International Journal of
Molecular Sciences

MDPI

Article

Machine Learning Assisted Approach for Finding Novel High Activity Agonists of Human Ectopic Olfactory Receptors

Amara Jabeen [1], Claire A. de March [2], Hiroaki Matsunami [2,3,*] and Shoba Ranganathan [1,*]

1 Applied BioSciences, Macquarie University, Sydney, NSW 2109, Australia; amara.jabeen@mq.edu.au
2 Department of Molecular Genetics and Microbiology, Duke University School of Medicine, Durham, NC 27710, USA; claire.de.march@duke.edu
3 Department of Neurobiology, Duke Institute for Brain Sciences, Duke University, Durham, NC 27710, USA
* Correspondence: hiroaki.matsunami@duke.edu (H.M.); shoba.ranganathan@mq.edu.au (S.R.)

Abstract: Olfactory receptors (ORs) constitute the largest superfamily of G protein-coupled receptors (GPCRs). ORs are involved in sensing odorants as well as in other ectopic roles in non-nasal tissues. Matching of an enormous number of the olfactory stimulation repertoire to its counterpart OR through machine learning (ML) will enable understanding of olfactory system, receptor characterization, and exploitation of their therapeutic potential. In the current study, we have selected two broadly tuned ectopic human OR proteins, OR1A1 and OR2W1, for expanding their known chemical space by using molecular descriptors. We present a scheme for selecting the optimal features required to train an ML-based model, based on which we selected the random forest (RF) as the best performer. High activity agonist prediction involved screening five databases comprising ~23 M compounds, using the trained RF classifier. To evaluate the effectiveness of the machine learning based virtual screening and check receptor binding site compatibility, we used docking of the top target ligands to carefully develop receptor model structures. Finally, experimental validation of selected compounds with significant docking scores through in vitro assays revealed two high activity novel agonists for OR1A1 and one for OR2W1.

Keywords: machine learning; random forest; molecular descriptors; virtual ligand screening; olfactory receptor; G protein-coupled receptors; luciferase assay

Citation: Jabeen, A.; de March, C.A.; Matsunami, H.; Ranganathan, S. Machine Learning Assisted Approach for Finding Novel High Activity Agonists of Human Ectopic Olfactory Receptors. *Int. J. Mol. Sci.* **2021**, *22*, 11546. https://doi.org/10.3390/ijms222111546

Academic Editors: Jung Hun Oh and Mingon Kang

Received: 30 September 2021
Accepted: 22 October 2021
Published: 26 October 2021

1. Introduction

G protein-coupled receptors (GPCRs, also known as seven transmembrane or 7TM receptors) represent the largest family of cell surface receptors. The myriad functional diversity of GPCRs has led them to be the largest family of proteins targeted by approved drugs. Primarily the drugs that target the GPCRs are small molecules and peptides [1]. Olfactory receptors (ORs) [2], first reported in 1991, represent the largest sub-group of G protein-coupled receptors (GPCRs) [3].

Initially, ORs were thought to be localised only to nasal tissue and responsible solely for the sense of olfaction, with odorant molecules combinatorically leading to the perception of smell [4]. However, some ORs are expressed in extra-nasal tissues such as mammalian germ cells [5], where they are implicated in different physiological and disease conditions. A recent study has reported the localization of a subset of ORs in various tissues including the brain, prostate, sperm, colon, breast, lungs and kidneys [6]. Functional characterization of these ectopic ORs in different tissues support their roles in cell-cell recognition, migration, proliferation, apoptosis, exocytosis, and novel alternate pathways. Ectopic ORs are also known to be associated with numerous diseases and disorders, including prostate cancer, melanoma, colon cancer, breast tumours, neurodegenerative disorders, obesity and anaemia [7]. Thus, ORs are potential therapeutic targets [8]. Recently, del Marmol et al., reported the experimental structure of an insect olfactory receptor through cryo-electron microscopy [9], which has an inverted topology to animal ORs. However,

there is no experimental structure for any animal OR. We have reviewed the challenges associated with the experimental structure determination for ORs elsewhere in detail [10]. Briefly, the absence of any experimentally determined animal OR structure is attributed to ORs being low abundance, tissue-specific hydrophobic membrane proteins, which are difficult to crystallize. Further, ORs show poor trafficking to the plasma membrane, due to mRNA retention when expressed heterologously in different cell types [11].

