

Technical University of Denmark



QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories

Abildgaard Rosenberg, Sine; D. Watt, Eric; Judson, Richard S. ; Simmons, S. O.; Paul Friedmann, Katie; Dybdahl, Marianne; Nikolov, Nikolai Georgiev; Wedebye, Eva Bay

Published in:
Computational Toxicology

Link to article, DOI:
[10.1016/j.comtox.2017.07.006](https://doi.org/10.1016/j.comtox.2017.07.006)

Publication date:
2017

Document Version
Peer reviewed version

[Link back to DTU Orbit](#)

Citation (APA):
Abildgaard Rosenberg, S., D. Watt, E., Judson, R. S., Simmons, S. O., Paul Friedmann, K., Dybdahl, M., ... Wedebye, E. B. (2017). QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories. *Computational Toxicology*, 4, 11-21. DOI: 10.1016/j.comtox.2017.07.006

DTU Library

Technical Information Center of Denmark

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories

Rosenberg, S.A.^a, Watt, E.D.^{b,c}, Judson, R.S.^b, Simmons, S.O.^b, Paul Friedman, K.^b, Dybdahl, M.^{a1},
Nikolov, N.G.^{a1}, and Wedebye, E.B.^{a1*}

- a. *Division of Diet, Disease Prevention and Toxicology, National Food Institute, Technical University of Denmark, Kemitorvet, Building 202, 2800 Kgs. Lyngby, Denmark*
- b. *National Center for Computational Toxicology, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA*
- c. *Current Address: Computational ADME Group, Department of Pharmacokinetics, Dynamics, and Metabolism, Pfizer Worldwide Research & Development, Groton, CT 06340, USA*

¹*Contributed equally*

**Corresponding author: ebawe@food.dtu.dk, +45 35887604*

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the U.S. Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Abstract

Thyroperoxidase (TPO) is the enzyme that synthesizes thyroid hormones (THs). TPO inhibition by chemicals can result in decreased TH levels and developmental neurotoxicity, and therefore identification of TPO inhibition is of high relevance in safety evaluation of chemicals. In the present study, we developed two global quantitative structure-activity relationship (QSAR) models for TPO inhibition *in vitro*. Rigorous cross- and blinded external validations demonstrated that the first model, QSAR1, built from a training set of 877 chemicals, was robust and highly predictive with balanced accuracies of 80.6% (SD = 4.6%) and 85.3%, respectively. The external validation test set was subsequently merged with the training set to constitute a larger training set totaling 1,519 chemicals for a second model, QSAR2, which underwent robust cross-validation with a balanced accuracy of 82.7% (SD = 2.2%). An analysis of QSAR2 identified the ten most discriminating structural features for TPO inhibition and non-inhibition, respectively. Both models were used to screen 72,524 REACH substances and 32,197 U.S. EPA substances, and QSAR2 with the expanded training set had an approximately 10% larger coverages compared to QSAR1. Of the substances predicted within QSAR2's applicability domain, 8,790 (19.3%) REACH substances and 7,166 (19.0%) U.S. EPA substances, respectively, were predicted to be TPO inhibitors. A case study on butyl hydroxyanisole (BHA), which is extensively used as an antioxidant, was included to exemplify how predictions from the developed QSAR2 model may aid in elucidating the modes of action in adverse outcomes of chemicals. Overall, predictions from QSAR2 can for example be used in priority setting of chemicals and in read-across cases or weight-of-evidence assessments.

Keywords

QSAR; thyroperoxidase (TPO) inhibition; Adverse Outcome Pathway (AOP); Screening; BHA; REACH;

Abbreviations

AD, applicability domain; AOP, adverse outcome pathway; AUR, Amplex®UltraRed; BHA, butylated hydroxyanisole; DNT, developmental neurotoxicity; DTU Food, Technical University of Denmark National Food Institute; EPA, Environmental Protection Agency; HTS, high-throughput screening; IATA, integrated approaches to testing and assessment; KE, key event; LPDM, Leadscope® Predictive Data Miner; MIE, molecular initiating event; NCCT, National Center for Computational Toxicology; OECD, Organisation for Economic Co-operation and Development; PLR, partial logistic regression; PRS, pre-registered substances; QSAR, quantitative structure-activity relationship; SD, standard deviation; TH, thyroid hormone; TPO, thyroperoxidase; WoE, weight-of-evidence

ACCEPTED MANUSCRIPT

1. Introduction

Thyroid hormones (THs) participate in multiple biological processes from early development and throughout adulthood [1–3]. In the fetus and neonate, THs play an essential role in neurodevelopment [4], where they are involved in neuron differentiation, proliferation and migration, dendritic branching and synaptogenesis, and myelination [1,5]. In early gestation, the fetus depends entirely on maternally-derived THs until the fetal thyroid gland becomes functional at approximately gestational week 12 in humans and gestational day 17-18 in rats [1,6,7]. Maternal THs continue to contribute to fetal TH levels throughout gestation in both humans and rats [1,6]. Studies have shown that even a moderate and transient decrease in maternal TH levels during pregnancy is associated with permanent adverse neurological changes in the offspring [8]. In animal models and humans altered cognition, socialization and motor function as well as hearing loss have been observed following moderate to severe hypothyroidism [6,9–17]. Even low levels of TH insufficiency during fetal development may result in measurable IQ deficits in children [9,13–18]. In adulthood, dysregulated TH levels can give reversible clinical symptoms of hypo- or hyperthyroidism [8] and are correlated with pathological processes involved in adverse outcomes such as cancer, obesity and type II diabetes mellitus [19,20].

Humans are exposed to tens of thousands of man-made chemicals through food, drugs, air, water and consumer products [21–24]. Large data gaps exist for most of these xenobiotics on their potential thyroid disrupting properties [25]. Xenobiotics can disturb TH homeostasis through many different mechanisms, including altered TH synthesis, transport, metabolism, and thyroid hormone receptor activation as well as disruption of the hypothalamus-pituitary-thyroid axis [10,25–28]. The same xenobiotic may act through more than one mechanism [25]. Because of the severity of the adverse effects that can be expected from chemical disruption of thyroid homeostasis, especially during early development, there is a need to develop a strategy for the identification and testing of thyroid-active compounds. As a step towards replacing expensive and time-consuming whole animal studies with alternative methods in chemical risk assessments, the Organisation for Economic Co-

operation and Development (OECD) launched a new program on the development of Adverse Outcome Pathways (AOPs) in 2012 [29]. An AOP describes the sequential chain of causally linked events at different levels of biological organization starting from a so-called molecular initiating event (MIE) going through a number of downstream linked key events (KEs), and ends at an adverse health or ecotoxicological effect [29,30]. According to the OECD, AOPs are the central element of a toxicological knowledge framework to support chemical risk assessment based on mechanistic reasoning. AOPs can help industry and regulators use results from alternative methods, such as *in vitro* and *in silico* methods, in chemical risk assessments [31], e.g. by applying the AOP in OECDs Integrated Approaches to Testing Assessment (IATA) context [29,32,33]. Multiple thyroid-related AOPs have been suggested [34,35]. One AOP under development determined to have a strong overall weight-of-evidence (WoE) describes a series of linked events from the MIE, thyroperoxidase (TPO) inhibition, leading to hypothyroxinemia, and resulting in altered neurodevelopment and neurological dysfunction in the offspring [36, see also 4 and 25]. TPO is a heme-containing multifunction enzyme essential in TH synthesis [37,38]. Recently, a high-throughput screening (HTS) *in vitro* assay for TPO inhibition was developed by the U.S. Environmental Protection Agency (EPA) National Center for Computational Toxicology (NCCT) [39] and used to screen 1,126 ToxCast Phase I and II chemicals including structurally diverse environmental chemicals and failed drugs [34,40,41]. The assay is based on microsomes from rat thyroid tissue and requires the amount from approximately one rat to assess quantitative TPO inhibition of 1.5 chemicals [39]. An additional set of 771 ToxCast chemicals (known as the 'Endocrine 1000' or 'E1K' set) [41,42] was subsequently screened in the same HTS TPO inhibition assay (Simmons *et al.*, in prep).

The goal of the present study was to use the ToxCast data to develop *in silico* models, and apply the models to large inventories of man-made chemicals to predict their potential to inhibit TPO. For this purpose, we first used experimental TPO inhibition results for 1,126 ToxCast Phase I and II chemicals to prepare a training set of 877 chemicals, which was then used to train and cross-validate a global binary Quantitative Structure-Activity Relationship (QSAR) model. QSARs are mathematical models

that relate chemical structure descriptors with an experimental continuous (e.g. EC₅₀) or categorical (e.g. positive/negative) activity. Once established, these *in silico* models can be used as a non-testing approach to predict the activities of untested chemical structures (an introduction to QSAR can e.g. be found in [43] and [44]). The E1K dataset was used to prepare a test set of 646 chemicals, which was applied to externally validate the QSAR model. Next, the test set was merged with the training set to form a larger training set of 1,519 chemicals, which was subsequently used for training and cross-validating a second QSAR model. An analysis of the structural features in the second QSAR model was performed to identify the top features that discriminated TPO inhibitors from non-inhibitors. Both QSAR models were used to screen two large EU and U.S. chemical inventories containing man-made substances potentially present in e.g. the environment and consumer products for their possible TPO inhibition activity. The screened EU inventory consist of 72,524 REACH pre-registered substances (PRS) extracted from the online Danish (Q)SAR Database structure set [45,46]. Briefly, REACH pre-registration concerns existing substances that companies plan to register under REACH as so-called phase-in substances and the full PRS list contains a total of 145,299 unique substances/entries [47]. The U.S. inventory was originally curated by the U.S. EPA as a part of the CERAPP project [48] and contains 32,464 unique structures to which humans are potentially exposed. The structures were curated from sources such as the ACToR CPCat database [21], the DSSTox database [49], the Canadian Domestic Substances List, the Endocrine Disruption Screening Program set and EPI Suite training and test sets [41,42,48]. Predictions from these screenings will inform a tiered approach to prioritize possible thyroid modulating chemicals for further evaluation and could be used, together with relevant AOP(s), in IATA WoE assessments [29,33,50]. We also conducted a case study to highlight how the developed QSAR models for TPO inhibition can support hypotheses regarding the mode of action for chemical-induced adverse outcomes observed in *in vivo* studies.

