1	**QSAR identifier**	
1.1	QSAR identifier (title)	AHR - Aryl Hydrocarbon Receptor model (ONTOX) 1.0.0
1.2	Other related models	
1.3	Software coding the model	VEGA QSAR 1.2.5 (https://www.vegahub.eu/portfolio-item/vega-qsar/)
2	**General Information**	
2.0	Abstract	A quantitative structure–activity relationship (QSAR) classification model was developed from bioactivity data of ChEMBL 33 database to predict compound disruption toward AHR - Aryl Hydrocarbon Receptor - as a molecular initiating event (MIE) upstream of cognitive functional defects. Predictions from this QSAR models could be used alone or being integrated with predictions for other MIEs of the same adversity to provide hints about the presence of chemicals with potential adverse effects. QSARs for MIEs represent a relevant first-tier in prioritizing chemicals for further targeted testing.
2.1	Date of QMRF	24/09/2024
2.2	QMRF author(s) and contact details	Domenico Gadaleta et al., Laboratory of Environmental Chemistry and Toxicology, Istituto Di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
2.3	Date of QMRF update(s)	
2.4	QMRF update(s)	
2.5	Model developer(s) and contact details	Domenico Gadaleta, Laboratory of Environmental Chemistry and Toxicology, Istituto Di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy
2.6	Date of model development and/or publication	Published: November 2023; released: October 2025
2.7	Reference(s) to main scientific papers and/or software package	Gadaleta, D., Garcia de Lomana, M., Serrano-Candelas, E., Ortega-Vallbona, R., Gozalbes, R., Roncaglioni, A., & Benfenati, E. (2024). Quantitative structure–activity relationships of chemical bioactivity toward proteins associated with molecular initiating events of organ-specific toxicity. Journal of Cheminformatics, 16(1), 122.
2.8	Availability of information about the model	The model is non-proprietary: full description of the model algorithm is available; training and test sets can be downloaded from VEGA QSAR v1.2.5 5 (https://www.vegahub.eu/portfolio-item/vega-qsar/)
2.9	Availability of another QMRF for exactly the same model	
3	**Defining the endpoint - OECD Principle 1: "A DEFINED ENDPOINT"**	**PRINCIPLE 1: "A DEFINED ENDPOINT". ENDPOINT refers to any physicochemical, biological, or environmental property/activity/effect that can be measured and therefore modelled. The intent of PRINCIPLE 1 (a (Q)SAR should be associated with a defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system and test conditions that is being modelled by the Q)SAR.**
3.1	Species	Different assays performed on *Homo Sapiens* isoform of the protein target.
3.2	Endpoint	Binding to AHR - Aryl Hydrocarbon Receptor (CHEMBL ID: 3201; UniProt Accession: P35869).

		Data were extracted from ChEMBL v33. Only entries with pChEMBL values, representing measures of half-maximal responses (molar IC50, XC50, EC50, AC50, Ki, Kd, potency, and ED50) expressed on a negative logarithmic scale were chosen. Selected pChEMBL data were converted into a binary classification (active, not active) based on the information reported in the 'Standard Value', 'Standard Relation' and 'Comment' fields' in ChEMBL: • When both a 'Standard Relation' and a 'Standard Value' were available and 'Standard unit' was 'nM', records were classified as not active if the 'Standard Relation' was "=", ">" or "≥" and the 'Standard Value' was higher or equal than 10,000 nM. Conversely, records were classified as active if the 'Standard Relation' was "=" or "<"and the 'Standard Value' was lower than 10,000 nM. • When a 'Standard Value' was available but a 'Standard Relation' was missing, records with a 'Standard Value' higher or equal to 10,000 nM were flagged as not active, while those with 'Standard Value' lower than 10,000 nM were flagged as active. • If neither the 'Standard Value' nor the 'Standard Relation' were available, the 'Comment' field was further searched for keywords suggesting an activity classification (e.g., active/not active, inhibitor/not inhibitor). • After assigning activity labels, the 'Comment' field was further searched for keywords suggesting an activity classification (e.g., active/not active, inhibitor/not inhibitor, etc.). If any keyword was found suggesting a contradictory activity with respect to the one previously assigned, the record was discarded.
3.3	Comment on endpoint	
3.4	Endpoint units	Categorical (active / non active) based on a threshold of 10,000 nM on IC50, XC50, EC50, AC50, Ki, Kd, potency, and ED50 data from ChEMBL(see section 3.2).
3.5	Dependent variable	
3.6	Experimental protocol	The endpoint is derived from heterogeneous data types reported in ChEMBL (IC50, EC50, Ki, Kd, etc.) and obtained from different assays. Details on the assay used to extract single datapoints may be reconstructible by associating the compound IDs from the training and test set and the corresponding assay IDs available from the raw datasets attached (see section 9.3).
3.7	Endpoint data quality and variability	Records with potential duplicates, flagged with "Data Validity Comment" or classified as 'inconclusive,' 'undetermined,' or 'not determined' in the "Comment" field were discarded. ChEMBL data undergo manual curation and multiple quality checks prior to inclusion, ensuring a high standard of reliability. These checks include the validation of assay annotations, harmonization of chemical structures, standardization of activity types and units, and the removal of obvious errors or inconsistencies. As a result, ChEMBL is recognized as one of the most extensively used public resources for bioactivity data in cheminformatics and toxicology. More detailed information on curation and quality control is available in the official ChEMBL documentation (https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/chembl-data-questions) and in dedicated publications [1, 2]
4	**Defining the algorithm - OECD Principle 2: "AN UNAMBIGUOUS**	**PRINCIPLE 2: "AN UNAMBIGUOUS ALGORITHM". The (Q)SAR estimate of an endpoint is the result of applying an ALGORITHM to a set of structural parameters which describe the chemical structure. The**

