International Journal of
Molecular Sciences

MDPI

Article

Development of a Robust Read-Across Model for the Prediction of Biological Potency of Novel Peroxisome Proliferator-Activated Receptor Delta Agonists

Maria Antoniou [1,2,3], Konstantinos D. Papavasileiou [1,2,3], Georgia Melagraki [4], Francesco Dondero [3,5], Iseult Lynch [3,6] and Antreas Afantitis [1,2,3,*]

1 Department of Chemoinformatics, NovaMechanics Ltd., 1046 Nicosia, Cyprus; antoniou@novamechanics.com (M.A.); papavasileiou@novamechanics.com (K.D.P.)
2 Department of ChemoInformatics, NovaMechanics MIKE, 18545 Piraeus, Greece
3 Entelos Institute, 6059 Larnaca, Cyprus; francesco.dondero@uniupo.it (F.D.); i.lynch@bham.ac.uk (I.L.)
4 Division of Physical Sciences & Applications, Hellenic Military Academy, 16672 Vari, Greece; georgiamelagraki@gmail.com
5 Department of Science and Technological Innovation, Università del Piemonte Orientale, 15121 Alessandria, Italy
6 School of Geography, Earth and Environmental Sciences, University of Birmingham Edgbaston, Birmingham B15 2TT, UK
* Correspondence: afantitis@novamechanics.com

Abstract: A robust predictive model was developed using 136 novel peroxisome proliferator-activated receptor delta (PPARδ) agonists, a distinct subtype of lipid-activated transcription factors of the nuclear receptor superfamily that regulate target genes by binding to characteristic sequences of DNA bases. The model employs various structural descriptors and docking calculations and provides predictions of the biological activity of PPARδ agonists, following the criteria of the Organization for Economic Co-operation and Development (OECD) for the development and validation of quantitative structure–activity relationship (QSAR) models. Specifically focused on small molecules, the model facilitates the identification of highly potent and selective PPARδ agonists and offers a read-across concept by providing the chemical neighbours of the compound under study. The model development process was conducted on Isalos Analytics Software (v. 0.1.17) which provides an intuitive environment for machine-learning applications. The final model was released as a user-friendly web tool and can be accessed through the Enalos Cloud platform's graphical user interface (GUI).

Keywords: PPARδ agonists; molecular docking; in silico modelling; machine learning; Isalos Analytics Platform

Citation: Antoniou, M.; Papavasileiou, K.D.; Melagraki, G.; Dondero, F.; Lynch, I.; Afantitis, A. Development of a Robust Read-Across Model for the Prediction of Biological Potency of Novel Peroxisome Proliferator-Activated Receptor Delta Agonists. *Int. J. Mol. Sci.* **2024**, *25*, 5216. https://doi.org/10.3390/ijms25105216

Academic Editor: Antonio Carrieri

Received: 1 April 2024
Revised: 2 May 2024
Accepted: 3 May 2024
Published: 10 May 2024

1. Introduction

PPARs are members of the nuclear receptor (NR) superfamily of proteins, whose functions are essential for cell signalling, survival, and proliferation, which comprises 48 members in humans [1] and function as ligand-activated transcription factors. Their role is central in the regulation of diverse biological processes, encompassing immune system function, development, reproduction, and homeostasis [2], involving the control of gene expression related to fatty acid utilisation and storage [3]. Target gene regulation is achieved by PPAR binding to characteristic sequences of DNA bases, called peroxisome proliferator response elements (PPREs). PPREs are active as heterodimers with the receptor for 9-cis-retinoic acid (retinoid X receptor or RXR), and thus play a critical role in modulating the actions of hormones and ligands. Furthermore, PPARs participate in

various cellular processes, including glucose utilisation, cell proliferation, cell differentiation, inflammatory responses, and adipogenesis [4]. Depending on their tissue expression, they are classified into three subtypes, namely PPARα, PPARγ, and PPARβ/δ [5], which also reflects their distinct physiological roles.

Although the PPAR subtypes exhibit a significant degree of amino acid sequence similarity, they vary in ligand selectivity and target genes in a species-specific manner [6]. For example, PPARδ exhibits significant expression levels in organs characterised by elevated rates of oxidative metabolism, such as the heart, skeletal muscle, and liver, while playing a regulatory role in the utilisation of fatty acids and glucose, as well as in antioxidant defence mechanisms [7].

Several attempts have been previously reported in the literature to construct QSARs for the establishment of statistically significant correlations for the prediction of PPAR agonists' behaviour [8]. Specifically, classical (1D) and (2D)-QSAR models [9–11] were developed using a dataset evaluated by Wickens et al. [12], linking molecular properties and structural characteristics, respectively, to the activity of the compounds, with the predictions being confirmed through docking methods. In addition, QSAR modelling was represented using the three-dimensional (3D) properties of the ligands to predict the biological potency of PPARδ receptors by exploiting methods such as comparative molecular field analysis (CoMFA), which provides a visual display of the active centres in compounds that indicates the fragments contributing maximally to the activity profile of the compounds, and comparative molecular similarity indices analysis (CoMSIA), which expresses the fields in terms of molecular similarity indices rather than the usually applied Lennard–Jones- and Coulomb-type potentials, as used in CoMFA [13,14]. Other studies [15,16] employed a different dataset [17] comprising 34 PPARδ partial agonists, for the establishment of hologram QSARs (HQSARs) by using molecular holograms as variables for their predictive schemes. Lastly, Daadoosh et al. [18] employed a machine-learning method, iterative stochastic elimination (ISE), to perform the virtual screening of over 1.5 million compounds and identified thirteen highly selective PPARδ agonists [19]. It was apparent from these studies that the inclusion of molecular docking calculations in the models appears to ameliorate their poor predictive performance in the absence of the molecular docking information.

In the present study, we introduce a robust predictive model utilising a set of 136 novel PPARδ agonists that, according to the new OECD definition, includes per- and polyfluoroalkyl substances (PFAS) [20]. The model integrates diverse structural descriptors with docking calculations to predict the biological activity of PPARδ agonists, adhering to the criteria outlined by the OECD for the development of QSAR and read-across models. Concentrating specifically on small molecules, the model aids in identifying highly potent and selective PPARδ compounds, employing a read-across concept to delineate the chemical neighbours of the compound under investigation and thus to classify it as active or non-active based on their biological potency score.

2. Results

The techniques mentioned in the Materials and Methods section for the development of the predictive model were implemented in the Isalos Analytics Platform. First, the initial dataset of 136 novel compounds was derived from the PubChem public repository using the Enalos+ KNIME node 'Main PubChem'. Each small molecule or compound was accompanied by an extensive set of 777 molecular descriptors that encode their structural, topological, and geometrical characteristics, along with a calculation of their binding affinity for the human PPARδ protein structure. The dimensionality of the data was reduced after numerous descriptors were excluded from the set using the 'Remove Column' function and a low variance filter (20%). The 245 remaining descriptors' values were transformed with a Gaussian normalisation function into a new set of values that lie on a similar scale and whose mean is zero and standard deviation is one. A clustering technique was employed for the distinction of the novel molecules into two classes that

represent biological potency. Through a greedy algorithm, the number of input descriptors was further reduced and the most relevant descriptors that exhibit optimal correlation to the target variable were distinguished. Following the pre-processing steps, a kNN classification algorithm was used as the modelling methodology, since it allows the observation of the five neighbouring instances of each test compound from the training set. This read-across approach allows the exploration of the adjacent chemical space of the compound under study [21], wherein the closest five neighbours are more likely to share physicochemical properties and structural patterns with the molecule of interest or test compound (Figure 1).