2. Materials and Methods

2.1 Experimental Datasets

We used two datasets provided by U.S. EPA NCCT with chemical structure information and HTS screening results for TPO inhibition *in vitro* to train and validate two QSAR models. The chemicals screened contained diverse chemical structures including environmental and industrial chemicals, as well as some failed drugs [41]. The chemicals in both datasets were not selected specifically for this project or based on suspected TPO inhibition activity, and the original datasets include internal replicated samples. The experimental results consisted of data from the HTS Amplex®UltraRed-thyroperoxidase (AUR-TPO) *in vitro* assay [39], which had further undergone a selectivity filtering procedure to identify potentially false positive results due to non-specific activity decrease in the AUR-TPO assay [34]. Briefly, all chemical structures were initially screened at a single, high concentration (~87.5µM). The chemicals associated with 20% or greater decreases in maximal TPO activity were subsequently screened for possible concentration-response. The concentration-response data were processed as described previously using the ToxCast data pipeline whereby each chemical was assigned a 'hit-call' of 1 if active in AUR-TPO, or a 'hit-call' of 0 if inactive in AUR-TPO [51]. Actives in the AUR-TPO assay were further processed through a selectivity filtering algorithm, which integrates results from cytotoxicity and luciferase inhibition assays to identify possible non-specific positive results in the AUR-TPO assay [34]. The chemical structures, assays, data analysis and selectivity filtering procedure have been described in more details previously [34,39,40,51]. We classified the chemicals into three categories (Figure 1): 1) chemicals that had a <20% activity decrease in the single, high concentration screening or had been assigned a 'hit-call' of 0 in the concentration-response AUR-TPO screening were classified as inactive in this assay; 2) chemicals with a 'hit-call' of 1 in AUR-TPO and a selectivity score greater than 1 were classified as active for TPO inhibition; and 3) chemicals with a 'hit-call' of 1 in AUR-TPO but with a selectivity score of 1 or less were classified as inconclusive for TPO inhibition.

The first dataset provided to the QSAR model developers at the Technical University of Denmark National Food Institute (DTU Food) consisted of structure information and experimental results for 1,126 ToxCast Phase I and II chemicals [34,40,41], including replicates, and was used for preparing a training set referred to as training set 1 (Figure 1). The second E1K dataset of an additional 771 chemicals from ToxCast [41,42], initially containing only structural information, was used for preparing a test set of 646 chemicals for external validation of the selected QSAR model built from training set 1 (see 2.3) (Figure 1). After determining the external validation statistics, the experimental results of the test set structures were made available to the model developers at DTU Food. The test set and training set 1 were then merged to form a second, larger training set referred to as training set 2 (Figure 2).

2.2 Structure Preparation

All chemical structures in the two U.S. EPA NCCT provided datasets had previously undergone an extensive quality control and structure curation procedure as part of the ToxCast program [41,52]. The QSAR software applied in this study handles organic chemical structures with an unambiguous 2D structure. We apply an overall definition of structures acceptable for QSAR processing in all our in-house QSAR software [45,46], as structures:

- containing at least two C atoms
- containing only the atoms H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and/or I; and,
- that are not mixtures consisting of two or more organic components

The structures that did not fulfill these criteria were removed from the two datasets. Further processing of the structural information included stripping off ions and neutralization of the organic parent structures, i.e. all structures were used in their non-ionized form (Figure 1).

Next, identical QSAR-ready structures within the first dataset were identified and their assigned experimental results were compared. For identical structures with concordant activities, only one of

the structures was kept. If a group of identical structures had discrepant activities then the whole group was removed from the dataset. Next, structures with inconclusive experimental results, i.e. 'hit-call' of 1 in AUR-TPO and a selectivity score of 1 or less, were removed and the dataset now constituted training set 1 (Figure 1). The same duplicates removal procedure was performed by U.S. EPA NCCT scientists on the DTU Food experimentally-blinded E1K set, which then constituted the test set (Figure 1). Some of the QSAR-ready structures in the test set were identical to structures in training set 1 and were therefore excluded from the external validation. When the test set experimental results were made available to DTU Food, and training set 2 was prepared by merging the test set and training set 1 (Figure 2), the experimental results of the identified structural duplicates were compared. Again, if they had concordant experimental result only one of the structures was kept, while all the structures were removed in case of disagreement between the experimental results.

2.3 QSAR Modeling and Selection

We used the commercial software Leadscape® Predictive Data Miner (LPDM), a component of Leadscape® Enterprise Server version 3.2.4 [53], to build the QSAR models. Briefly, for each chemical structure in a training set LPDM automatically performs a systematic sub-structural analysis using a template library of more than 27,000 predefined structural features and calculates nine molecular descriptors (AlogP, Hydrogen Bond Acceptors and Donors, Lipinski Score, Molecular Weight, Parent Atom Number, Parent Molecular Weight, Polar Surface Area, Rotatable Bonds) [54]. The structural features and molecular descriptors are included in a default descriptor set. In addition, the user may call a functionality in LPDM to generate and add new training set-dependent structural features (scaffolds) to the descriptor set. The pre-defined structural features, added scaffolds and numeric molecular descriptors are included in an initial descriptor set. From the initial descriptor set, an automatic descriptor selection procedure in LPDM selects the top 30% descriptors according to Yates χ^2 -test for a binary response variable. For the current training set 1 and 2 with binary responses, predictive models were built using partial logistic regression (PLR) with further selection

of descriptors in an iterative procedure, and selection of the optimum number of PLR factors based on least predictive residual sum of squares. LPDM has the option of building composite models, a type of ensemble model [55], for training sets with an imbalanced distribution of actives and inactives. With this option a number of sub-models are created by specifying the desired ratio of actives to inactives per sub-model training set, so that each of the sub-models contains the smaller class and a sample of the bigger class. The positive prediction probability (see 2.4) for a query chemical from a composite model is defined as the average of the positive prediction probabilities of all sub-models having the test chemical in the applicability domain (AD) [56]. This is in contrast to single models where the entire training set is used to train one single model.

Multiple modeling approaches were applied in LPDM to build seven predictive models for TPO inhibition first using training set 1 (Figure 2):

- 1) single (i.e., non-composite)
- 2) single with scaffolds
- 3) single with scaffolds and a reduced set of structural features
- 4) composite
- 5) composite with scaffolds
- 6) composite with scaffolds and a reduced set of structural features
- 7) composite model combining model 3 and the sub-models from model 6

In 1 and 4, the descriptors were selected among the default descriptors, i.e. the molecular descriptors and the predefined structural features, and used to build a single model and a composite model, respectively. Next, scaffolds were generated in LPDM for the training set structures and added to the initial descriptor set, which subsequently was used for descriptor selection for models 2 and 5. In models 3 and 6, the scaffold-enriched descriptor set was reduced using a built-in function in LPDM (i.e., 'Remove most features – (removes less similar features)') that removed certain similar structural features before the descriptor selection. This step was employed to achieve a higher-quality set of fewer structural features, eliminate highly similar or redundant ones, and reduce the

risk of overfitting. In model 7, the single model 3 and the sub-models from composite model 6 were combined to constitute a new composite model with equal weight of all its sub-models.

All seven models underwent a ten times two-fold cross-validation by the LPDM algorithm. The algorithm transfers knowledge of the selected descriptor set from the parent model when building the cross-validation models, and we therefore do not use it for our measures of absolute predictive performance, but only to guide relative performance-based selection between the seven preliminary models. Among the seven predictive models built from training set 1, we selected the model with the highest performance from the LPDM cross-validation for further validation and screening studies (Figure 2). The selected model, called QSAR1, was then closed for further development.

2.4 Applicability Domain Definition

The definition of the AD applied in this project consists of two components: 1) the definition of a structural domain in LPDM, and 2) a DTU Food in-house class probability refinement on the output from LPDM:

1) For a test compound to be within LPDM's structural domain it was required that: all molecular descriptors used in the model could be calculated, it contained at least one structural feature used in the model, and it had at least 30% Tanimoto similarity with a training set compound [56]. The 30% Tanimoto similarity was a default cut-off in the LPDM software. For a test compound outside this structural domain no prediction call (active/inactive) was generated by LPDM. For test compounds within the LPDM structural domain, a positive prediction probability, p , between 0 and 1, was given together with the prediction call; actives having a $p \geq 0.5$ and inactives having a $p < 0.5$ [56].