	ALGORITHM"	**intent of PRINCIPLE 2 (a (Q)SAR should be associated with an unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. In this context, algorithm refers to any mathematical equation, decision rule or output approach.**
4.1	Type of model	Extreme Gradient Boosting
4.2	Explicit algorithm	Extreme gradient Boosting (XGBoost) Classifier as available in Python's sklearn 1.3.0. Model parameters were selected based on hyperparameter tuning and five-fold cross-validation performed on training data obtained from the first "master" split (see section 7.9). Balanced accuracy (BA) was used as the objective function. Bayesian Optimization was employed for parameter selection. The model parameters are gamma: 0; lambda: 1; alpha: 1, eta: 0.165, maximum tree depth: 8.
4.3	Descriptors in the model	A total of six descriptors is used in the model. The list of numerical indices identifying the descriptors is not provided as they have no explicit chemical or biological meaning.
4.4	Descriptor selection	The initial pool of 510 Continuous data-driven descriptors (CDDD) was pruned by descriptors with by low variance (st. dev <0.01) and by highly correlated descriptors (absolute pair correlation >0.90. For each pair of correlated descriptors, the one with the most correlated descriptors is kept and the other is removed. A reduced pool of 488 descriptor was obtained. Additional descriptor selection was made with the R package VSURF, based on a Random Forest (RF) algorithm that eliminates redundant or irrelevant variables [3]. The reduced set included six variables.
4.5	Algorithm and descriptor generation	Continuous data-driven descriptors (CDDD) are generated by deep neural networks, and are derived from the embedding learned by a model trained to translate semantically different molecular representations. Additional information on the algorithm for descriptor generation are available from https://github.com/jrwnter/cddd and from the reference publication [4].
4.6	Software name and version for descriptor generation	Descriptors were calculated using the Python code available at https://github.com/jrwnter/cddd
4.7	Chemicals/Descriptors ratio	Training set: 619. Descriptors: 6.
5	**Defining the applicability domain - OECD Principle 3: "A DEFINED DOMAIN OF APPLICABILITY"**	**PRINCIPLE 3: "A DEFINED DOMAIN OF APPLICABILITY". APPLICABILITY DOMAIN refers to the response and chemical structure space in which the model makes predictions with a given reliability. Ideally the applicability domain should express the structural, physicochemical and response space of the model. The CHEMICAL STRUCTURE (x variable) space can be expressed by information on physicochemical properties and/or structural fragments. The RESPONSE (y variable) can be any physicochemical, biological or environmental effect that is being predicted. According to PRINCIPLE 3 a (Q)SAR should be associated with a defined domain of applicability. Section 5 can be repeated (e.g., 5.a, 5.b, 5.c, etc) as many times as necessary if more than one method has been used to assess the applicability domain.**
5.1	Description of the applicability domain of	The applicability domain of the model implemented in VEGA v. 2.5.0 is assessed using an Applicability Domain