Linking the olfactory stimulus repertoire, consisting of more than a trillion, to its counterpart ORs, is a challenging task. To date, only 21% of human ORs have been matched (or deorphanized) with active ligands [10]. OR deorphanization using olfactory sensory neurons (OSNs) can be vastly facilitated by computational approaches. In silico methods coupled with in vitro approaches have proven useful in deorphanizing some ORs [12]. Recent excellent studies using pharmacophore based virtual screening [13] and machine learning (ML) [14] have resulted in expanding the chemical space of a few ORs including the prostate specific G protein receptors (PSGRs: OR51E1 and OR51E2). Although there are a wide variety of ML algorithms, no single algorithm has been capable of solving every problem [15]. ML has now been extensively used to solve various bioinformatics problems [16,17], including GPCR research [18,19]. Recently, support vector machines (SVMs) were used to predict agonists for OR51E1, OR1A1, OR2W1 and MOR256-3 from commonly used odorants [14]. 7/18 predicted ligands for OR1A1, 2/5 for OR2W1, 5/13 for MOR256-3 and 2/4 for OR51E1 were found to be the true ligands when verified through in vitro luciferase assays. The OR51E1 homology model was then used in the reported study to elaborate binding cavity, mutating the residues predicted by molecular docking resulted in receptor response termination in vitro. Homology modelling and molecular docking are thus powerful tools to study receptor ligand interactions in the absence of experimental 3D structure. Many studies have been reported that couple homology modelling, molecular docking and site-directed mutagenesis to elucidate the binding cradle of different ORs with various odorants [20–24]. The mutational dataset for human ORs is now available through an interactive webserver, the Human Olfactory Receptor Mutation Database (hORMdb) [25]. Virtual screening using homology models has also resulted in the discovery of novel ligands. Recently, the human metabolome database was screened against the homology model of OR51E2 and resulted in identification of 24 novel agonists and one antagonist verified experimentally. In a benchmarking study [26], homology models of 19 GPCRs were used for ligand based virtual screening and 10 models showed comparable performance to X-ray structures depicting the applicability of homology models for the identification of novel ligands. We recently compared the performance of four classifiers, based on agonist and non-agonist datasets for OR1G1, with the naïve Bayes classifier performing better than SVM, random forest [27] and neural networks (NN) for agonist prediction [28].

Tunyasuvunakool et al., have reported highly accurate protein structure prediction for the entire human proteome, including ORs [29]. However, the AlphaFold models generated in this study need a lot of adjustment in the orientation of the transmembrane helices (Jabeen and Ranganathan, unpublished data), to recover the OR binding sites that were validated by published mutational studies [20–24]. Instead, we recently showed that using a biophysical approach, Bio-GATS, for template selection generates an excellent homology model for OR1A1 [30].

In the current study, we have focused our attention on two ectopic ORs with broad ligand spectrum, OR1A1 and OR2W1. OR1A1, reported in gut enterochromaffin cells, was implicated in serotonin release [31]. Recurrent mutations in OR1A1 were identified in lung adenocarcinoma [32]. Further, OR1A1 was detected in HepG2 liver cells [33], and implicated in hepatic triglyceride metabolism modulation. OR2W1 has recurrent mutations reported in small lung cancer [34]. Both receptors, along with a few other ORs, are also proposed to have roles in fatigue attenuation [35].

The aim of our study was to expand the chemical space for these two ORs with potential clinical importance, by predicting and experimentally testing novel agonists.