2) To exclude less reliable predictions, i.e. those with a positive prediction probability close to the cutoff $p = 0.5$, we required $p \geq 0.7$ for active prediction calls and $p \leq 0.3$ for inactive prediction calls. Predictions within the LPDM structural domain but with an associated positive prediction probability in the interval 0.3 to 0.7 were thus defined as outside of the AD and excluded from the statistical analyses.

2.5 Validation of the Models

Next, the closed QSAR1 model underwent an external validation blinded to DTU Food using the test set to evaluate its predictive performance (Figure 2). U.S. EPA NCCT compared the DTU Food generated test set prediction calls within the AD (see 2.4) with the corresponding experimental results and calculated sensitivity, specificity, balanced accuracy and coverage. Sensitivity is the percentage of experimental actives correctly predicted, specificity is the percentage of the experimental inactives correctly predicted, and balanced accuracy is the average of the sensitivity and specificity [57]. The coverage is the proportion of test set compounds that had predictions within the model's AD.

The assigned experimental activities for the test set were then made available to DTU Food, who merged the test set with training set 1 to constitute the larger training set 2 (see 2.2). Training set 2 was used to build seven predictive models using the same modeling and LPDM cross-validation approaches described for training set 1 in section 2.3, and of these the best performing model was selected (Figure 2). The selected model, called QSAR2, was closed for further development.

As described above, the LPDM cross-validation algorithm was, due to the issue with transfer of knowledge to the cross-validation models, only used to guide the selection of the best performing model among the seven models built from training set 1 and 2, respectively. The two selected and closed models, QSAR1 and QSAR2, were each subsequently subjected to a DTU Food in-house five times two-fold stratified cross-validation procedure to further estimate their robustness and predictive performance (Figure 2). This was done by randomly removing 50% of the structures from the training set, preserving the ratio of actives and inactives. Then a cross-validation model was built on the reduced training set using the same modeling approach as the full, parent model, but without transferring any established information such as selected descriptors from the parent model. The cross-validation model was applied to predict the 50% of the training set that had been removed. Likewise, a cross-validation model was made using the removed 50% of the training set, and this

model was used to predict the remaining 50%. This procedure was performed five times resulting in ten cross-validation models. Sensitivity, specificity and balanced accuracy were calculated for the in-AD predictions for each of the ten cross-validation models, and the mean and standard deviation (SD) were computed to give overall statistical measures of the predictive performance and robustness of the parent model based on the full-training set. The coverage, i.e. the mean percentage of how many of the predicted substances that had predictions within the AD of the ten cross-validation models, was also calculated.

2.6 Structural Features in QSAR2

To identify structural features in QSAR2 related to TPO inhibition or non-inhibition, respectively, all features in the model were sorted in descending order by:

$$|0.5 - \bar{x}| \cdot n$$

where n is the number of training set 2 structures containing the given feature, and \bar{x} is the mean TPO inhibition experimental activity (1 for actives and 0 for inactives) of the n training set structures.

With this metric the QSAR2 structural features that discriminate well between the two classes, i.e. actives and inactives, and are contained in the largest number of training set 2 structures are given the highest ranking. Based on this sorting, the top ten structural features with an $\bar{x} \geq 0.8$, i.e. structural features associated with activity, and an $\bar{x} \leq 0.02$, i.e. structural features associated with inactivity, respectively, were identified (Figure 2). The cutoff of $\bar{x} \leq 0.02$ was chosen instead of 0.2, which would have been symmetric to the $\bar{x} \geq 0.8$ cutoff for activity associated structural features, due to the larger proportion of inactive structures in the training set.

2.7 Screening Large Chemical Inventories

The structures in the REACH-PRS inventory were originally curated from deliverable 3.4 of the OpenTox EU project and had previously been processed through the structure preparation steps described in section 2.2 [58]. The 72,524 QSAR-ready REACH-PRS structures included structural

duplicates, and the REACH-PRS set thus contained a total of 60,281 unique structures (Figure 2). The U.S. EPA inventory was also previously processed through the structure preparation steps described in section 2.2 and 32,197 unique QSAR-ready structures remained. Both the REACH-PRS set and the U.S. EPA set were screened through the QSAR1 and QSAR2 TPO inhibition models to identify substances with the potential to inhibit TPO. We applied both QSAR1 and QSAR2 to be able to assess the effect of adding the test set structures to training set 2 with regard to their coverages and the prevalence of predicted TPO inhibitors in the two models. While QSAR2 is likely to provide a better coverage of the inventories, the lack of an external validation of QSAR2 may for some purposes suggest that QSAR1 is a more appropriate tool.

The overlaps in substances as well as unique structures between U.S. EPA and REACH-PRS were identified (Figure 2). The proportion of the QSAR-predicted U.S. EPA and REACH-PRS substances within the AD of QSAR1 and QSAR2 and the activity distributions of the predictions were calculated.

3. Results and Discussion

This is to our knowledge the first study to develop global binary QSAR models for TPO inhibition and apply them to predict two large and structurally diverse chemical inventories containing man-made substances for their TPO inhibiting potential.

3.1 The Training and Test Sets

The number of QSAR-ready structures and the distribution of active and inactive experimental results in training set 1, the test set and training set 2 are summarized in Table 1. The numbers given in the table reflect the situation after removing structures that were either unsuited for QSAR processing in the applied software, structural duplicates or had inconclusive experimental results. In training set 1 this resulted in the removal of 72 structures due to structural QSAR criteria, i.e. structures not acceptable for QSAR processing, 21 due to structural duplicates (four of these due to conflicting experimental results), and 156 due to inconclusive experimental results; in total 249 out

of the 1,126 initial structure entries. In the external validation test set, a total of 125 out of the 771 initial E1K structure entries were removed; 14 due to structural QSAR criteria, 23 due to overlap with training set 1 structures, 14 due to internal structural duplicates (two of these due to conflicting experimental results), and 74 due to inconclusive experimental results. When merging training set 1 and the test set, which at this point was un-blinded to DTU Food, the experimental results of the 23 structures removed from the test set due to overlap with training set 1 structures were compared with their corresponding training set 1 experimental results. In four cases the experimental results disagreed, and these structures were therefore removed from the final training set 2 (Table 1).

The chemical structures in the provided datasets had undergone thorough quality control and curation [41,52]. In addition, since the datasets originated from the same source, i.e. U.S. EPA NCCT, and all chemicals had been screened in the same testing protocols and undergone the same data processing, this has likely contributed to decrease the experimental variability. The quality of the AUR-TPO assay has been assessed previously [34,39] and indicated excellent performance with robust Z-prime factor from 0.77 to 0.83, where values above 0.5 generally indicate excellent performance to distinguish between actives and inactives, and high intralaboratory repeatability with the robust coefficient of variance being 3–4%. The data in training set 1 and 2 and the test set were therefore assessed to be of high quality [34,39] and expected to be a good basis for QSAR model development. The AUR-TPO assay measures the fluorescence intensity from the commercial peroxidase substrate, Amplex®UltraRed (AUR), which is converted to Amplex UltroxRed by a peroxidase in the presence of hydrogen peroxide. A decrease in fluorescence intensity in response to a chemical is an indirect measure of TPO inhibition. The reaction chemistry and oxidation product of AUR is proprietary and the exact reaction(s) inhibited and its reversibility cannot be identified [34]. Therefore, the AUR-TPO assay read out has multiple potential confounders, including: non-specific enzyme inhibition; reactive, autofluorescent or fluorescence quenching chemicals; and other sources of interference with the peroxidase reaction [34,39]. When comparing results from the AUR-TPO assay with results from the lower throughput orthogonal guaiacol oxidation assay, the AUR-TPO

assay was previously found to have a sensitivity of 86% and a specificity of 39% [34]. Part of the high sensitivity of AUR-TPO could be due to a higher rate of false positive results from confounding non-specific activity decrease, a known problem with loss-of signal assays. Identification and removal of such potentially AUR-TPO false positive TPO inhibitors in the datasets was attempted by the application of the selectivity score filter [34] and the inconclusive category, i.e. AUR-TPO positives with a selectivity score less than 1, see section 2.1. However, not all mechanisms potentially causing non-specific activity decrease, e.g. fluorescence quenching, have been addressed in the selectivity score [34] and so the presence of false positive TPO inhibitors in the training and test sets cannot be excluded. Furthermore, the tiered screening approach in AUR-TPO with a cutoff of 20% activity decrease in the initial single, high-concentration screening [34] may have produced some false negatives as it cannot be excluded that a portion of the chemicals causing an activity decrease below the 20% cutoff would have been positive if screened for concentration-response. In addition to the potential confounding effects in the raw experimental outputs, the models applied for the 'hit-call' assignment and the selectivity score algorithm are also subject to some degree of uncertainty in their results.

3.2 QSAR Modeling and Selection

Table 2 shows the LPDM cross-validation results for the seven models built from training set 1 and 2, respectively. As mentioned above, the LPDM cross-validation was used to guide relative performance-based selection between the seven preliminary models. As can be seen in Table 2, the composite modeling approaches 4 to 7 outperformed the single models 1, 2 and 3 in the LPDM cross-validation with regard to the balanced accuracy (Table 2). This is most likely an effect of the imbalanced distribution of actives and inactives in both training sets with a ratio of approximately 1:6 (Table 1). The composite model feature in LPDM was implemented to handle such imbalanced training sets to include also a high proportion of the bigger class and thereby optimize the size of the AD [56].