	the model	Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several indices, each one taking the calculation of the analogues found in the training and test set of the model, the similarity of analogues with respect of the predicted chemical and the accuracy of analogue predictions. For each index, including the final ADI, three fixed intervals are defined, with the first interval corresponding to a positive evaluation, the second to a suspicious evaluation and the last to a negative evaluation.
5.2	Method used to assess the applicability domain	The ADI is calculated by grouping the following indices: • Similar molecules with known experimental value. This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are: - 1 >= index > 0.9 strongly similar compounds with known experimental value in the training set have been found. - 0.9 >= index > 0.75 only moderately similar compounds with known experimental value in the training set have been found. - index <= 0.75 no similar compounds with known experimental value in the training set have been found. • Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are: - index < 0.5 accuracy of prediction for similar molecules found in the training set is good - 0.5 <= index < 1.0 accuracy of prediction for similar molecules found in the training set is not optimal. - index > 1.0 accuracy of prediction for similar molecules found in the training set is not adequate. • Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules). This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are: - index < 0.5 similar molecules found in the training set have experimental values that agree with the target compound predicted value. - 0.5 <= index < 1.0 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value. - index > 1.0 similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value. • Maximum error of prediction among similar molecules. This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds fall in an

		area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are: - index < 0.5 the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability. - 0.5 <= index < 1.0 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability. - index >= 1.0 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability. Atom-centered fragments similarity check: This index checks if atom centered fragments in the test chemical are represented in the training set. A value near 0 means that a prominent number of atom-centered fragments of the compound have not been found in the compounds of the training set or are rare fragments. Defined intervals are: - index < 0.5 a prominent number of atom-centered fragments of the compound have not been found in the compounds of the training set or are rare fragments. - 0.5 <= index < 1.0 few atom-centered fragments of the compound have not been found in the compounds of the training set - index >= 1.0 all the atom-centered fragments of the compound have been found in the compounds of the training set. • Global AD Index. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are: - 1 >= index > 0.85 predicted substance is into the Applicability Domain of the model. - 0.85 >= index > 0.75 predicted substance could be out of the Applicability Domain of the model. - index <= 0.75 predicted substance is out of the Applicability Domain of the model.
5.3	Software name and version for applicability domain assessment	VEGA QSAR 1.2.5 (https://www.vegahub.eu/portfolio-item/vega-qsar/)
5.4	Limits of applicability	Given the nature of the indexes used to determine the ADI, the model is likely to return reliable predictions only for chemicals that are structurally similar to the training samples. Chemicals for which CDDD descriptors cannot be calculated (e.g., inorganics and organometallics) will be not predicted by the model.
6	**Defining goodness-of-fit and robustness (internal validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"**	**PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. GOODNESS-OF-FIT and ROBUSTNESS refer to the internal model performance.**
6.1	Availability of the training set	The dataset is available from VEGA QSAR 1.2.5 and attached to the present QMRF (see section 9.3). Given that multiple training-test splitting iterations were made to evaluate model performance and the model was retrained on the entire dataset, the subdivision between training and test is not provided.

6.2	Available information for the training set	1) ID (internal identifier); 2) ChEMBL_ID (ChEMBL compound identifier); 3) SMILES (Kekulized SMILES by mean of RDKit Kekulizer in KNIME 4.7.5); 4) Category (binary endpoint); 5) CDDD descriptors with progressive numeric IDs
6.3	Data for each descriptor variable for the training set	Available for download from VEGA 1.2.5
6.4	Data for the dependent variable for the training set	Available for download from VEGA 1.2.5
6.5	Other information about the training set	Number of training chemicals: 556 (average across 100 splitting iterations) (see section 7.9).
6.6	Pre-processing of data before modelling	A semi-automated curation procedure [5] was applied to SMILES strings retrieved from ChEMBL to neutralize ionized chemical structures, remove counterions, and discard inorganics, organometallics, and mixtures. Duplicate structures were verified automatically at the InChI level. Duplicates with contradictory activity labels were removed.
6.7	Statistics for goodness-of-fit	
6.8	Robustness - Statistics obtained by leave-one-out cross-validation	
6.9	Robustness - Statistics obtained by leave-many-out cross-validation	
6.10	Robustness - Statistics obtained by Y-scrambling	
6.11	Robustness - Statistics obtained by bootstrap	
6.12	Robustness - Statistics obtained by other methods	
7	**Defining predictivity (external validation) – OECD Principle 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY"**	**PRINCIPLE 4: "APPROPRIATE MEASURES OF GOODNESS-OF-FIT, ROBUSTENESS AND PREDICTIVITY". PRINCIPLE 4 expresses the need to perform validation to establish the performance of the model. PREDICTIVITY refers to the external model validation. Section 7 can be repeated (e.g., 7.a, 7.b, 7.c, etc) as many times as necessary if more validation studies need to be reported in the QMRF.**
7.1	Availability of the external validation set	The dataset is available from VEGA QSAR 1.2.5 and attached to the present QMRF (see section 9.3). Given that multiple training-test splitting iterations were made to evaluate model performance and the model was retrained on the entire dataset, the subdivision between training and test is not provided.
7.2	Available information for the external validation set	1) ID (internal identifier); 2) ChEMBL_ID (ChEMBL compound identifier); 3) SMILES (Kekulized SMILES by mean of RDKit Kekulizer in KNIME 4.7.5); 4) Category (binary endpoint); 5) CDDD descriptors with progressive numeric IDs
7.3	Data for each descriptor variable for the external validation set	Available for download from VEGA 1.2.5