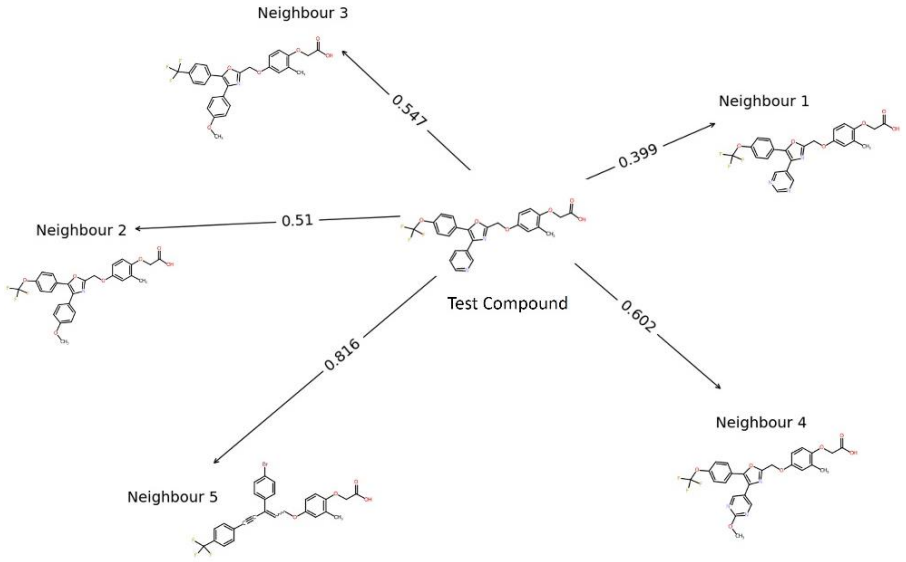

Figure 1. Network of a test compound (PubChem CID: 44627413) with Euclidean distances from its five closest neighbours.

2.1. Interpretation of the Selected Descriptors

As mentioned above, the variables that were most pertinent to the modelling target were selected from a pool of 777 molecular descriptors after the 'BestFirst' function in Isalos was applied to the training dataset. Since the descriptors are mathematical representations of the molecules [22], the interpretation of the selected variables grants insight into the most significant factors that control the behaviour of chemicals against the PPARδ nuclear receptor. The eleven favoured descriptors, presented in Table S1, encode information mainly on the compounds' bulk characteristics, their autocorrelation, and topological indices.

Firstly, the Broto–Moreau spatial autocorrelation descriptor (ATSe,7), which emerged as the most significant descriptor overall, is a measure whereby the atoms of a molecule are represented by an atomic property such as the Sanderson electronegativity [23,24]. It provides information on how the atomic property is distributed on the topological structure of the molecule, thus a higher electronegativity distribution within the molecule contributes to the biological activity of PPARδ. The total information content (TIC_m) was also selected, which quantifies the complexity of a knowledge graph. Higher values amount to higher molecular graph complexity and highest effect concentrations (EC_{50}s), evidenced from the positive correlation coefficient between the descriptor and the associated biological activity of the test compound.

The Burden eigenvalues are also among the highly influential descriptors. This descriptor is computed as a solution to the characteristic equation of the Burden connectivity matrix (B), whose elements correspond to a topological distance between pairs of atoms [25], and its diagonal elements (Bii) are given by the van der Waals volume values. Another important descriptor correlated with PPARδ activity is the sum of

topological distances between the vertices of oxygen atoms and fluorine or sulphur atoms, calculated as the row sum of the distance matrix. Topological distances are the number of edges along the shortest path between two specific atoms, measuring the number of involved bonds [26].

Last but not least, the Kier shape index (S2k) was proven as a valuable variable that describes the shape of the molecule in terms of counts of two bond paths [27]. It captures the degree of star graph-likeness and provides information about the branching and the flexibility of the molecular structure. Higher S2k values indicate a greater degree of flexibility within the molecules. According to Xu et al. [28], PPAR activation is effective when the linked compound is flexible, thus less pliable compounds exhibit reduced potency (and thus have higher EC_{50} values). Even though we highlight the features that describe the biological system in an effective manner, further validation against experimental data is needed to establish meaningful correlations between the above-mentioned descriptors and the biological potency.

2.2. Model Validation

2.2.1. Metrics and Statistics

Assessing the predictive performance of the model using several statistical criteria ensures that it can classify the instances effectively. After the implementation of the classification machine-learning algorithm, different statistics were employed for the evaluation of the model, based on the number of correct predictions and misclassifications of the test set [29]. Provided that the model aims to predict the potency class of a target compound, characterising it as either "active" or "inactive", the problem boils down to a binary classification one.

Therefore, a confusion matrix for the test set is presented, which is essentially a table recording the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions in comparison with the actual classes of the agonists (Table 1).

Table 1. Confusion Matrix summarising the number of correct and incorrect predictions from the test set.

Class	Predicted Active	Predicted Inactive
Actual Active	20	3
Actual Inactive	2	16

Based on the confusion matrix, various classification performance indications can be obtained, including accuracy, sensitivity, and precision, all synopsized in Table 2. The measurements for the goodness-of-fit and predictivity were higher than 0.80 when applying the kNN algorithm, with an optimised value of k = 5 to the test set, which denotes the ability of the model to accurately capture patterns and return reliable predictions.

Table 2. Accuracy statistics of the predictive model.

Metric	Metric Formula	Metric Value
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	87.8%
Sensitivity	$\frac{TP}{TP + FN}$	87.0%
Precision	$\frac{TP}{TP + FP}$	90.9%
F1-Score	$\frac{2TP}{2TP + FP + FN}$	88.9%
Matthews Correlation Coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	0.755

Cohen's kappa	$\frac{2(TP \times TN - FP \times FN)}{(TP + FP)(FP + TN) + (TP + FN)(TN + FN}$	0.754

2.2.2. Internal and External Validation

Internal validation was performed through the Y-randomisation procedure in order to ensure the robustness of the predictive model [30]. Specifically, the observed target feature's values are randomly assigned to other compounds; thus, the original values of the descriptors now correspond to a different endpoint variable. Provided that the original model is robust, when it is applied on the test set it is expected that the predicted values are not close to the confounding ones, which is verified through the inadequate performance of the model. This technique was performed using the 'Y-randomization' node in KNIME contained in Enalos+ [31]. Calculations were repeated for five randomisations, ensuring that the model was not based on chance correlation and overfitting. When the algorithm was trained on disarranged targets, the predictive performance of the obtained models was statistically low, whereas the validation measurements of the original model were adequate (Figure 2). Specifically, the accuracy values derived fluctuated between 41.5% and 61.0% and were significantly lower compared to the accuracy value of 88.9% of the original model.

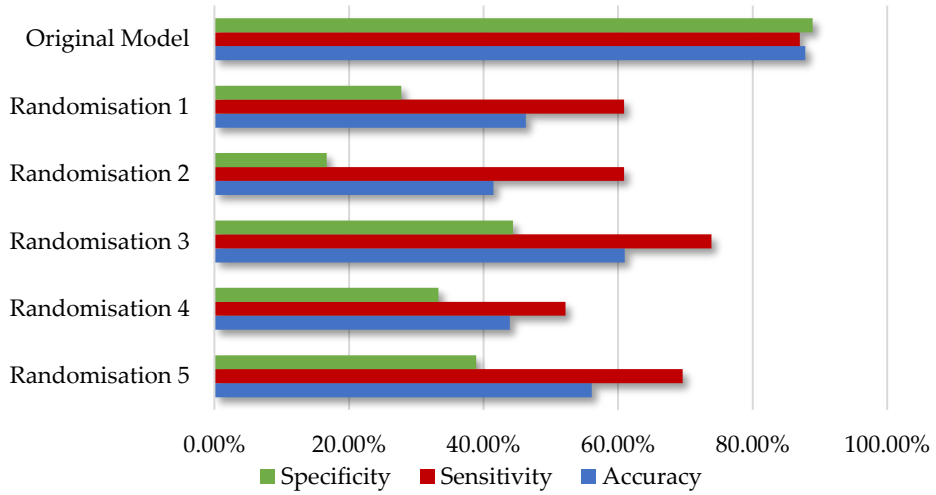

Figure 2. The predictive power of the original model compared with the models obtained from the five Y-randomization tests.

For external validation, the original subset was partitioned into the training set, which was used during model development, and the test set, which was used solely for validation. More precisely, the developed model was applied to the test set, which was not included in the development process and was later involved during the model's performance assessment. This technique validates that the read-across model's performance is satisfactory on unseen data that were not involved in the construction of the classifier.

2.3. Applicability Domain

The domain of applicability (APD) is defined after model validation and determines the area of reliable or unreliable predictions. It is essential for describing the limitations of a model and the degree of similarity between the compound of interest and the model training set, as determined by different approaches. A distance-based method is used in this work, which involves similarity measurements based on the Euclidean distances among all training data, compared to a predefined APD threshold [32,33].