In this work we employed a new approach where a single, unbalanced model (i.e., model 3) was added as a sub-model, together with the balanced sub-models from a composite model (i.e., model 6), to form a new composite model (i.e., model 7). This addition caused a significant reduction in the number of false positive predictions (FPs) produced in the LPDM cross-validation as compared to model 6 alone (see Table 2). For both training set 1 and 2 this resulted in a remarkable increase in the LPDM cross-validation specificity while causing only a smaller reduction in sensitivity (Table 2), and together this explains why model 7, in both cases, outperformed the other composite models 4, 5 and 6. To conclude, model 7 was the best performing among the seven models for both training set 1 and 2, and therefore selected for both training sets, and these models were named QSAR1 and QSAR2, respectively (Table 3).

3.3 Predictive Performance of the QSAR Models

The two selected and final models, QSAR1 and QSAR2, underwent a five times two-fold DTU Food in-house cross-validation procedure to evaluate their predictive performance and robustness. QSAR1 also underwent a DTU Food blinded external validation with the test set. The results from the validation studies are presented in Table 3 and demonstrate high predictive performance, i.e. balanced accuracies of 85.3% by external validation for QSAR1 and 82.7% by cross-validation for QSAR2, respectively.

Adding the test set to training set 1 to build QSAR2 served multiple purposes. One purpose was to explore how much the added test set would enlarge the AD of the model and thereby increase the coverages of the two large chemical screening inventories, U.S. EPA and REACH-PRS. The coverage of QSAR2 was roughly 6% larger in the cross-validation (Table 3) and 10% larger for both screening inventories (Table 5) than the respective coverages of QSAR1. A second purpose of adding the test set in QSAR2 was to explore the possible improvements in predictive performance. To do this, we first built the smaller QSAR1 model and performed both a rigorous five times two-fold cross-validation procedure and an external validation with the test set. As can be seen in Table 3 the

validation procedures show that QSAR1 has high predictive performance and is a robust model, i.e. a balanced accuracy of 85.3% in external validation and 80.6% with an SD of 4.6% in the cross-validation. A comparison of the statistical parameters from the two validation methods indicates that the rigorous cross-validation procedure applied does not overestimate the model's predictive performance, but rather, outputs conservative estimates. This conservative nature of the cross-validation is likely due to the rigorous procedure of removing 50% of the full training set to build the cross-validation models. Such a procedure is especially hard on the proportionally few actives in training set 1, i.e. 130 out of 877 (Table 1), which is also reflected in the relatively high SD of 10% in the sensitivity of the ten QSAR1 cross-validation models (Table 3) as well as its lower mean value (72.3%) compared to the sensitivity from the external validation (79.7%). The structures in the test set used for the DTU-blinded external validation of QSAR1 were not selected due to specific TPO inhibition concerns or to serve as a representative test set for QSAR1, but instead selected because they are included in the U.S. EPA regulatory ToxCast universe based on potential for exposure, and not because of prior concern about endocrine disruptive effects [41,42].

The procedure of performing both independent and robust cross-validation and a large, representative and prospective external validation is optimal when evaluating a model's predictive performance, but external validation has the disadvantage of withholding what may in many cases be valuable data from the model itself [43,59]. Adding all available data to a training set can, in addition to expanding the AD, also result in a model with a higher predictive performance, depending on the characteristics of the added data. The QSAR2 model could not undergo an external validation procedure due to lack of another external test set. Previous studies have shown that robust cross-validations give reliable estimates of a model's predictive performance (e.g. [59,60]). For QSAR1, the applied two-fold cross-validation procedure gave a conservative measure of performance in comparison to the external validation. Based on this, we anticipate that QSAR2 will have a similar or higher predictive performance if it underwent a robust external validation with a test set generated using the same protocol and data processing. As can be seen from Table 3, the

cross-validation sensitivity was slightly increased in QSAR2 (75.6%) compared to QSAR1 (72.3%) and the sensitivity SD was reduced from 10.1% to 5%. This is most likely the effect of an increase in actives from 130 in training set 1 to 230 in training set 2, which renders the 50% exclusion in the cross-validation procedure less influential on the sensitivity. As there were already many inactives in training set 1, the addition of more inactives to training set 2 did, as expected, not have the same high impact on the specificity, which went from 89.0% (SD 2.8%) in QSAR 1 to 89.8% (SD 1.5%) in QSAR2.

3.4 Top Structural Features in QSAR2

The ten most frequent and discriminating predictive structural features associated with actives and inactives, respectively, in QSAR2 are shown in Figure 3. Among the highest ranking structural features associated with activity were versions of phenols, anisole and aniline. The most frequent structural features associated with inactivity included ethers, esters, aryl halides and a tertiary amine. To our knowledge structural docking or pharmacophore studies for TPO have not been performed (Simmons *et al.*, in prep).

3.5 The Screening Results

We found a total of 27,444 substances present in both the U.S. EPA and the full REACH-PRS inventories. There were 19,279 unique structures in common in the two inventories (Table 4). To our knowledge this is the first study that has quantified the overlap between these two inventories, both with regard to overall substance and unique structure overlap. The high overlap between the U.S. EPA set and the REACH-PRS set was not surprising since both inventories represent collections of man-made, environmental chemicals in the U.S. and EU, respectively.

Both the U.S. EPA and REACH-PRS inventories were screened using QSAR1 and QSAR2 for TPO inhibition. In Table 5 the coverage of the two substance inventories, i.e. the proportion of the full set predicted within the AD of the model, and the number of active and inactive predictions are presented for each model. As expected, the coverage of QSAR2 was larger than QSAR1 of both

screening sets. The percentage of chemicals in the two inventories with active predictions in the AD of the two models ranged from 16.5% to 19.3% (Table 5). Although slightly higher this is not very different from the percentage of experimentally determined actives of 14.8% to 15.5% in the training and test sets (Table 1). As mentioned earlier, the chemicals in the experimental datasets were not selected on the basis of expected TPO inhibition effects. It is not known to what extent the slightly higher percentages of TPO inhibitors in the two predicted screening sets are due to false positive predictions or if they reflect a true TPO inhibitor prevalence. The validation studies showed that both QSAR1 and QSAR2 have specificities >10% higher than their respective sensitivities (Table 3), and therefore both models are expected to, in a balanced universe, make relatively more false negative than false positive predictions.

3.6 Butylated Hydroxyanisole as a Potential Thyroid Hormone Disruptor

We searched the two chemical inventories for possible examples of human-relevant chemicals with known indications for adverse neurodevelopmental outcomes. Included in both the U.S. EPA and the REACH-PRS set were the two isomers of butylated hydroxyanisole (BHA, CASN 25013-16-5), 2-*tert*-Butyl-4-hydroxyanisole (2-BHA, CASN 88-32-4) and 3-*tert*-Butyl-4-hydroxyanisole (3-BHA, CASN 121-00-6) (Figure 4). BHA is manufactured and/or imported to the EU in a total of 100-1,000 tonnes per year and is used as an antioxidant and preservative in e.g. food, food contact materials, cosmetics, and pharmaceuticals [61–63]. It is an anticipated human carcinogen [64] and is has been noted to have published evidence of developmental neurotoxicity (DNT) in mammals [65,66]. Both *in vitro* and *in vivo* published studies indicate that the BHA isomers have endocrine-modulating potential, with most evidence for estrogenic and androgenic effects [63,67–72]. Based on this, BHA is on both the EU list of potential endocrine disruptors [73,74] and on the SIN (Substitute It Now!) List [75,76]. However, more data is needed to fully elucidate BHA's potential as an endocrine disruptor and its mode of action(s) in DNT [63].

Both 2- and 3-BHA were predicted active for TPO inhibition by QSAR2, and 3-BHA was included in the QSAR2 training set as a TPO inhibitor. Studies in rats and pigs indicate that exposure to BHA (mixture of the two isomers) *in utero* can cause effects such as changed thyroxine serum levels, altered thyroid gland function and histology, and altered brain weight and behavior in the offspring [66,67,72]. TPO inhibition is as mentioned above identified to be the MIE in an AOP for thyroid-related neurodevelopmental adverse effects (under development) [41]. The three common top activity-associated structural features from QSAR2 in the two isomers were identified as described in section 2.6 and is shown in Figure 4. Two of the features, “Scaffold 297” and “benzene, 1-alkoxy-,4-hydroxy” were among the top ten structural features associated with activity in QSAR2 (Figure 3). “Scaffold 297” was present in eleven training set 2 structures of which nine were experimentally active for TPO inhibition. The “benzene, 1-alkoxy-,4-hydroxy” structural feature was present in five training set 2 structures that were all experimentally positive.

The QSAR2 training set including flags for the test set structures of QSAR1 as well as the full experimental background data set used to prepare the QSAR sets are available for download at the following link: http://qsar.food.dtu.dk/download/TPO_inhibition_QSAR_training_set.zip. Work is underway to make the training sets available from the U.S. EPA ToxCast website. Furthermore, predictions for around 640,000 structures in QSAR2, including the 72,524 REACH-PRS structures, will be made available from the online Danish (Q)SAR Database [46]. QSAR2 will also be made available for prediction of user-submitted structures in a coming free online Danish (Q)SAR Models sister-site to the Danish (Q)SAR database at the DTU homepage [46].