7.4	Data for the dependent variable for the external validation set	Available for download from VEGA 1.2.5
7.5	Other information about the external validation set	Number of test chemicals: 63 (average across 100 splits) (see section 7.9)
7.6	Experimental design of test set	Datasets were randomly partitioned into training and test sets (80:20 ratio) using stratified sampling to ensure a uniform distribution of active and not active samples between the two datasets for each endpoint.
7.7	Predictivity - Statistics obtained by external validation	Macro-averaged performance over 100 splits: TP: 36.12; FP: 7.06; TN: 15.94, FN: 3.88; SEN: 0.903; SPE: 0.693; BA: 0.798; MCC: 0.621; AUC: 0.845
7.8	Predictivity - Assessment of the external validation set	The distribution of active and not active samples between training and test chemicals was kept constant during the splitting based on a stratified sampling of activities.
7.9	Comments on the external validation of the model	A "master" splitting between training and test set was employed to determine optimal model parameters (based on balanced accuracy from five-fold cross validation). The optimal parameters identified from the master were kept constant, then the splitting procedure was repeated 100 times and statistics on the test sets were collected for each iteration and macro-averaged. In the end, the model was retrained on the entire dataset keeping the optimal parameters constant.
8	**Providing a mechanistic interpretation - OECD Principle 5: "A MECHANISTIC INTERPRETATION, IF POSSIBLE"**	
8.1	Mechanistic basis of the model	No mechanistic basis of the model is possible because descriptors are mathematical indices that cannot be interpreted from a chemical / biological perspective.
8.2	A priori or a posteriori mechanistic interpretation	
8.3	Other information about the mechanistic interpretation	
9	**Miscellaneous information**	
9.1	Comments	This model can be used together with other models available in VEGA QSAR 1.2.5 for other molecular initiating events upstream of cognitive functional defects aiming at prioritizing chemicals for their potential systemic effect, based on the assumption that chemicals activating multiple MIEs are more likely to extert toxicity [6].
9.2	Bibliography	(1) Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., ... & Leach, A. R. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic acids research, 52(D1), D1180-D1192. (2) Zdrazil, B. (2025). Fifteen years of ChEMBL and its role in cheminformatics and drug discovery. Journal of Cheminformatics, 17(1), 1-9. (3) Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31:2225‑2236 (4) Winter R, Montanari F, Noe F, Clevert D-A (2019) Learning continuous and data-driven molecular

		descriptors by translating equivalent chemical representations. Chem Sci 10:1692‑1701 (5) Gadaleta D, Lombardo A, Toma C, Benfenati E (2018) A new semiautomated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminform 10(1):1–13 (6) Leist M, Ghallab A, Graepel R, Marchan R, Hassan R, Bennekou SH, Limonciel A, Vinken M, Schildknecht S, Waldmann T, Danen E, van Ravenzwaay B, Kamp H, Gardner I, Godoy P, Bois FY, Braeuning A, Reif R, Oesch F, Drasdo D, Höhme S, Schwarz M, Hartung T, Braunbeck T, Beltman J, Vrieling H, Sanz F, Forsby A, Gadaleta D, Fisher C, Kelm J, Fluri D, Ecker G, Zdrazil B, Terron A, Jennings P, Burg BVD, Dooley S, Meijer AH, Willighagen E, Martens M, Evelo C, Mombelli E, Taboureau O, Mantovani A, Hardy B, Koch B, Escher S, van Thriel C, Cadenas C, Kroese D, Water BVD, Hengstler JG (2017) Adverse outcome pathways: opportunities, limitations and open questions. Arch Toxicol 91:3477–3505
9.3	Supporting information	