Therefore, we have developed an ML-based workflow (Figure 1) to predict agonists for the two ORs by scanning the huge chemical compounds databases available online. We selected three methods, RF, SVM and NB, based on their performance for OR1G1 [28]. We filtered the predicted compounds using knowledge-based homology models of the two receptors, based on the best Bio-GATS template [30]. From the shortlisted predictions, we validated some of the randomly selected predicted compounds using an in vitro functional assay.

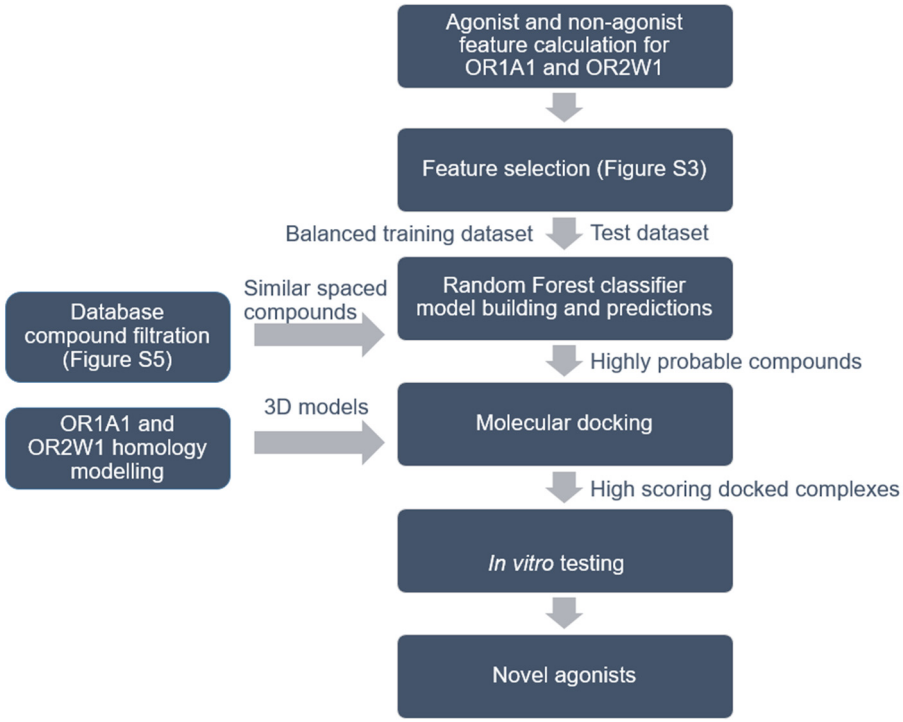

Figure 1. Workflow for agonist identification based on machine learning, molecular docking and in vitro testing. Details of feature selection is shown in Supplementary Figure S3 and the process for database filtration is shown in Supplementary Figure S5. Compounds from five different databases were downloaded and classified as agonists or non-agonists for OR1A1 and OR2W1.

2. Methods

2.1. Ligand Dataset Collection for the Receptors and Chemical Descriptors Calculation

The experimentally tested compounds against the OR1A1 and OR2W1 were retrieved from the literature. Among the 365 compounds tested against OR1A1, 51 are agonists, 263 are non-agonists, and 51 have conflicting information (Supplementary Table S1). While for the OR2W1, 292 compounds have been experimentally tested, of which 64 are agonists, 198 are non-agonists, seven are antagonists, and 22 have conflicting information (Supplementary Table S2). As the antagonists are few in number and the compounds with conflicting information could not be classified uniquely, these two categories were not considered for further analysis. Molecular descriptors for the agonists and non-agonists of both the receptors were calculated using an open-source and free software, Mordred v1.2.0 [36]. Overall, 1443 and 1505 2D and 3D molecular descriptors were calculated for OR1A1 and OR2W1, respectively.