At first, the average value of all Euclidean distances is calculated and then the set whose distances are lower than the average value are excluded from further calculations. Next, a new average value (d) and the standard deviation (σ) of the remaining distances sets is determined, thus the APD threshold is calculated as:

$$APD = z\sigma + d, \quad (1)$$

where, z is an empirical parameter whose default value is 0.5. In the case that the distance from an external compound to its nearest neighbour (among the test set data) is smaller than the APD threshold, then the prediction is considered reliable. The APD thresholding was performed in the Isalos platform, Statistics → Domain—APD. The selected APD model, developed from the training subset, was employed from Analytics → Existing Model Utilisation in order to be applied to the test subset. The obtained APD threshold value was equal to 3.682, while the predictions were regarded as reliable for all compounds included in the test set. Table S2 of the Supplementary Information File includes the selected descriptors, an indication of the actual class of each compound in the testing set, and the prediction obtained from the model.

2.4. Model Availability

In order to accelerate the assessment of small molecules and their activity towards the PPARδ nuclear receptor, the read-across predictive model was disseminated as a publicly available web application in the Enalos Cloud Platform. Several fully validated cheminformatics models [34,35] are hosted by the Enalos Cloud Platform, supporting the scientific community by making the predictive workflows easily accessible to anyone interested. The model's functionality can be easily accessed through a user-friendly interface that requires limited input, and no coding skills, in order to provide predictions.

Figure 3 portrays the initial interface, where a brief description of the model development is given, along with three different ways to insert compounds and initiate predictions. Either the SMILES notations can be entered manually or an SDF file that contains the structure of one or multiple compounds can be browsed and uploaded by the user. As a further option, users can use a drawing interface (Figure 3d) to design the molecule of interest. In the sketcher field, the user can also transform the initial molecule by adding different functional groups such as alkanes, amines and amides, benzene rings etc., or more complex chemical structures such as steroids and amino acids.

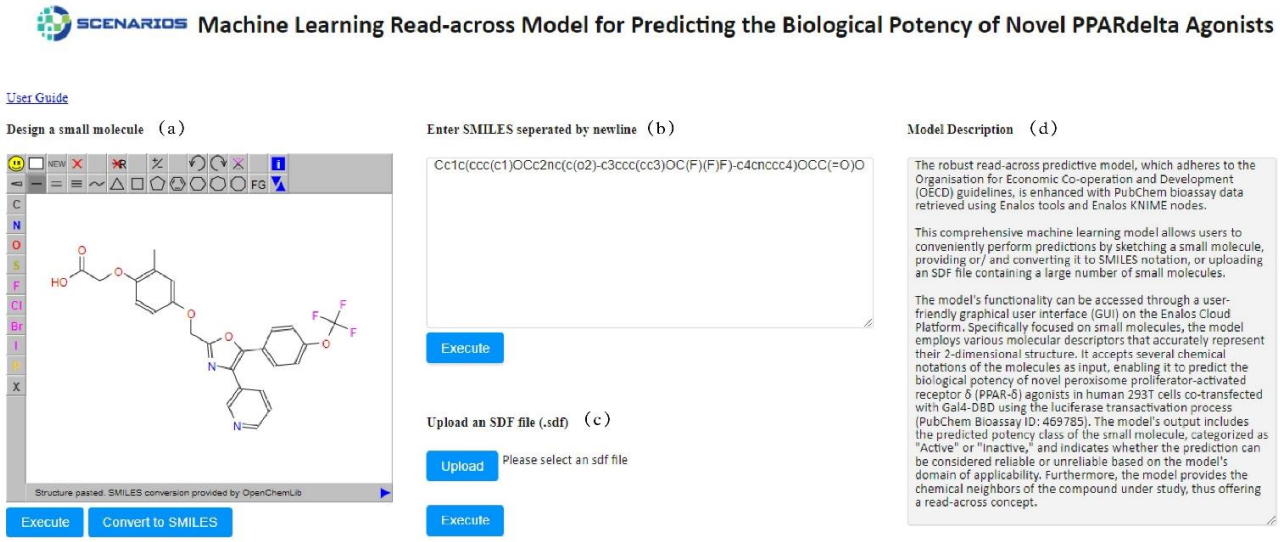

Figure 3. PPARδ environment in the Enalos Cloud Platform: The Design Molecule field for input compounds (**a**), the SMILES (**b**) and the SDF (**c**) field for input compounds, along with a brief description of the model (**d**).

For the demonstration of the tool's functionality, five compounds of interest—including two substances with a perfluorinated methyl group (−CF_3), i.e., PFAS—were selected from the PubChem library (CIDs: 155547595, 54764927, 51346913, 46230234, 137464756). The selected chemicals share at least 95% Tanimoto similarity with active compounds from the initial dataset, and their SMILES notations (Figure 4a) were extracted with the Enalos+ 'Main PubChem' node in KNIME. A prediction is generated within seconds, and the output includes a table that presents the classification of the compound's activity, the five nearest neighbours of the input compound from the training set, and the Euclidean distances from each of the neighbours (Figure 4). The distance of each submitted compound calculated according to the APD of the model is presented, along with an indication of the reliability of each prediction. As presented in Figure 4b, when the calculated domain of the small molecule is higher than the APD threshold, the web application highlights that the prediction is not reliable. As seen from this case study, the read-across model can be used within a virtual screening framework to identify whether similar chemicals can be potentially used as activators of the PPARδ receptor.

In order to enhance the accessibility and programmability of the predictive model, a Representational State Transfer (REST) application programming interface (API) was incorporated (https://enaloscloud.novamechanics.com/scenarios/swagger-ui/index.html, accessed on 11 January 2024). This method is useful as it allows seamless integration of the computational workflow into various systems and platforms and enables users to explore the capabilities of the model without direct access to the original workflow. Users are, therefore, able to incorporate the model into their own workflows through the API (Figure 5). It is further used to communicate with the Isalos Analytics Platform to exchange data for the straightforward execution of the model. The API was implemented using the POST request method, since it is suitable for transferring substantial amounts of structured input data securely. The submission of a tuple of data input (i.e., containing either a single or multiple SMILES string(s) of the desired compound(s)) in JSON format is needed to use the PPARδ agonists bioactivity API:

```
[
{
"smiles": "Cc1c(ccc(c1)OCc2nc(c(o2)-c3ccc(cc3)OC(F)(F)F)-c4cnccc4)OCC(=O)O"
}
]
```

The user is able to make a request through a data transfer software such as Client URL:

```
curl -X POST "https://enaloscloud.novamechanics.com/scenarios/apis/ppardelta/smiles" -H "accept: application/json" -H "Content-Type: application/json" -d "[ { \"smiles\": \"Cc1c(ccc(c1)OCc2nc(c(o2)-c3ccc(cc3)OC(F)(F)F)-c4cnccc4)OCC(=O)O\" }]"
```

and obtain the corresponding results of the GUI environment, as seen in Figure 4. The returned response includes class prediction, the closest neighbours, and the Euclidean distances from the molecule in question, and the APD indicating the reliability of the prediction:


```
[
{
"id": "cluster_0",
"idNN1": "Entry 20",
"distNN1": 0.7024641202204505,
"idNN2": "Entry 89",
"distNN2": 0.8202558963827292,
"idNN3": "Entry 80",
"distNN3": 0.8326502973790398,
"idNN4": "Entry 52",
```

"distNN4": 0.8342452655891001,
"idNN5": "Entry 5",
"distNN5": 0.8364009145055961,
"idNN6": "Entry 46",
"distNN6": 0.8772952879358424,
"domain": 2.714170703054797,
"apd": 3.4716837408236625,
"predictionReliablility": "reliable",
"knnprediction": "inactive"
}
]