4. Conclusions

The present study reports the development, validation, and application of two global, binary composite QSAR models for TPO inhibition *in vitro*. The first model, QSAR1, showed high predictive performance in both cross-and external validation with balanced accuracies of 80.6% (SD = 4.6%) and 85.3%, respectively. QSAR2, the second model, enlarged with the external test set of QSAR1,

showed improved robustness and predictive performance in cross-validation compared to QSAR1, i.e. a balanced accuracy of 82.7% (SD = 2.2%), and this was largely driven by an increase in sensitivity from 72.3% (SD = 10.1%) of QSAR1 to 75.6% (SD 5.0%) of QSAR2. The top-10 structural features in QSAR2 related to TPO enzyme inhibition and non-inhibition, respectively, were identified. The two QSAR models were used to screen two large chemical inventories from the U.S. and EU containing structurally diverse man-made chemicals to which humans are potentially exposed. QSAR2 showed an increase in coverage of ~10% for both inventories relative to QSAR1, and of the substances predicted within QSAR2's AD, 8,790 (19.3%) REACH-PRS substances and 7,166 (19.0%) U.S. EPA substances, respectively, were predicted to be TPO inhibitors. Among the predicted TPO inhibitors were the two isomers of BHA, which have previously been shown to cause both TH and neurological effects in animal studies. These QSAR predictions may contribute to elucidating the mode of action by which BHA results in these altered TH levels and neurological adverse outcomes. Overall, predictions from the two models can be used to prioritize chemicals for further testing and in considerations of possible concerns for downstream adverse outcomes (e.g., DNT) [77,78]. They may also be used e.g. in read-across cases or in IATA WoE assessments.

Conflict of Interest Statement

The authors declare that they have no conflict of interest in relation with this paper.

Acknowledgements

We would like to thank the Danish 3R Center and the Danish Environmental Protection Agency for supporting the project.

References

- [1] G.R. Williams, Neurodevelopmental and Neurophysiological Actions of Thyroid Hormone, *J. Neuroendocrinol.* 20 (2008) 784–794. doi:10.1111/j.1365-2826.2008.01733.x.
- [2] P.M. Yen, Physiological and molecular basis of thyroid hormone action., *Physiol. Rev.* 81 (2001) 1097–1142. <http://www.ncbi.nlm.nih.gov/pubmed/11427693>.

- [3] R.T. Zoeller, S.W. Tan, R.W. Tyl, General Background on the Hypothalamic-Pituitary-Thyroid (HPT) Axis, *Crit. Rev. Toxicol.* 37 (2007) 11–53. doi:10.1080/10408440601123446.
- [4] R.T. Zoeller, K.M. Crofton, Mode of Action: Developmental Thyroid Hormone Insufficiency—Neurological Abnormalities Resulting From Exposure to Propylthiouracil, *Crit. Rev. Toxicol.* 35 (2005) 771–781. doi:10.1080/10408440591007313.
- [5] E. Cuevas, E. Ausó, M. Telefont, G.M. de Escobar, C. Sotelo, P. Berbel, Transient maternal hypothyroxinemia at onset of corticogenesis alters tangential migration of medial ganglionic eminence-derived neurons, *Eur. J. Neurosci.* 22 (2005) 541–551. doi:10.1111/j.1460-9568.2005.04243.x.
- [6] K.L. Howdeshell, A Model of the Development of the Brain as a Construct of the Thyroid System, *Environ. Health Perspect.* 110 (2002) 337–348. doi:10.1289/ehp.02110s3337.
- [7] J. Kratzsch, F. Pulzer, Thyroid gland development and defects, *Best Pract. Res. Clin. Endocrinol. Metab.* 22 (2008) 57–75. doi:10.1016/j.beem.2007.08.006.
- [8] M.D. Miller, K.M. Crofton, D.C. Rice, R.T. Zoeller, Thyroid-Disrupting Chemicals: Interpreting Upstream Biomarkers of Adverse Outcomes, *Environ. Health Perspect.* 117 (2009) 1033–1041. doi:10.1289/ehp.0800247.
- [9] P. Berbel, J.L. Mestre, A. Santamaría, I. Palazón, A. Franco, M. Graells, A. González-Torga, G.M. de Escobar, Delayed Neurobehavioral Development in Children Born to Pregnant Women with Mild Hypothyroxinemia During the First Month of Gestation: The Importance of Early Iodine Supplementation, *Thyroid.* 19 (2009) 511–519. doi:10.1089/thy.2008.0341.
- [10] K.M. Crofton, Developmental Disruption of Thyroid Hormone: Correlations with Hearing Dysfunction in Rats, *Risk Anal.* 24 (2004) 1665–1671. doi:10.1111/j.0272-4332.2004.00557.x.
- [11] E.S. Goldey, L.S. Kehn, G.L. Rehnberg, K.M. Crofton, Effects of Developmental Hypothyroidism on Auditory and Motor Function in the Rat, *Toxicol. Appl. Pharmacol.* 135 (1995) 67–76. doi:10.1006/taap.1995.1209.
- [12] L. Kooistra, S. Crawford, A.L. van Baar, E.P. Brouwers, V.J. Pop, Neonatal Effects of Maternal Hypothyroxinemia During Early Pregnancy, *Pediatrics.* 117 (2006) 161–167. doi:10.1542/peds.2005-0227.
- [13] Y. Li, Z. Shan, W. Teng, X. Yu, Y. Li, C. Fan, X. Teng, R. Guo, H. Wang, J. Li, Y. Chen, W. Wang, M. Chawinga, L. Zhang, L. Yang, Y. Zhao, T. Hua, Abnormalities of maternal thyroid function during pregnancy affect neuropsychological development of their children at 25-30 months, *Clin. Endocrinol. (Oxf).* 72 (2010) 825–829. doi:10.1111/j.1365-2265.2009.03743.x.
- [14] G. Morreale de Escobar, M. Jesús Obregón, F. Escobar del Rey, Is Neuropsychological Development Related to Maternal Hypothyroidism or to Maternal Hypothyroxinemia? 1, *J. Clin. Endocrinol. Metab.* 85 (2000) 3975–3987. doi:10.1210/jcem.85.11.6961.
- [15] V.J. Pop, E.P. Brouwers, H.L. Vader, T. Vulsma, A.L. van Baar, J.J. de Vijlder, Maternal hypothyroxinaemia during early pregnancy and subsequent child development: a 3-year follow-up study, *Clin. Endocrinol. (Oxf).* 59 (2003) 282–288. doi:10.1046/j.1365-2265.2003.01822.x.
- [16] V.J. Pop, J.L. Kuijpers, A.L. van Baar, G. Verkerk, M.M. van Son, J.J. de Vijlder, T. Vulsma, W.M. Wiersinga, H.A. Drexhage, H.L. Vader, Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy, *Clin.*