2.2. Data Pre-Processing, Feature Selection and Class Balancing

Since a large number of descriptors were calculated for both receptors, the pre-processing and feature selection techniques were employed to get an optimal number of features. Initially, all the features with any missing value were eliminated then a near zero filter was applied to exclude features having low variance. A correlation filter was

applied to reduce the collinearity among the descriptors [37]. The threshold value for the correlation coefficient (r) was set to 0.95 as previously setup for OR1A1 and OR2W1 features [14] and for the moth odorant receptor [38]. Afterwards, three different methods were used for feature selection including a wrapper method: recursive feature elimination, a filter method: Gini, and an embedded method: random forest feature selection, were applied for selection of relevant subset of molecular descriptors. The dataset, comprising known agonists and non-agonists with the selected features, was split into 80% training and 20% test sets using random sampling. Pre-processing, feature selection and data splitting were carried out using the R programming language [39].

Since the two classes (agonists and non-agonists) are imbalanced (one agonist: ~five non-agonists for OR1A1 and one agonist: ~four non-agonists for OR2W1), we used the synthetic minority over-sampling technique (SMOTE) [40] embedded as a node in Knime 3.6.0 [41] on the training-set to have balanced datasets. The 5th nearest neighbour was considered for synthetic sampling.

2.3. Classifiers

We generated RF, SVM, and NB classifier models using the R programming language. 10-fold cross validation on the training dataset was used as a resampling method for each classifier. The CV was repeated three times to avoid any bias during the creation of CV data splits. A brief description of each model is provided below.

2.3.1. Random Forest

The RF approach utilizes the decision trees and creates various models through random partitioning. The final output is based on majority voting [42]. In our model, the number of trees and variables randomly sampled as candidates at each split were hyper-parametrized to obtain the optimal RF model. The final RF classification model was based on 300 trees with 5 variables randomly sampled as candidates at each split for OR1A1 and 3 variables randomly sampled as candidates at each split for OR2W1.

2.3.2. Support Vector Machine

SVM is based on calculating the maximal marginal hyperplane to separate positives from the negatives [43]. In the current study, the SVM classifier was built using radial basis kernel function. The two parameters that were hyper-parametrized are sigma and cost. Sigma was held constant at the value of 0.2326189 and the accuracy metric was used to select the optimal model using the largest value. The final value for sigma was 0.2326189 and 0.5 for cost.

2.3.3. Naïve Bayes

NB is the commonly used, simple and computationally less expensive ML method [44]. NB is based on Bayes rule as mentioned in Equation (1):

$$P(y|x) = P(y)P(x|y)/P(x) \quad (1)$$

where y represents the class and x represents the data points.

The NB classifier assumes that all features are independent of each other so P(y) P(x | y) can be re-written as Equation (2):

$$P\left(y_j\right)P(x|y_j) = P\left(y_j\right)\prod_{i=1}^{n}P(x_i|y_j) \quad (2)$$

where $P(y_j)$ is the prior of the classes and $P(x_i \mid y_j)$ is the distribution for one feature and one class. The Gaussian distribution was used in this study for NB classification model.

2.4. Model Validation

Prediction performance of each classifier was assessed by the test set for each OR and by 10-fold cross validation of the training data. 20% of the dataset was reserved as a test set and was not used for training the model. Therefore, this test set was unseen for the classifier and can be considered as a blind test set, as no suitable external validation set is available. Two statistical tests namely, p-value and Cohen's kappa coefficient (κ) were also used to evaluate the models. Further, accuracy, sensitivity, specificity and the F1 score measures were used to evaluate the classifiers. The values for accuracy, sensitivity, and specificity were calculated using Equations (3)–(5):

$$\text{Accuracy} = {}^{\text{TP + TN}}/_{\text{TP + TN + FP + FN}} \quad (3)$$

$$\text{Sensitivity/recall} = {}^{\text{TP}}/_{\text{TP + FN}} \quad (4)$$

$$\text{Specificity} = {}^{\text{TN}}/_{\text{TN + FP}} \quad (5)$$

where TP, TN, FP and FN refer to true positive, true negative, false positive, and false negative.