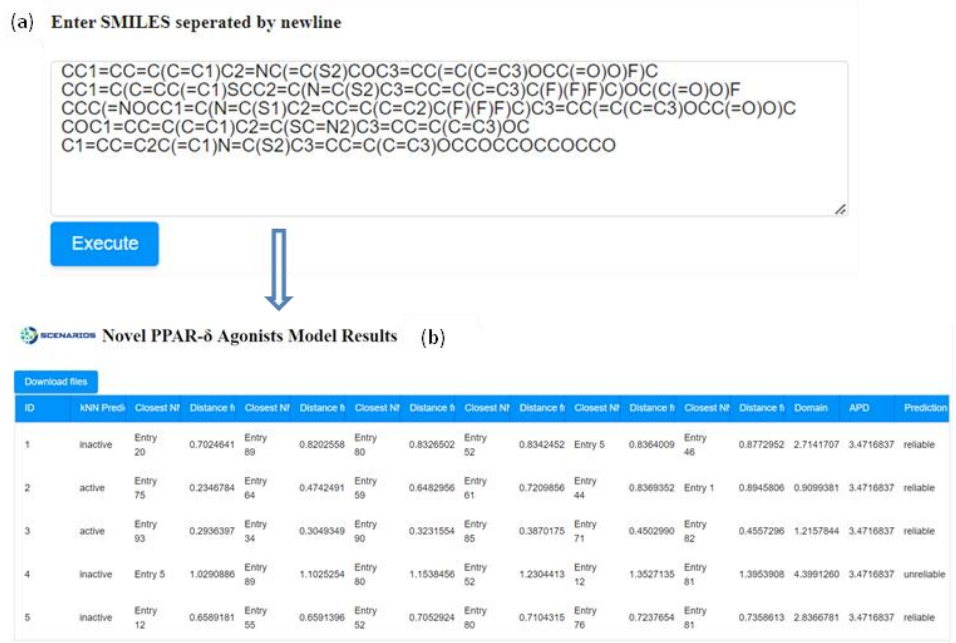

Figure 4. Entering the SMILES notations of five different compounds as input to the web application (**a**) and the generated output page (**b**). Out of the five compounds tested, the kNN algorithm identified only the two PFAS congeners (CID 54764927 and 51346913) as active.

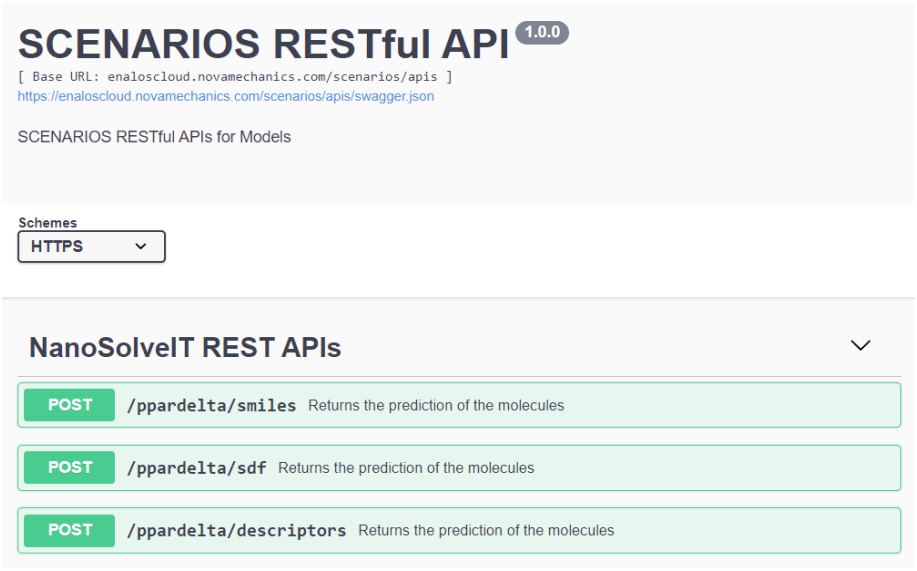

Figure 5. The REST API environment (accessed on 11 January 2024) for the PPARδ agonist bioactivity prediction.

3. Discussion

In summary, in the present study an in silico predictive model that correlates novel PPARδ agonists' two-dimensional chemical structures to their biological potencies in terms of nuclear receptor activation (PPARδ) was successfully developed. PPARδ, a subtype of the nuclear receptor superfamily, plays a pivotal role in regulating cellular metabolic functions and in modulating diseases associated with changes in lipid and glucose homeostasis. The search for highly potent and selective compounds that act as PPARδ activators is still ongoing [36]; hence, the development of computational methods that assist in the identification of such compounds is crucial.

The predictive model in this study uses an initial dataset sourced from the PubChem BioAssay public repository that consists of 136 novel molecules tested in human 293T cells co-transfected with Gal4-DBD via a process called luciferase transactivation. The chemical structure of each compound of the dataset is represented through a comprehensive set of 777 molecular descriptors, generated using the EnalosMold2 specialised module in KNIME. Apart from the structural properties of the molecules, docking calculations were included as a supplementary variable. All analysis steps, including the normalisation of the descriptors, the selection of the most correlated variables, the algorithm selection, and the validation of the final model were executed within the Isalos Analytics software (v. 0.1.17), an advanced platform for machine-learning applications. The model underwent internal and external validation, through the use of different subsets for training and testing and Y-randomization tests, demonstrating strong performance. Fully adhering to the guidelines posed by the OECD, the domain of applicability was described, defining the region in the chemical space where the generated predictions can be trusted. The interpretation of the molecular descriptors' influence on the compounds' biological activity is discussed. While the descriptors' effect on the biological potency was emphasised, full comprehension of their effects on the biological potency requires additional experimental validation. The model is fully documented via a QMRF (S1) report, which was prepared for the reporting of the key information on this read-across model for regulatory use.

Although successful attempts to derive statistically significant relationships on the biological activity of PPARδ have been reported in the past, the present work applies a different modelling approach. In comparison to the use of traditional QSAR methodologies, this work enables the categorising of an unknown compound into 'active' or 'inactive' and introduces a read-across paradigm that provides information on the five closest instances (neighbours) from the training data. The separation of the compounds into two distinct classes facilitates rapid decision-making in early drug discovery. While the current study is tailored to distinguishing compounds for the activation of PPARδ, the read-across methodology can be adapted for the identification of small molecules that can act as agonists or antagonists against other biological targets. The proposed read-across methodology can be extended across other proteins or enzymes to describe structure–activity relationships with potential regulators. A similar in silico approach can be implemented for other nuclear receptors, such as the PPARα or PPARγ ligand-activated transcription factors, starting from the identification of experimental datasets that identify the agonists and antagonists of the target nuclear receptors. Recently, a computational tool [37] was developed for the prediction of chemical molecules' binding class to multiple nuclear receptors, but the previous tool does not employ the read-across framework and does not distinguish between agonist and antagonist compounds regarding the PPARδ receptor.

Additionally, the present work enables export of the read-across model as a web tool via the Enalos Cloud Platform. The web tool can be easily accessed through the following link: http://www.enaloscloud.novamechanics.com/scenarios/ppardelta/ (accessed on 11 January 2024). This comprehensive model allows users to provide input data by sketching a small molecule, entering and converting it to SMILES notation, or by uploading an SDF file containing a large number of small molecules. The web application offers the

possibility to use the predictive capabilities of the model from anywhere, assisting scientists and researchers in the continuing process of detecting PPARδ activators.

4. Materials and Methods

The comprehensive analysis for the development of the read-across predictive model was performed with the Isalos Analytics Platform [38]. Isalos is a simple, straightforward software developed by NovaMechanics Ltd (Nicosia, Cyprus) (https://isalos.novamechanics.com/, accessed on 4 December 2024), which allows the implementation of machine-learning workflows without requiring coding skills. The Isalos Platform provides a practical interface through the use of menus, tabs, and buttons, while each tab acts as a node and allows the transformation and transition of data in tabular form. All analysis steps, including data preparation, feature selection, algorithm building, and model validation, were performed using the special functions encoded in the software. Leveraging the built-in functions, along with the Enalos+ proprietary nodes [31] accessible through the KNIME Analytics Platform, results in a combined workflow, as illustrated in Figure 6. The final predictive model is fully validated according to the OECD guidelines [39] and its key information was summarised and reported using the QSAR Model Reporting Format (QMRF) template, following the guidance of the Joint Research Centre and the European Centre for Validation of Alternative Methods [40]. The completed reporting template can be found in the electronic Supplementary Information File (ESI S1).