- Endocrinol. (Oxf). 50 (1999) 149–155. doi:10.1046/j.1365-2265.1999.00639.x.
- [17] R.T. Zoeller, J. Rovet, Timing of Thyroid Hormone Action in the Developing Brain: Clinical Observations and Experimental Findings, *J. Neuroendocrinol.* 16 (2004) 809–818. doi:10.1111/j.1365-2826.2004.01243.x.
- [18] J.E. Haddow, G.E. Palomaki, W.C. Allan, J.R. Williams, G.J. Knight, J. Gagnon, C.E. O’Heir, M.L. Mitchell, R.J. Hermos, S.E. Waisbren, J.D. Faix, R.Z. Klein, Maternal Thyroid Deficiency during Pregnancy and Subsequent Neuropsychological Development of the Child, *N. Engl. J. Med.* 341 (1999) 549–555. doi:10.1056/NEJM199908193410801.
- [19] E.N. Pearce, Thyroid hormone and obesity, *Curr. Opin. Endocrinol. Diabetes Obes.* 19 (2012) 408–413. doi:10.1097/MED.0b013e328355cd6c.
- [20] C. Wang, The Relationship between Type 2 Diabetes Mellitus and Related Thyroid Diseases, *J. Diabetes Res.* 2013 (2013) 1–9. doi:10.1155/2013/390534.
- [21] K.L. Dionisio, A.M. Frame, M.-R. Goldsmith, J.F. Wambaugh, A. Liddell, T. Cathey, D. Smith, J. Vail, A.S. Ernstoff, P. Fantke, O. Jolliet, R.S. Judson, Exploring consumer exposure pathways and patterns of use for chemicals in the environment, *Toxicol. Reports.* 2 (2015) 228–237. doi:10.1016/j.toxrep.2014.12.009.
- [22] P.P. Egeghy, R. Judson, S. Gangwal, S. Mosher, D. Smith, J. Vail, E.A. Cohen Hubal, The exposure data landscape for manufactured chemicals, *Sci. Total Environ.* 414 (2012) 159–166. doi:10.1016/j.scitotenv.2011.10.046.
- [23] R. Judson, A. Richard, D.J. Dix, K. Houck, M. Martin, R. Kavlock, V. Dellarco, T. Henry, T. Holderman, P. Sayre, S. Tan, T. Carpenter, E. Smith, The Toxicity Data Landscape for Environmental Chemicals, *Environ. Health Perspect.* 117 (2009) 685–695. doi:10.1289/ehp.0800168.
- [24] M.-R. Goldsmith, C.M. Grulke, R.D. Brooks, T.R. Transue, Y.M. Tan, A. Frame, P.P. Egeghy, R. Edwards, D.T. Chang, R. Tornero-Velez, K. Isaacs, A. Wang, J. Johnson, K. Holm, M. Reich, J. Mitchell, D.A. Vallero, L. Phillips, M. Phillips, J.F. Wambaugh, R.S. Judson, T.J. Buckley, C.C. Dary, Development of a consumer product ingredient database for chemical exposure screening and prioritization, *Food Chem. Toxicol.* 65 (2014) 269–279. doi:10.1016/j.fct.2013.12.029.
- [25] A.J. Murk, E. Rijntjes, B.J. Blaauboer, R. Clewell, K.M. Crofton, M.M.L. Dingemans, J. David Furlow, R. Kavlock, J. Köhrle, R. Opitz, T. Traas, T.J. Visser, M. Xia, A.C. Gutleb, Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals, *Toxicol. Vitro.* 27 (2013) 1320–1346. doi:10.1016/j.tiv.2013.02.012.
- [26] K.M. Crofton, E.S. Craft, J.M. Hedge, C. Gennings, J.E. Simmons, R.A. Carchman, W.H. Carter Jr., M.J. DeVito, Thyroid-Hormone-Disrupting Chemicals: Evidence for Dose-Dependent Additivity or Synergism, *Environ. Health Perspect.* 113 (2005) 1549–1554. doi:10.1289/ehp.8195.
- [27] R.L. Divi, D.R. Doerge, Mechanism-Based Inactivation of Lactoperoxidase and Thyroid Peroxidase by Resorcinol Derivatives, *Biochemistry.* 33 (1994) 9668–9674. doi:10.1021/bi00198a036.
- [28] M. V. Kirthana, F. Nawaz Khan, P.M. Sivakumar, M. Doble, P. Manivel, K. Prabakaran, V. Krishnakumar, Antithyroid agents and QSAR studies: inhibition of lactoperoxidase-catalyzed iodination reaction by isochromene-1-thiones, *Med. Chem. Res.* 22 (2013) 4810–4817.

- doi:10.1007/s00044-013-0475-x.
- [29] OECD, PROPOSAL FOR A TEMPLATE, AND GUIDANCE ON DEVELOPING AND ASSESSING THE COMPLETENESS OF ADVERSE OUTCOME PATHWAYS, 2012.
<http://www.oecd.org/chemicalsafety/testing/49963554.pdf> (accessed January 13, 2017).
- [30] AOP-Wiki, The AOP-Wiki homepage, (2017). <https://aopwiki.org/> (accessed March 13, 2017).
- [31] N.C. Kleinstreuer, K. Sullivan, D. Allen, S. Edwards, D.L. Mendrick, M. Embry, J. Matheson, J.C. Rowlands, S. Munn, E. Maull, W. Casey, Adverse Outcome Pathways: From Research to Regulation Scientific Workshop Report, *Regul. Toxicol. Pharmacol.* 76 (2016) 39–50. doi:10.1016/j.yrtph.2016.01.007.
- [32] OECD, Workshop on Integrated Approaches to Testing and Assessment, 2008.
[http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2008\)10&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2008)10&doclanguage=en) (accessed January 13, 2017).
- [33] K.E. Tollefsen, S. Scholz, M.T. Cronin, S.W. Edwards, J. de Knecht, K. Crofton, N. Garcia-Reyero, T. Hartung, A. Worth, G. Patlewicz, Applying Adverse Outcome Pathways (AOPs) to support Integrated Approaches to Testing and Assessment (IATA), *Regul. Toxicol. Pharmacol.* 70 (2014) 629–640. doi:10.1016/j.yrtph.2014.09.009.
- [34] K. Paul Friedman, E.D. Watt, M.W. Hornung, J.M. Hedge, R.S. Judson, K.M. Crofton, K.A. Houck, S.O. Simmons, Tiered High-Throughput Screening Approach to Identify Thyroperoxidase Inhibitors Within the ToxCast Phase I and II Chemical Libraries, *Toxicol. Sci.* 151 (2016) 160–180. doi:10.1093/toxsci/kfw034.
- [35] AOPs, AOPs in AOP-Wiki as of March 2017, (2017). <https://aopwiki.org/aops> (accessed March 13, 2017).
- [36] AOP-42, Inhibition of Thyroperoxidase and Subsequent Adverse Neurodevelopmental Outcomes in Mammals, (2017). <https://aopwiki.org/aops/42> (accessed March 13, 2017).
- [37] R.S. Fortunato, E.C. Lima de Souza, R.A. Hassani, M. Boufraqueh, U. Weyemi, M. Talbot, O. Lagente-Chevallier, D.P. de Carvalho, J.-M. Bidart, M. Schlumberger, C. Dupuy, Functional Consequences of Dual Oxidase-Thyroperoxidase Interaction at the Plasma Membrane, *J. Clin. Endocrinol. Metab.* 95 (2010) 5403–5411. doi:10.1210/jc.2010-1085.
- [38] J. Ruf, P. Carayon, Structural and functional aspects of thyroid peroxidase, *Arch. Biochem. Biophys.* 445 (2006) 269–277. doi:10.1016/j.abb.2005.06.023.
- [39] K.B. Paul, J.M. Hedge, D.M. Rotroff, M.W. Hornung, K.M. Crofton, S.O. Simmons, Development of a Thyroperoxidase Inhibition Assay for High-Throughput Screening, *Chem. Res. Toxicol.* 27 (2014) 387–399. doi:10.1021/tx400310w.
- [40] D.J. Dix, K.A. Houck, M.T. Martin, A.M. Richard, R.W. Setzer, R.J. Kavlock, The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals, *Toxicol. Sci.* 95 (2007) 5–12. doi:10.1093/toxsci/kfl103.
- [41] A.M. Richard, R.S. Judson, K.A. Houck, C.M. Grulke, P. Volarath, I. Thillainadarajah, C. Yang, J. Rathman, M.T. Martin, J.F. Wambaugh, T.B. Knudsen, J. Kancherla, K. Mansouri, G. Patlewicz, A.J. Williams, S.B. Little, K.M. Crofton, R.S. Thomas, ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology, *Chem. Res. Toxicol.* 29 (2016) 1225–1251. doi:10.1021/acs.chemrestox.6b00135.

- [42] EDSP21 Work Plan, The Incorporation of In Silico Models and In Vitro High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening, (2011). https://www.epa.gov/sites/production/files/2015-07/documents/edsp21_work_plan_summary_overview_final.pdf (accessed March 13, 2017).
- [43] ECHA, Guidance on information requirements and chemical safety assessment - Chapter R.6: QSARs and grouping of chemicals, (2008). https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf (accessed March 16, 2017).
- [44] OECD, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, 2 (2007) 1–154. [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono\(2007\)2&doclanguage=en](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en) (accessed December 8, 2016).
- [45] QSAR, User Manual for the Danish (Q)SAR Database, (2015). http://qsar.db.food.dtu.dk/Danish_QSAR_Database_Draft_User_manual.pdf (accessed March 28, 2017).
- [46] QSARDB, Danish (Q)SAR Database, (2015). <http://qsar.food.dtu.dk/> (accessed March 14, 2017).
- [47] ECHA, Pre-registered substances, 2016. (n.d.). <https://echa.europa.eu/information-on-chemicals/pre-registered-substances> (accessed January 13, 2017).
- [48] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I. V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ. Health Perspect.* 124 (2016) 1023–1033. doi:10.1289/ehp.1510267.
- [49] A.M. Richard, C.R. Williams, Distributed structure-searchable toxicity (DSSTox) public database network: a proposal, *Mutat. Res. Mol. Mech. Mutagen.* 499 (2002) 27–52. doi:10.1016/S0027-5107(01)00289-5.
- [50] Z.A. Collier, K.A. Gust, B. Gonzalez-Morales, P. Gong, M.S. Wilbanks, I. Linkov, E.J. Perkins, A weight of evidence assessment approach for adverse outcome pathways, *Regul. Toxicol. Pharmacol.* 75 (2016) 46–57. doi:10.1016/j.yrtph.2015.12.014.
- [51] D.L. Filer, P. Kothiya, W.R. Setzer, R.S. Judson, M.T. Martin, The ToxCast™ Analysis Pipeline: An R Package for Processing and Modeling Chemical Screening Data, 2015. https://www.epa.gov/sites/production/files/2015-08/documents/pipeline_overview.pdf (accessed January 11, 2017).
- [52] U.S. EPA, ToxCast Chemical Inventory: Data Management and Data Quality Considerations, 2014. https://www.epa.gov/sites/production/files/2015-08/documents/toxcast_chemicals_qa_qc_management_141204.pdf (accessed January 13, 2017).
- [53] Leadscope, Leadscope, Inc, (2016). <http://www.leadscope.com/> (accessed March 23, 2017).
- [54] G. Roberts, G.J. Myatt, W.P. Johnson, K.P. Cross, P.E. Blower, LeadScope † : Software for