The F1 score is defined in Equation (6) as:

$$\text{F1} = {}^{\text{Precision*recall}}/_{\text{Precision + recall}} \quad (6)$$

where precision is calculated as Equation (7) and recall is calculated as Equation (4):

$$\text{Precision} = {}^{\text{TP}}/_{\text{TP + FP.}} \quad (7)$$

2.5. Filtration of Compounds for Virtual Screening on the Basis of Chemical Similarity

For virtual screening using the built classifiers, we downloaded the compounds from ZINC [45], human metabolome database (HMDB) [46], ChEBI [47], Cancer Odor Database (COD) [48], OdorDB [49]. Those compounds that have already been experimentally tested against OR1A1 and OR2W1were filtered out from the list. We used PubChem fingerprints coupled with Tanimoto index for scanning similar spaced compounds from the above-mentioned databases. Only compounds with Tanimoto index of at least 85% were selected for screening. The final list of selected compounds was then evaluated as potential agonists of OR1A1 and OR2W1, using the RF classifier.

2.6. OR1A1 and OR2W1 Homology Modelling

The 3D models of OR1A1 (UniProtID: Q9P1Q5) and OR2W1 (UniProtID: Q9Y3N9) were built using homology modelling approach as described previously [50]. The X-ray crystal structure of bovine rhodopsin (PDB ID: 1U19) [51] was used as the template for homology modelling of human OR1A1 and OR2W1. Briefly, the sequences of OR1A1 and OR2W1 were aligned with bovine rhodopsin, based on conserved GPCR motifs (Supplementary Figure S1). The predicted transmembrane domains in both receptors were based on the GRoSS sequence alignment of all known GPCRs sequences [52], as implemented by Bio-GATS [30]. Homology modelling was performed using MODELLER 9.18 [53]. The resulting models were assessed using the Modeller objective function, which reflects the quality of the model and the presence of a disulphide bond between Cys97 and Cys179. The selected models were also evaluated using the Ramachandran plot and favoured rotamers, on the Molprobity webserver [54]. The side chains of the built models were refined using SCWRL4 [55] to improve rotamer geometry.

2.7. Molecular Docking of Highly Probable Predicted Compounds

The compounds having similar space as agonists of OR1A1 and OR2W1 were classified as agonists and non-agonists through the trained RF classifier. Compounds with prediction probability of 1.0 for being agonists alone were considered for molecular dock-

ing. The binding pockets of both the receptors were predicted using ICMPocketFinder embedded in ICM package [56]. The binding pockets were selected based on site-directed mutagenesis data of different ORs. Induced fit docking was then performed using ICM. Ten conformations were generated for each predicted ligand and the control molecule. The docking effort was set to 3, as the developers of ICM benchmark the accuracy at this effort level. The conformation with the lowest ICM-score was selected for binding analysis.

2.8. Cell Culture

Hana3A cells [57] were maintained in minimal essential medium [34] containing 10% FBS (vol/vol) with penicillin-streptomycin and amphotericin B (1/200 each vol/vol) at 37 °C and 5% CO_2. Hana3A cells are derived from HEK293T [57] and are optimized for OR studies as shown in several studies [58–66], compared to earlier OR expression in Sf9 insect cells in Gat et al. [67], and in *Xenopus* oocytes, COS-7, PC12h and CHO-K1 cells in Katada et al. [68].