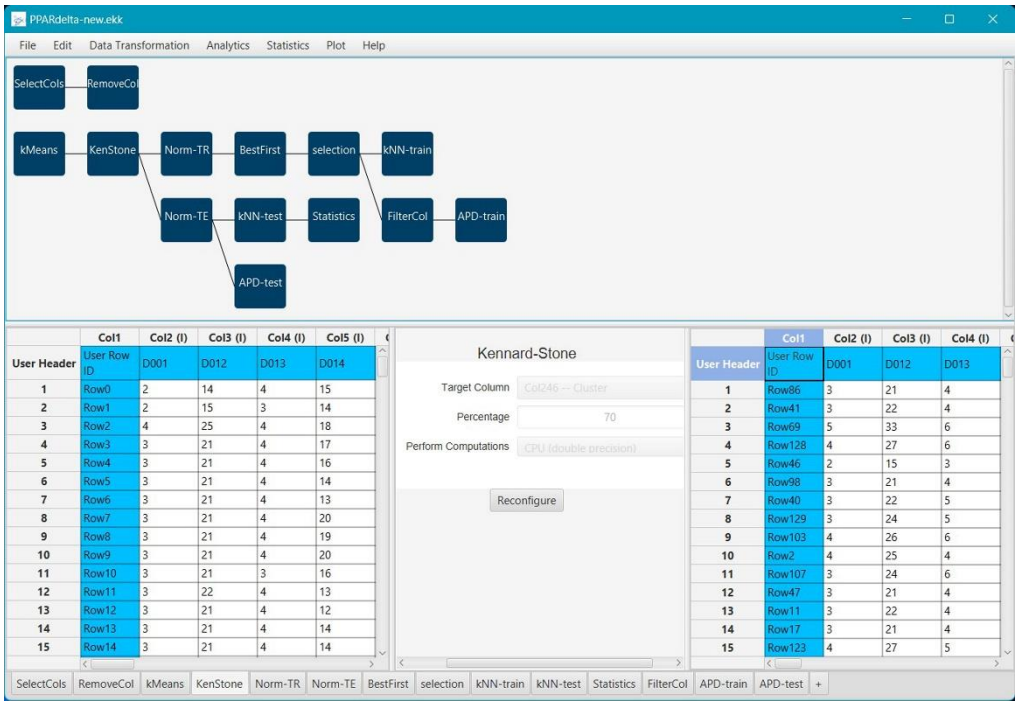

Figure 6. Implementation of the model development process in Isalos Analytics Platform.

4.1. Dataset

Epple et al. [41] performed a high throughput screening (HTS) of approximately 1 million chemical compounds, defining hits as molecules that induced luciferase activity and utilising this assessment as an indicator of agonist activity against the human PPARδ ligand binding domain. The luciferase gene encodes a 61-kDa enzyme that oxidises D-luciferin in the presence of ATP, oxygen, and Mg^{2+}, yielding a fluorescent product that can be quantified by measuring the released light via a luminescence assay. The molecules were tested in a human embryonic kidney cell line, 293T, co-transfected with a chimeric plasmid with the yeast GAL 4 DNA-binding domain (DBD). The dataset was retrieved

from PubChem BioAssay, a public repository for the biological activities of small molecules and small interfering RNAs hosted by the National Institutes of Health (NIH), under the numeric identifier AID 469785 [42]. All 136 retrieved oxazole-based compounds from the initial dataset are accompanied by a standardised measure of potency, the half maximal effective concentration (EC_{50}), which determines the agonist concentration needed to elicit half of its maximum biological effect, in this case cytotoxicity to human embryonic kidney cells. The EC_{50} value is inversely related to a compound's potency [43]. It is important to note that Garcia et al. [13] and Nandy et al. [44] utilised the same experimental dataset to apply 2D and 3D QSAR methodologies for the assessment of the biological activity of PPARδ agonists. However, they used a subset that comprised just above 100 compounds, in contrast to the entirety of the dataset as used in this work. The inclusion of the complete dataset ensures the generalizability of our model and provides a broader applicability domain. Additionally, while the other studies focus on the derivation of regressive QSARs, aiming to predict the value of a potency metric, the read-across model developed in this study deploys a different modelling approach, classifying the small compounds into the 'Active' and 'Inactive' categories and enabling prediction of which class an unknown small molecule fits into, based on the applicability domain.

4.2. Calculation of Descriptors

The Mold2 [45] software package (version 2.0), accessed through the Enalos+ node 'EnalosMold2' [31] node available in KNIME, was used for the retrieval of molecular descriptors representing characteristics of the small molecules. Requiring only the 'Simplified Molecular Input Line Entry System' (SMILES) notations as input, in the Structure Data File (SDF) format, Mold2 calculates 777 molecular descriptors based on the one-dimensional (1D) and two-dimensional (2D) structure of each compound. The calculated 1D descriptors are related to counts of atoms, and the 2D descriptors mainly refer to bonds and functional groups, physicochemical properties, autocorrelation, charge, connectivity, and topological features of the molecules.

4.3. Data Pre-Processing

Data modification is a crucial step in data analysis, as it allows the cleaning, reduction, and transformation of data as a means of eliminating noisy data and improving the performance of machine-learning algorithms. Firstly, duplicates containing more than 80% of the same repeated values of particular parameters were removed from the dataset using the Isalos' 'Remove Column' function. Furthermore, the extracted raw data were pre-processed with a low variance filter in order to reduce the dimensionality of the dataset and filter out descriptors that have least impact on the target variable. An upper bound of 20% was chosen; thus, descriptors whose variance fell below the threshold were excluded from the following steps.

4.4. Clustering into Distinct Classes

Since the initial dataset's potency score, EC50, was available as a continuous variable without predefined classes, an unsupervised clustering method was used to explore the natural groupings of the unlabelled compounds. The k-means algorithm was chosen to perform an initial analysis and divide the instances into appropriate groups according to their activity indicators. The k-means is a useful method for partitioning variables into k separated clusters, where each cluster is represented by its centroid average [46]. The algorithm begins by randomly selecting k observation compounds as the initial centroids, then proceeds by assigning each observation to the cluster whose centroid is closest to it. Euclidean distance is used as a distance-calculating method. The centroids are then recomputed as the average of the observations allocated to the cluster, and this process is repeated until the assignment of observations to clusters no longer changes [47].

By employing a clustering method, an initial partitioning is created, assuming that k-means partitions (k = 2) now distinguish all data as belonging to one of two biological activity classes (active or inactive) based on their log-transformed EC_{50} values given in micromolar (μM) units. The logarithmic transformation of the values was preferred, in order to reduce the skewness of the data [48]. Each cluster represents an activity class; therefore, the original regression problem is reduced to a simpler classification problem, assigning a class label to each observation. The higher the EC_{50} value, the more the concentration of a compound is required to obtain a 50% effect inducement, and the lower the potency. This designates that the compounds assigned in the cluster with a centroid of log(EC50) = 0.224 are considered inactive and those included in the cluster with a centroid of log(EC50) = −1.674 are regarded as active. The coverage of the two clusters is 58 and 78 compounds, respectively, portraying a broadly balanced dataset.

Feature scaling is another fundamental pre-processing step, which is performed after the k-mean clustering and normalises the range of the independent attributes. The selected method for normalisation is z-score scaling, used to transform the data to have a mean of zero and a unit standard deviation (Gaussian-distributed). In the Isalos Analytics Platform, Gaussian standardisation is available in Data Transformation → Data Manipulation → Z-score. Prior to further modelling, the collected data were also divided into two subsets, the training and test datasets, as an external validation procedure. The two representative sets were split 70/30%, respectively, using the Kennard–Stone algorithm [49,50], available in Isalos.

4.5. Variable Selection

The set of descriptors produced by the Mold2 Enalos+ is characterised by its considerable size and diversity, indicating the presence of numerous descriptors that may be redundant or unrelated to the forthcoming analysis. This issue renders feature selection a necessary step prior to modelling. By using 'Best First' as a search method, the most important variables out of the Mold2-derived descriptors are selected (from the 777 available descriptors) based on the training set to be included in the model. This method uses a greedy algorithm, starting with an empty feature set and iteratively adding or removing features based on certain criteria, in order to choose the successor out of all combinations, and it is implemented using the Isalos Analytics Platform through Analytics → Feature selection. Table S1 in the Electronic Supplementary Information (ESI) describes the 11 selected attributes as per the Handbook of Molecular Descriptors [22]. As an additional descriptor, the molecular docking scores of the small molecules to the PPARδ receptor were also taken into consideration. Including the binding affinity values in the dataset seems to improve the results and contributes to the variability of the data, since it increases the dataset and provides a more thorough examination of the interactions with the receptor.

