Publication date

17 January 2025

10.3 Keywords

iSafeRat®; Endocrine modality; thyroid receptors; SVM.

10.4 Comments

None.