2.9. Dual-Glo Luciferase Reporter Gene Assay

The Dual-Glo luciferase assay system (Promega, Madison, WI, USA) was used to evaluate the functionality of wild-type OR1A1 and OR2W1 in an in vitro system [69,70]. The open reading frames of ORs were amplified using Phusion polymerase (Thermo Fisher Scientific, Waltham, MA, USA). Amplified fragments were cloned into pCI expression vector (Promega, Madison, WI, USA) containing the sequence encoding the first 20 amino acids of human rhodopsin (Rho-tag) at N-terminal [71]. Hana3A cells have been cultured and plated the day before transfection with 6 mL at 1/10 of a 100% confluence 100 mm plate into 96-well plates coated with poly D-lysine. After overnight incubation, the required genes were transfected using, for each plate, 5 ng SV40-RL, 10 ng CRE-Luc, 5 ng human RTP1S [72], 2.5 ng M3 receptor [73] and 5 ng of receptor (OR1A1, OR2W1 or empty vector Rho-pCI) plasmid. After around 18 h of transfection, cells were stimulated during 3.5 h by 25 µL of odorant diluted in CD293 + 1% glutamine + 30 µM $CuCl_2$. Odorants were obtained from Sigma Aldrich (St. Louis, MO, USA) and diluted at 1 M concentration in DMSO as stock solutions. Dose response curves were determined with concentrations of 0, 1, 3.16, 10, 31.6, 100, and 316 µM obtained by dilution of the DMSO stock solution in CD293 + 1% glutamine + 30 µM $CuCl_2$. The luminescence of Firefly (Luc) and Renilla (Rluc) luciferase, were then sequentially monitored by injecting the corresponding substrate following the supplier's protocol. The activity in each well was normalized as (Luc-400)/(Rluc-400). The response of the receptor was also normalized to its basal activity as (NL_X/NL_0)-1 where NL_0 is the normalized luminescence value at 0 µM of odorant and NL_X the value at X µM. The cell response upon odorant stimulation was attributed to an OR if the empty vector control showed no response, assuring that the cell response is not due to other parameters than the presence of the OR at the cell surface. Raw results were first analyzed with Excel (Microsoft Corporation, Albuquerque, NM, USA) and dose response curves, max efficacy and EC50 have been determined with GraphPrism 6 software (GraphPad Software, La Jolla, CA, USA). Areas under the curves (AUC) were calculated in Excel by summing all the OR responses at different concentrations for each odorant.

2.10. Cell Surface Expression Evaluation by Flow Cytometry

The cell surface expression of the studied ORs has been evaluated by flow cytometry, which provides more quantitative cell surface expression data than conventional immunostaining, using the following protocol [74]. Human embryonic kidney variant 293T (HEK293T) cells were grown to confluency, resuspended and seeded onto 35 mm plates at 25% confluency. The cells were cultured overnight. The OR, RTP1s and GFP were transfected using Lipofectamine 2000. After 18–24 h, the cells were resuspended by cell stripper and then kept in 5 mL round bottom polystyrene (PS) tubes (Falcon 2052, Corning, Corning, NY, USA) on ice. The cells were spun down at 4 °C and resuspended in phosphate-buffered saline (PBS) containing 15 mM NaN_3, and 2% foetal bovine serum

(FBS) to wash the cell stripper. They were incubated in ice with primary antibody (mouse anti-Rho4D2 [75]) and then washed, and stained with phycoerythrin (PE)-conjugated donkey anti-mouse antibody (Jackson Immunologicals, West Grove, PA, USA) in the dark. To stain dead cells, 7-aminoactinomycin D (Calbiochem, MilliporeSigma, Burlington, MA, USA) was added. The cells were analyzed using FACS (BD FACSCanto II, Bio-Rad Laboratories, Hercules, CA, USA) with gating, allowing for GFP positive, single, spherical, viable cells, and the measured PE fluorescence intensities were analyzed and visualized using Flowjo v10.0.8 [76]. We also added Olfr539, which is robustly expressed on the cell surface, and Olfr541, which shows no detectable cell surface expression, as positive and negative controls of OR cell surface expression [74], respectively.