4.6. Molecular Docking Calculations

Molecular docking calculations were conducted using the Vina-GPU 2.0 software [51,52], on the set of compounds retrieved from PUBCHEM, by employing the PPARδ homo sapiens structure (PDB ID: 3TKM [53]). The structural preparation of PPARδ involved the Enalos Asclepios KNIME pipeline [54] (Figure 7), encompassing tasks such as the addition of missing residues, removal of heteroatoms, replacement of non-standard residues, and addition of heavy atoms using the PDBFixer software (version 1.9) [55]. Hydrogen atoms were subsequently added with the pdb4amber utility of AmberTools21 [56]. Ligand preparation was performed using the Enalos Asclepios KNIME pipeline, incorporating steps such as the addition of missing hydrogen atoms via Open Babel [57], setting the pH value at 7.4, and conversion of 2D structures to 3D through energy minimization using AsclepiosGenerate3DCooords [54]. The Enalos Asclepios KNIME nodes and workflow used in this study are proprietary to NovaMechanics Ltd and require a licensing agreement for access.

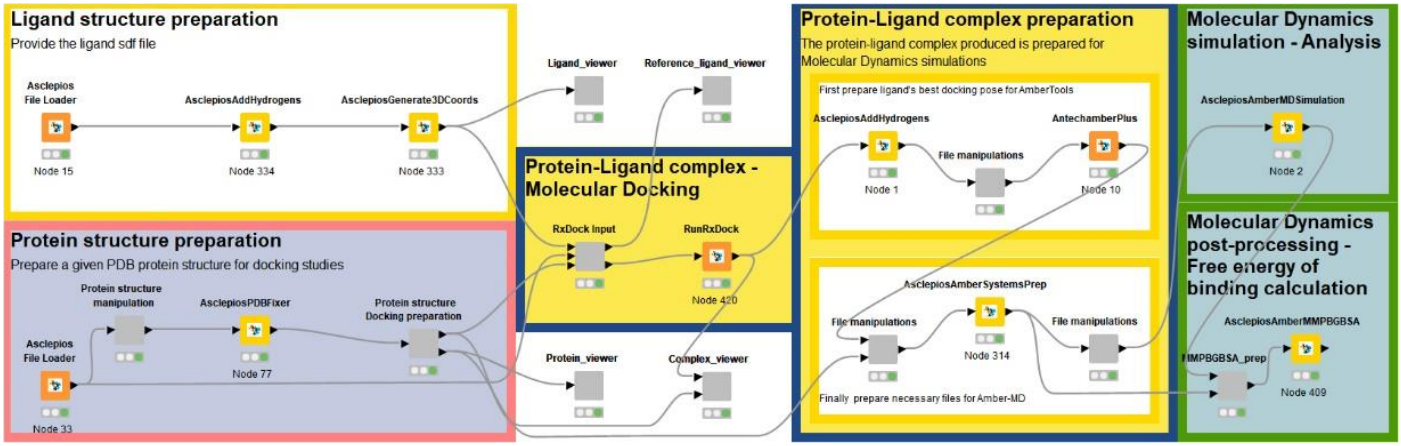

Figure 7. The Enalos Asclepios KNIME pipeline for automation of the drug discovery pipeline, applied here to screening PPAR biological activity.

The PPARδ structure and 136 compounds for docking were prepared using AutoDock Tools 1.5.7 python libraries [58,59], with partial atomic charges assigned based on the Kollman United Atom and Gasteiger–Marsili schemes, respectively, by previously merging non-polar hydrogens to heavy atoms. For the ligands, the torsion tree and the rotatable/non-rotatable bonds present were also set [58,59]. Calculations involved a docking box with dimensions set at 25 Å × 25 Å × 25 Å and 1 Å spacing, placing the centre of the grid at the centre of mass coordinates of the crystallographic ligand, which was used as a reference. The number of threads was set equal to 8000.

4.7. Model Development

Among the methodologies tested for the establishment of a correlation between the structural properties of, and the biological response to, PPARs, the k-Nearest Neighbours (kNN) classification algorithm emerged as the most appropriate. It is an easily implemented supervised machine-learning technique that is utilised in resolving problems for both continuous and categorical endpoints. The kNN algorithm operates on the principle of identifying the k-number of training data points that are most proximate to a new, unclassified observation based on Euclidean distances, and subsequently assigning the class label that is most frequently represented among the k-nearest neighbours [60].

kNN is considered a 'lazy' learning algorithm since it simply uses the training data for classification instead of building a new model beforehand for new data points and can be used under a read-across framework [61,62]. The optimised value of k, which denotes the number of nearest neighbours to consider, was set at k = 5 and the inverted distance was used as the weighting factor for the nearest k points. An overview of the flowchart conducted in Isalos is presented in Figure 6, where all the significant steps of the analysis were implemented with the software's specified functions.

5. Conclusions

In this study, a dataset of novel thiazole- and oxazole-based compounds was enhanced with binding affinity calculations for the development of a read-across QSAR model to predict their activation of nuclear receptor PPAR. The predictive model attempts to predict the biological activity of small molecules and helps to identify highly potent and selective compounds that act as PPARδ agonists. Using a combination of molecular descriptors that correspond to the different physicochemical, topological, and structural characteristics of a compound, and defining the domain of applicability for new predictions, this model was carefully developed, validated, and documented to adhere to OECD guidelines for QSARs, including provision of a QMRF report. The read-across

model is readily available as a public web application within the Enalos Cloud Platform, a valuable resource for predictive workflows for the assessment of small molecules. A REST API environment is also provided in order to complement the model and offer the users a means to augment its potential.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25105216/s1.

Author Contributions: Conceptualization, A.A. and G.M.; methodology, M.A. and K.D.P.; software, M.A. and K.D.P.; formal analysis, M.A. and K.D.P.; writing—original draft preparation, M.A. and K.D.P.; writing—review and editing, F.D., I.L. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge financial support by the EU H2020 project SCENARIOS (grant agreement No. 101037509). This work was supported by computing time awarded on the Cyclone supercomputer of the High-Performance Computing Facility of The Cyprus Institute under preparatory and production project IDs p114 and pr001017, respectively.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available via Zenodo (https://zenodo.org). The latest version of the curated dataset and the data enrichment attributes can be downloaded free of charge, using the following DOI: https://doi.org/10.5281/zenodo.10566883 The Enalos Asclepios KNIME nodes and workflow used in this study are proprietary to NovaMechanics Ltd. and require a licensing agreement for access.

Conflicts of Interest: Authors M.A., K.D.P. and A.A. are employed by NovaMechanics Ltd., a cheminformatics company.

Abbreviations list

ACC	Accuracy
APD	Applicability domain
API	Application Programming Interfaces
CoMFA	Comparative molecular field analysis and
CoMSIA	Comparative molecular similarity indices analysis
DBD	DNA-binding domain
EC50	Half maximal effective concentration
FN	False negative
FP	False positive
GUI	Graphical user interface
HQSAR	Hologram quantitative structure–activity relationships
ISE	Iterative stochastic elimination
MCC	Matthews correlation coefficient
NIH	National Institutes of Health
NR	Nuclear receptor
OECD	Organisation for Economic Co-operation and Development
PFAS	Per- and polyfluoroalkyl substances
PPAR	Peroxisome proliferator-activated receptor
PPRE	Peroxisome proliferator response element
QMRF	QSAR model reporting format
QSAR	Quantitative structure–activity relationship
REST	Representational state transfer
RXR	Retinoid X receptor
S2k	Kier shape index (fixed length k = 2)
SDF	Structure-data file
SMILES	Simplified molecular input line entry system
TN	True negative
TP	True positive
kNN	k-nearest neighbours
κ	Cohen's kappa