- Exploring Large Sets of Screening Data, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1302–1314. doi:10.1021/ci0000631.
- [55] L. Breiman, Bagging Predictors, *Mach. Learn.* 24 (1996) 123–140. doi:10.1023/A:1018054314350.
- [56] L.G. Valerio, C. Yang, K.B. Arvidson, N.L. Kruhlak, A structural feature-based computational approach for toxicology predictions, *Expert Opin. Drug Metab. Toxicol.* 6 (2010) 505–518. doi:10.1517/17425250903499286.
- [57] J.A. Cooper II, R. Saracci, P. Cole, Describing the validity of carcinogen screening tests, *Br. J. Cancer.* 39 (1979) 87–89.
- [58] OpenTox, Final database with additional content, (2011). <http://opentox.org/data/documents/development/opentoxreports/opentoxreportd34/view> (accessed October 14, 2016).
- [59] M. Gütlein, C. Helma, A. Karwath, S. Kramer, A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR, *Mol. Inform.* 32 (2013) 516–528. doi:10.1002/minf.201200134.
- [60] S.A. Rosenberg, M. Xia, R. Huang, N.G. Nikolov, E.B. Wedebye, M. Dybdahl, QSAR development and profiling of 72,524 REACH substances for PXR activation and CYP3A4 induction, *Comput. Toxicol.* (2017). doi:10.1016/j.comtox.2017.01.001.
- [61] ECHA, Substance information: tert-butyl-4-methoxyphenol, (2016). <https://echa.europa.eu/da/substance-information/-/substanceinfo/100.042.315>.
- [62] EFSA, Scientific opinion on the re-evaluation of butylated hydroxyanisole – BHA (E 320) as a food additive, 2011. doi:10.2903/j.efsa.2011.2392.
- [63] A. Pop, B. Kiss, F. Loghin, Endocrine disrupting effects of butylated hydroxyanisole (BHA - E320)., *Clujul Med.* 86 (2013) 16–20. <http://www.ncbi.nlm.nih.gov/pubmed/26527908> (accessed December 12, 2016).
- [64] NTP, Butylated Hydroxyanisole, 2016. <https://ntp.niehs.nih.gov/pubhealth/roc/index-1.html>.
- [65] W. Mundy, S. Padilla, M. Gilbert, J. Breier, J. Cowden, K. Crofton, D. Herr, K. Jensen, K. Raffaele, N. Radio, K. Schumacher, Building a Database of Developmental Neurotoxicants: Evidence from Human and Animal Studies, *Toxicol.* 108. (2009). http://www.fluoridealert.org/wp-content/uploads/epa_mundy.pdf (accessed December 12, 2016).
- [66] C. V. Vorhees, R.E. Butcher, R.L. Brunner, V. Wootten, Developmental Neurobehavioral Toxicity of Butylated Hydroxyanisole (BHA) in Rats, *Neurobehav. Toxicol. Teratol.* 3 (1981) 321–329.
- [67] S.-H. Jeong, B.-Y. Kim, H.-G. Kang, H.-O. Ku, J.-H. Cho, Effects of butylated hydroxyanisole on the development and functions of reproductive system in rats, *Toxicology.* 208 (2005) 49–62. doi:10.1016/j.tox.2004.11.014.
- [68] S. Jobling, T. Reynolds, R. White, M.G. Parker, J.P. Sumpter, A variety of environmentally persistent chemicals, including some phthalate plasticizers, are weakly estrogenic, *Environ. Health Perspect.* 103 (1995) 582–587. doi:10.1289/ehp.95103582.

- [69] H.G. Kang, S.H. Jeong, J.H. Cho, D.G. Kim, J.M. Park, M.H. Cho, Evaluation of estrogenic and androgenic activity of butylated hydroxyanisole in immature female and castrated rats, *Toxicology*. 213 (2005) 147–156. doi:10.1016/j.tox.2005.05.027.
- [70] A.M. Soto, C. Sonnenschein, K.L. Chung, M.F. Fernandez, N. Olea, F.O. Serrano, The E-SCREEN Assay as a Tool to Identify Estrogens: An Update on Estrogenic Environmental Pollutants, *Environ. Health Perspect.* 103 (1995) 113–122. doi:10.1289/ehp.95103s7113.
- [71] M.G.R. ter Veld, B. Schouten, J. Louisse, D.S. van Es, P.T. van der Saag, I.M.C.M. Rietjens, A.J. Murk, Estrogenic Potency of Food-Packaging-Associated Plasticizers and Antioxidants As Detected in ER α and ER β Reporter Gene Cell Lines, *J. Agric. Food Chem.* 54 (2006) 4407–4416. doi:10.1021/jf052864f.
- [72] G. Würtzen, P. Olsen, BHA study in pigs, *Food Chem. Toxicol.* 24 (1986) 1229–1233. doi:10.1016/0278-6915(86)90311-X.
- [73] DK-EPA, List of Undesiable Substances 2009, 2009. <http://www2.mst.dk/udgiv/publications/2011/05/978-87-92708-95-3.pdf>.
- [74] DK-EPA, The EU list of potential endocrine disruptors, (2016). <http://eng.mst.dk/topics/chemicals/endocrine-disruptors/the-eu-list-of-potential-endocrine-disruptors/> (accessed December 6, 2016).
- [75] U. Hass, S. Christiansen, M. Axelstad, J. Boberg, A. Andersson, N.E. Skakkebæk, K. Bay, H. Holbech, K.L. Kinnberg, P. Bjerregaard, Evaluation of 22 SIN List 2 . 0 substances according to the Danish proposal on criteria for endocrine disruptors, (2012) 1–141. http://eng.mst.dk/media/mst/67169/SIN_report_and_Annex.pdf (accessed December 13, 2016).
- [76] SIN, SIN List result for CAS number 25013-16-6, (2016). <http://sinlist.chemsec.org/search/search?query=25013-16-5> (accessed December 6, 2016).
- [77] EC, Commission Regulation (EU) 2015/282 of 20 February 2015 amending Annexes VIII, IX and X to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), (2015).
- [78] EFSA, OECD/EFSA Workshop on Developmental Neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes, (2016). <https://www.efsa.europa.eu/en/events/event/161018b> (accessed February 9, 2017).

Tables and Figures

Table 1. Number of structures in the QSAR-ready training sets 1 and 2, and test set with the distribution of active and inactive experimental results for TPO inhibition.

Datasets	Total number of unique structures	Active (%)	Inactive (%)
Training set 1	877	130 (14.8)	747 (85.2)
Test set*	646	100 (15.5)	546 (84.5)
Training set 2**	1519	230 (15.1)	1289 (84.9)

*The experimental results of the test set were masked to DTU Food model developers until after being predicted in QSAR1.

** some of the training set 1 structures were tested again together with the test set structures, and of these four structures had different activities compared to the training set 1 activity. The three training set 1 structures were removed from training set 2.

Table 2. The results from the LPDM cross-validation of the seven built models from training set 1 and 2, respectively.

Model	LPDMs 10 times two-fold cross-validation results						
	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	TP*	FP*	TN*	FN*
Training set 1							
1	43.0	96.8	69.9	49	21	626	65
2	48.2	96.0	72.1	55	26	621	59
3	50.0	96.3	73.2	57	24	623	57
4	72.9	82.7	77.8	94	105	502	35
5	81.4	78.2	79.8	105	136	498	24
6	84.5	80.3	82.4	109	123	502	20
7	74.6	92.5	83.6	97	55	676	33
Training set 2							
1	46.5	96.9	71.2	99	40	1153	114
2	49.8	96.1	73.0	106	46	1147	107
3	46.5	96.7	71.6	99	39	1154	114
4	79.1	79.9	79.5	182	233	928	48
5	75.7	79.5	77.6	174	240	931	56
6	76.1	78.4	77.3	175	253	918	55
7	71.3	92.6	82.0	164	95	1187	66

*TP: true positives, FP: false positives, TN: true negatives, FN: false negatives. The numbers are averages of the ten iterations as given by LPDM.

Table 3. Modeling approach applied and the predictive performances for QSAR1 and QSAR2.

Model	Statistical Parameter	Cross-Validation*, % (SD, %)	External Validation**, % (actual numbers)
QSAR1 Approach 7 Sub-models: 7	Sensitivity	72.3 (10.1)	79.7 (47/(47 + 12))
	Specificity	89.0 (2.8)	90.8 (266/(266 + 27))
	Balanced accuracy	80.6 (4.6)	85.3 ((79.7 + 90.8)/2)
	Coverage	51.6 (4.7)	54.5 (352/646)
QSAR2 Approach 7 Sub-models: 7	Sensitivity	75.6 (5.0)	-
	Specificity	89.8 (1.5)	-
	Balanced accuracy	82.7 (2.2)	-
	Coverage	57.8 (5.4)	-

*A five times two-fold cross-validation, ** A blinded external validation with the experimental results of the test set being masked to the model developers at DTU Food.

Table 4. The overlap in substances and unique structures between the U.S. EPA and REACH-PRS inventories.

Overlap analysis	U.S. EPA*	REACH-PRS**	Total number	In common		Unique to a set	
				REACH-PRS in U.S. EPA	U.S. EPA in REACH-PRS	REACH-PRS	U.S. EPA
Structure entries	32,197	72,524	104,721	27,444	19,279	45,080	12,918
Unique structures	32,197	60,281	92,478	19,279	19,279	41,002	12,918

*U.S. EPA: QSAR-ready structures from an U.S. EPA selected inventory of man-made chemical structures to which humans are potentially exposed, ** REACH-PRS: QSAR-ready structures from the REACH pre-registered substances list

Table 5. The coverage (AD) and the number of active/inactive predictions of the U.S. EPA and REACH-PRS inventories in QSAR1 and QSAR2.

	Total	QSAR 1			QSAR2		
		In AD (%)	Active (%)	Inactive (%)	In AD (%)	Active (%)	Inactive (%)
U.S. EPA*	32,197	16,898 (52.5)	2855 (16.9)	14,043 (83.1)	19,392 (60.2)	3201 (16.5)	16,191 (83.5)
REACH-PRS**	72,524	38,661 (53.3)	7,128 (18.4)	31,533 (81.6)	45,540 (62.8)	8,790 (19.3)	36,750 (80.7)
REACH-PRS unique	60,281	32,334 (53.6)	5,879 (18.2)	26,455 (81.8)	37,784 (62.7)	7,166 (19.0)	30,618 (81.0)

*U.S. EPA: QSAR-ready structures from an U.S. EPA selected inventory of man-made chemical structures to which humans are potentially exposed, ** REACH-PRS: QSAR-ready structures from the REACH pre-registered substances list

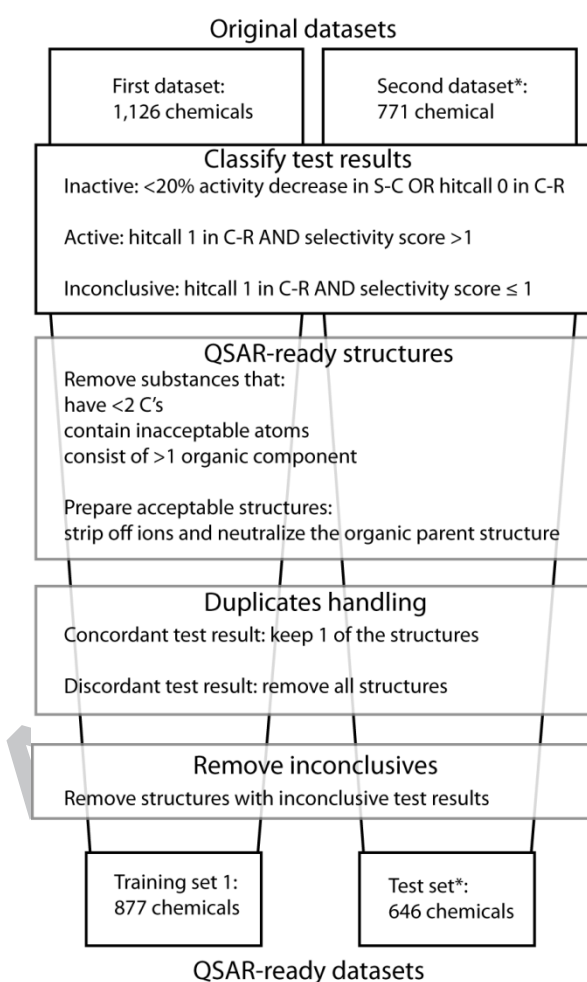


Figure 1. An overview of the dataset preparation procedure. S-C, single concentration screening; C-R, concentration response screening. *The experimental results of the dataset were blinded to the modelers at DTU Food

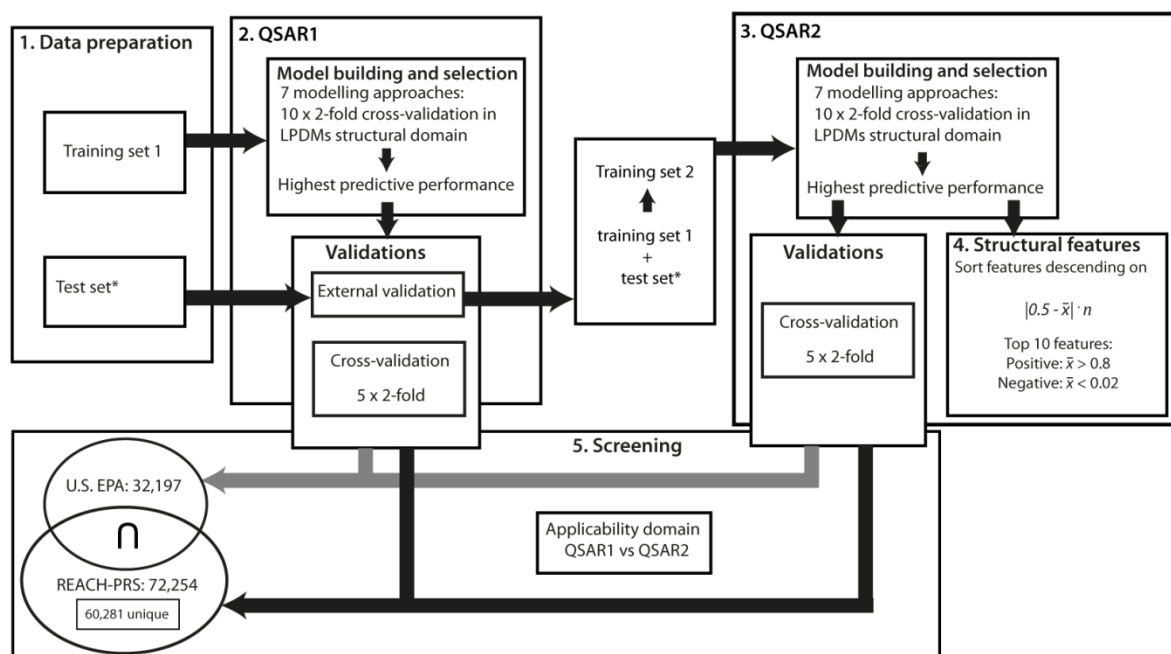


Figure 2. An overview of the datasets, modeling, structural feature sorting and screening. \bar{x} is the mean TPO inhibition experimental activity and n is the number of training set structures.

* The experimental results of the dataset were blinded to the modelers at DTU Food until after external validation had been performed.

13/0 benzene, 1,3-dihydroxy-	13/2 Scaffold 288	11/1 benzene, 1-alkyl-,4-amino(NH ₂)-	9/0 benzene, 1,2-dihydroxy-	9/2 Scaffold 297
6/0 alcohol, alkenyl-	7/1 Scaffold 576	5/0 benzene, 1-alkoxy-,4-hydroxy-	5/0 Scaffold 306	6/1 Scaffold 574
0/71 Scaffold 110	1/62 Scaffold 342	1/57 Scaffold 210	0/52 Scaffold 253	0/49 Scaffold 303

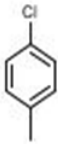
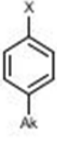
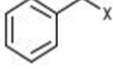
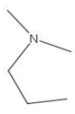
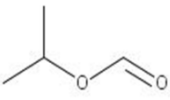
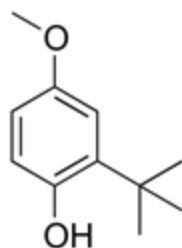
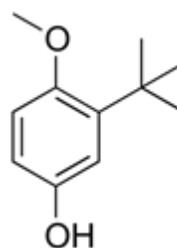
0/47 	0/44 	0/41 	0/36 	0/35 
Scaffold 108	benzene, 1-alkyl-,4-halo-	halide, benzyl-	Scaffold 454	Scaffold 194

Figure 2. The structural features used in QSAR2 were sorted on $|0.5 - \bar{x}(\text{TPO inhibition activity})| \cdot n$, and the ten most frequent and discriminating structural features alerting for activity ($\bar{x}(\text{TPO inhibition activity}) > 0.8$) and inactivity ($\bar{x}(\text{TPO inhibition activity}) < 0.02$) are shown here. Ak matches saturated carbon and X matches the halogen atoms Cl, Br, I or F. Numbers in the upper left corners display the ratio of TPO inhibitors/non-inhibitors in training set 2 for the specific structural feature.

2-*tert*-Butyl-4-hydroxyanisole (2-BHA)



3-*tert*-Butyl-4-hydroxyanisole (3-BHA)*



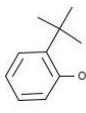
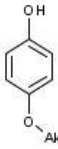
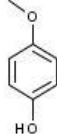
9/2 	5/0 	3/0 
Scaffold 297	benzene, 1-alkoxy-,4-hydroxy-	benzene, 1-hydroxy-,4-methoxy-

Figure 3. The two isomers of BHA and the three predictive structural features alerting for activity in QSAR2 selected based on highest $|0.5 - \bar{x}(\text{TPO inhibition activity})| \cdot n$ and an $\bar{x} > 0.8$. *3-BHA (CASN 121-00-6) was included in the training set and is the closest analog to 2-BHA (CASN 88-32-4).

Highlights: QSAR Models for Thyroperoxidase Inhibition and Screening of U.S. and EU Chemical Inventories

- Development of two QSAR models for TPO inhibition
- Highly performing and robust models according to cross- and external validations
- Predictions of two large U.S. and EU chemical inventories for TPO inhibition
- Identification of structural features associated with TPO inhibition

ACCEPTED MANUSCRIPT