References

1. Weikum, E.R.; Liu, X.; Ortlund, E.A. The Nuclear Receptor Superfamily: A Structural Perspective. *Protein Sci.* **2018**, *27*, 1876–1892. https://doi.org/10.1002/pro.3496.
2. Tyagi, S.; Sharma, S.; Gupta, P.; Saini, A.; Kaushal, C. The Peroxisome Proliferator-Activated Receptor: A Family of Nuclear Receptors Role in Various Diseases. *J. Adv. Pharm. Technol. Res.* **2011**, *2*, 236. https://doi.org/10.4103/2231-4040.90879.
3. Georgiadi, A.; Kersten, S. Mechanisms of Gene Regulation by Fatty Acids. *Adv. Nutr.* **2012**, *3*, 127–134. https://doi.org/10.3945/an.111.001602.
4. Ferré, P. The Biology of Peroxisome Proliferator-Activated Receptors. *Diabetes* **2004**, *53*, S43–S50. https://doi.org/10.2337/diabetes.53.2007.S43.
5. Schoonjans, K.; Martin, G.; Staels, B.; Auwerx, J. Peroxisome Proliterator-Activated Receptors, Orphans with Ligands and Functions. *Curr. Opin. Lipidol.* **1997**, *8*, 159–166. https://doi.org/10.1097/00041433-199706000-00006.
6. Desvergne, B.; Wahli, W. Peroxisome Proliferator-Activated Receptors: Nuclear Control of Metabolism. *Endocr. Rev.* **1999**, *20*, 649–688. https://doi.org/10.1210/edrv.20.5.0380.
7. Palioura, D.; Mellidis, K.; Mouchtouri, E.-T.; Mavroidis, M.; Lazou, A. PPARβ/δ at the Crossroads of Energy Metabolism, Mitochondrial Quality Control and Redox Balance. *J. Biol. Res.-Thessalon.* **2022**, *29*, 12. https://doi.org/10.26262/jbrt.v29i0.8787.
8. Abuhammad, A.; Taha, M.O. QSAR Studies in the Discovery of Novel Type-II Diabetic Therapies. *Expert Opin. Drug Discov.* **2016**, *11*, 197–214. https://doi.org/10.1517/17460441.2016.1118046.
9. Lather, V.; Fernandes, M.X. QSAR Models for Prediction of PPARδ Agonistic Activity of Indanylacetic Acid Derivatives. *QSAR Comb. Sci.* **2009**, *28*, 447–457. https://doi.org/10.1002/qsar.200810092.
10. Maltarollo, V.G.; Homem-de-Mello, P.; Honorio, K.M. Role of Physicochemical Properties in the Activation of Peroxisome Proliferator-Activated Receptor δ. *J. Mol. Model.* **2011**, *17*, 2549–2558. https://doi.org/10.1007/s00894-010-0935-x.
11. Maltarollo, V.G.; Silva, D.C.; Honório, K.M. Advanced QSAR Studies on PPARδ Ligands Related to Metabolic Diseases. *J. Braz. Chem. Soc.* **2012**, *23*, 78–84. https://doi.org/10.1590/S0103-50532012000100013.
12. Wickens, P.; Zhang, C.; Ma, X.; Zhao, Q.; Amatruda, J.; Bullock, W.; Burns, M.; Cantin, L.-D.; Chuang, C.-Y.; Claus, T.; et al. Indanylacetic Acids as PPAR-δ Activator Insulin Sensitizers. *Bioorganic Med. Chem. Lett.* **2007**, *17*, 4369–4373. https://doi.org/10.1016/j.bmcl.2007.03.057.
13. Garcia, T.S.; Silva, D.C.; Gertrudes, J.C.; Maltarollo, V.G.; Honorio, K.M. Molecular Features Related to the Binding Mode of PPAR δ Agonists from QSAR and Docking Analyses. *SAR QSAR Environ. Res.* **2013**, *24*, 157–173. https://doi.org/10.1080/1062936X.2012.751453.
14. Liu, Y.-Y.; Ding, T.-T.; Feng, X.-Y.; Xu, W.-R.; Cheng, X.-C. Virtual Identification of Novel Peroxisome Proliferator-Activated Receptor (PPAR) α/δ Dual Antagonist by 3D-QSAR, Molecule Docking, and Molecule Dynamics Simulation. *J. Biomol. Struct. Dyn.* **2020**, *38*, 4143–4161. https://doi.org/10.1080/07391102.2019.1673211.
15. Maltarollo, V.G.; Araujo, S.C.; Trossini, G.H.G.; Honorio, K.M. Understanding PPAR-δ Affinity and Selectivity Using Hologram Quantitative Structure–Activity Modeling, Molecular Docking and GRID Calculations. *Future Med. Chem.* **2016**, *8*, 1913–1926. https://doi.org/10.4155/fmc-2016-0061.
16. Garcia, T.S.; Honório, K.M. Two-Dimensional Quantitative Structure-Activity Relationship Studies on Bioactive Ligands of Peroxisome Proliferator-Activated Receptor δ. *J. Braz. Chem. Soc.* **2011**, *22*, 65–72. https://doi.org/10.1590/S0103-50532011000100008.
17. Shearer, B.G.; Patel, H.S.; Billin, A.N.; Way, J.M.; Winegar, D.A.; Lambert, M.H.; Xu, R.X.; Leesnitzer, L.M.; Merrihew, R.V.; Huet, S.; et al. Discovery of a Novel Class of PPARδ Partial Agonists. *Bioorganic Med. Chem. Lett.* **2008**, *18*, 5018–5022. https://doi.org/10.1016/j.bmcl.2008.08.011.
18. Da'adoosh, B.; Marcus, D.; Rayan, A.; King, F.; Che, J.; Goldblum, A. Discovering Highly Selective and Diverse PPAR-Delta Agonists by Ligand Based Machine Learning and Structural Modeling. *Sci. Rep.* **2019**, *9*, 1106. https://doi.org/10.1038/s41598-019-38508-8.
19. Kadayat, T.M.; Shrestha, A.; Jeon, Y.H.; An, H.; Kim, J.; Cho, S.J.; Chin, J. Targeting Peroxisome Proliferator-Activated Receptor Delta (PPARδ): A Medicinal Chemistry Perspective. *J. Med. Chem.* **2020**, *63*, 10109–10134. https://doi.org/10.1021/acs.jmedchem.9b01882.
20. OECD. *Reconciling Terminology of the Universe of Per- and Polyfluoroalkyl Substances: Recommendations and Practical Guidance*; OECD Series on Risk Management; No. 61; OECD Publishing: Paris, France, 2021. Available online: https://www.oecd.org/chemicalsafety/portal-perfluorinated-chemicals/terminology-per-and-polyfluoroalkyl-substances.pdf (accessed on 4 December 2023).
21. European Chemicals Agency. *Read-Across Assessment Framework (RAAF)*; European Chemicals Agency: Helsinki, Finland, 2017.
22. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley: Hoboken, NJ, USA, 2000; ISBN 9783527299133.
23. Moreau, G.; Broto, P. The Auto-Correlation of a Topological-Structure—A New Molecular Descriptor. *New J. Chem.* **1980**, *4*, 359–360.
24. Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures–Perception, Auto-Correlation Descriptor and SAR Studies - Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66–70.
25. Burden, F.R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct.-Act. Relatsh.* **1997**, *16*, 309–314. https://doi.org/10.1002/qsar.19970160406.

26. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73. https://doi.org/10.1021/ci00046a002.
27. Kier, L.B. Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relatsh.* **1986**, *5*, 1–7. https://doi.org/10.1002/qsar.19860050102.
28. Xu, H.E.; Lambert, M.H.; Montana, V.G.; Parks, D.J.; Blanchard, S.G.; Brown, P.J.; Sternbach, D.D.; Lehmann, J.M.; Wisely, G.B.; Willson, T.M.; et al. Molecular Recognition of Fatty Acids by Peroxisome Proliferator–Activated Receptors. *Mol. Cell* **1999**, *3*, 397–403. https://doi.org/10.1016/S1097-2765(00)80467-0.
29. Naser, M.Z.; Alavi, A.H. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Archit. Struct. Constr.* **2021**, *3*, 499–517. https://doi.org/10.1007/s44150-021-00015-8.
30. Faulon, J.-L.; Bender, A. *Handbook of Chemoinformatics Algorithms*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2010; ISBN 9781420082999.
31. NovaMechanics Ltd. Enalos + KNIME Nodes. 2017. Available online: http://enalosplus.novamechanics.com/ (accessed on 4 December 2023).
32. Afantitis, A.; Melagraki, G.; Koutentis, P.A.; Sarimveis, H.; Kollias, G. Ligand-Based Virtual Screening Procedure for the Prediction and the Identification of Novel β-Amyloid Aggregation Inhibitors Using Kohonen Maps and Counterpropagation Artificial Neural Networks. *Eur. J. Med. Chem.* **2011**, *46*, 497–508. https://doi.org/10.1016/j.ejmech.2010.11.029.
33. Melagraki, G.; Afantitis, A.; Sarimveis, H.; Igglessi-Markopoulou, O.; Koutentis, P.A.; Kollias, G. In Silico Exploration for Identifying Structure–Activity Relationship of MEK Inhibition and Oral Bioavailability for Isothiazole Derivatives. *Chem. Biol. Drug Des.* **2010**, *76*, 397–406. https://doi.org/10.1111/j.1747-0285.2010.01029.x.
34. Varsou, D.D.; Melagraki, G.; Sarimveis, H.; Afantitis, A. MouseTox: An Online Toxicity Assessment Tool for Small Molecules through Enalos Cloud Platform. *Food Chem. Toxicol.* **2017**, *110*, 83–93. https://doi.org/10.1016/j.fct.2017.09.058.
35. Melagraki, G.; Ntougkos, E.; Rinotas, V.; Papaneophytou, C.; Leonis, G.; Mavromoustakos, T.; Kontopidis, G.; Douni, E.; Afantitis, A.; Kollias, G. Cheminformatics-Aided Discovery of Small-Molecule Protein-Protein Interaction (PPI) Dual Inhibitors of Tumor Necrosis Factor (TNF) and Receptor Activator of NF-KB Ligand (RANKL). *PLoS Comput. Biol.* **2017**, *13*, e1005372. https://doi.org/10.1371/journal.pcbi.1005372.
36. Kamata, S.; Honda, A.; Ishii, I. Current Clinical Trial Status and Future Prospects of PPAR-Targeted Drugs for Treating Nonalcoholic Fatty Liver Disease. *Biomolecules* **2023**, *13*, 1264. https://doi.org/10.3390/biom13081264.
37. Ramaprasad, A.S.E.; Smith, M.T.; McCoy, D.; Hubbard, A.E.; La Merrill, M.A.; Durkin, K.A. Predicting the Binding of Small Molecules to Nuclear Receptors Using Machine Learning. *Brief. Bioinform.* **2022**, *23*, bbac114. https://doi.org/10.1093/bib/bbac114.
38. Varsou, D.-D.; Tsoumanis, A.; Papadiamantis, A.G.; Melagraki, G.; Afantitis, A. Isalos Predictive Analytics Platform: Cheminformatics, Nanoinformatics, and Data Mining Applications. In *Machine Learning and Deep Learning in Computational Toxicology*; Springer: Cham, Switzerland, 2023; pp. 223–242.
39. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD: Paris, France, 2014; ISBN 9789264085442.
40. European Commission, Joint Research Centre (JRC). *JRC QSAR Model Database*; European Commission, Joint Research Centre (JRC) [Dataset] PID; Joint Research Centre (JRC): Brussels, Belgium, 2020. Available online: http://data.europa.eu/89h/E4ef8d13-D743-4524-A6eb-80e18b58cba4 (accessed on 4 December 2023).
41. Epple, R.; Cow, C.; Xie, Y.; Azimioara, M.; Russo, R.; Wang, X.; Wityak, J.; Karanewsky, D.S.; Tuntland, T.; Nguyên-Trân, V.T.B.; et al. Novel Bisaryl Substituted Thiazoles and Oxazoles as Highly Potent and Selective Peroxisome Proliferator-Activated Receptor δ Agonists. *J. Med. Chem.* **2010**, *53*, 77–105. https://doi.org/10.1021/jm9007399.
42. National Center for Biotechnology Information. PubChem Bioassay Record for AID 469785, Source: ChEMBL. Available online: https://pubchem.ncbi.nlm.nih.gov/bioassay/469785 (accessed on 4 December 2023).
43. Singh, A.; Raju, R.; Mrad, M.; Reddell, P.; Münch, G. The Reciprocal EC50 Value as a Convenient Measure of the Potency of a Compound in Bioactivity-Guided Purification of Natural Products. *Fitoterapia* **2020**, *143*, 104598. https://doi.org/10.1016/j.fitote.2020.104598.
44. Nandy, A.; Roy, K.; Saha, A. Exploring Molecular Fingerprints of Selective PPARδ Agonists through Comparative and Validated Chemometric Techniques. *SAR QSAR Environ. Res.* **2015**, *26*, 363–382. https://doi.org/10.1080/1062936X.2015.1039576.
45. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344. https://doi.org/10.1021/ci800038f.
46. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100–108.
47. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining*; Morgan Kaufmann: Burlington, MA, USA, 2017.
48. West, R.M. Best Practice in Statistics: The Use of Log Transformation. *Ann. Clin. Biochem. Int. J. Lab. Med.* **2022**, *59*, 162–165. https://doi.org/10.1177/00045632211050531.
49. Kennard, R.W.; Stone, L.A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137–148. https://doi.org/10.1080/00401706.1969.10490666.
50. Daszykowski, M.; Walczak, B.; Massart, D.L. Representative Subset Selection. *Anal. Chim. Acta* **2002**, *468*, 91–103. https://doi.org/10.1016/S0003-2670(02)00651-7.

51. Tang, S.; Chen, R.; Lin, M.; Lin, Q.; Zhu, Y.; Ding, J.; Hu, H.; Ling, M.; Wu, J. Accelerating AutoDock Vina with GPUs. *Molecules* **2022**, *27*, 3041. https://doi.org/10.3390/molecules27093041.
52. Trott, O.; Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. https://doi.org/10.1002/jcc.21334.
53. Batista, F.A.H.; Trivella, D.B.B.; Bernardes, A.; Gratieri, J.; Oliveira, P.S.L.; Figueira, A.C.M.; Webb, P.; Polikarpov, I. Structural Insights into Human Peroxisome Proliferator Activated Receptor Delta (PPAR-Delta) Selective Ligand Binding. *PLoS ONE* **2012**, *7*, e33643. https://doi.org/10.1371/journal.pone.0033643.
54. Papadopoulou, D.; Drakopoulos, A.; Lagarias, P.; Melagraki, G.; Kollias, G.; Afantitis, A. In Silico Identification and Evaluation of Natural Products as Potential Tumor Necrosis Factor Function Inhibitors Using Advanced Enalos Asclepios KNIME Nodes. *Int. J. Mol. Sci.* **2021**, *22*, 10220. https://doi.org/10.3390/ijms221910220.
55. Eastman, P.; Swails, J.; Chodera, J.D.; McGibbon, R.T.; Zhao, Y.; Beauchamp, K.A.; Wang, L.-P.; Simmonett, A.C.; Harrigan, M.P.; Stern, C.D.; et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659. https://doi.org/10.1371/journal.pcbi.1005659.
56. Case, D.A.; Aktulga, H.M.; Belfon, K.; Ben-Shalom, I.; Brozell, S.R.; Cerutti, D.S.; Cheatham, T.E., III; Cruzeiro, V.W.D.; Darden, T.A.; Duke, R.E.; et al. Amber 2021: Reference Manual; University of California: San Francisco, CA, USA, 2021.
57. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33. https://doi.org/10.1186/1758-2946-3-33.
58. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. https://doi.org/10.1002/jcc.21256.
59. Sanner, M.F. Python: A Programming Language for Software Integration and Development. *J. Mol. Graph. Model.* **1999**, *17*, 57–61.
60. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. https://doi.org/10.1109/TIT.1967.1053964.
61. Varsou, D.; Afantitis, A.; Tsoumanis, A.; Papadiamantis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. Zeta-Potential Read-Across Model Utilizing Nanodescriptors Extracted via the NanoXtract Image Analysis Tool Available on the Enalos Nanoinformatics Cloud Platform. *Small* **2020**, *16*, 1906588. https://doi.org/10.1002/smll.201906588.
62. Varsou, D.D.; Ellis, L.J.A.; Afantitis, A.; Melagraki, G.; Lynch, I. Ecotoxicological Read-across Models for Predicting Acute Toxicity of Freshly Dispersed versus Medium-Aged NMs to Daphnia Magna. *Chemosphere* **2021**, *285*, 131452. https://doi.org/10.1016/j.chemosphere.2021.131452.