3. Results

3.1. Chemical Diversity Analysis

The chemical space for OR1A1 and OR2W1 is highly diverse and comprised of aldehydes, alcohols and esters among others (Supplementary Figure S2). The diversified nature of the collected experimentally known compounds against the two receptors were verified using principal component analysis (PCA). The first two principal components were plotted and are clearly indicative of the diversified chemical space for the two receptors (Figure 2A). Splitting the experimentally tested data into 80% training and 20% test datasets showed considerable overlap between the two sets (Figure 2B) indicating that the classifiers are being validated on the basis of similar spaced compounds.

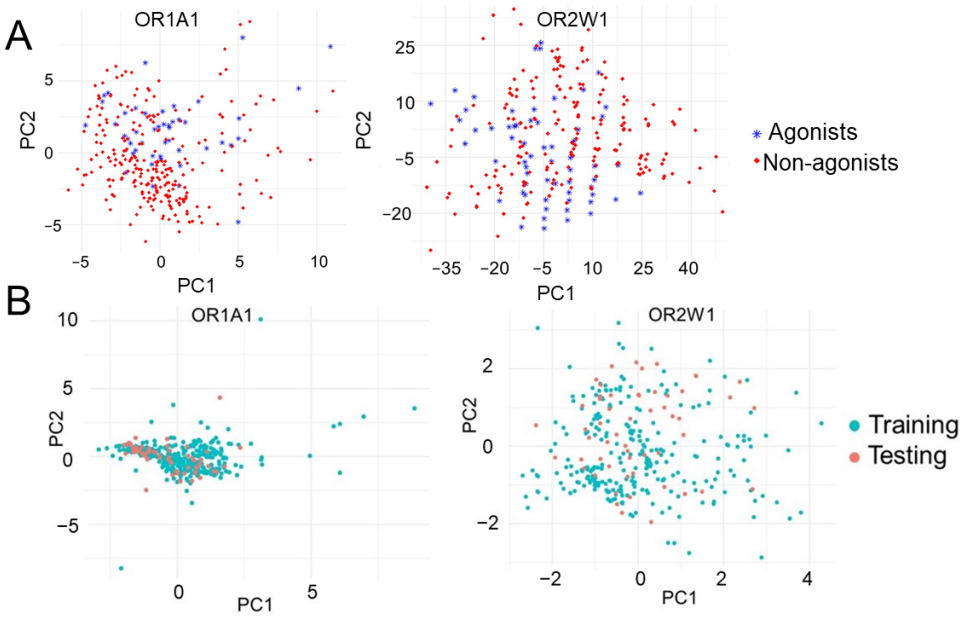

Figure 2. First two principal components of (**A**) OR1A1 and OR2W1 agonists and non-agonists showing the diversified nature of agonists for both ORs and (**B**) training and testing datasets showing the considerable overlap between training and testing set.

3.2. Feature Selection and Performance of the Classifiers

The strategy for identifying novel agonists for OR1A1 and OR2W1 required training of the classifiers with the chemical descriptors (or features) of the known agonists and non-agonists. Mordred calculated 1505 for OR1A1 and 1443 features for OR2W1. Feature selection is an important step for efficient dimensionality reduction to gain quality classifiers [77]. Broadly, three methods for feature selection in use are filter methods, wrapper methods and embedded methods. However, it is hard to determine any one specific method as the most accurate [78]. As the accuracy of machine learning approaches is highly dependent on the selected features [79], we used a combination of data-driven filter methods and other feature selection methods (recursive feature elimination, Gini index,

and random forest feature selection) to select the optimal features. Initially, 1096 features for OR1A1 and 1034 features for OR2W1 respectively, were eliminated, using filter methods (Supplementary Figure S3). Subsequently, three different methods for feature selection were applied to each selected feature dataset. A wrapper method: recursive feature elimination, a filter method: Gini index, and an embedded method: random forest feature selection were used. The top 20 features from each approach were compared. For OR1A1, 13 consensus features were obtained by the three methods, namely VR2_A, AATSC0v, ATSC6d, ATSC7d, ATSC8c, BCUTm.1l, C3SP2, EState_VSA4, JGI5, JGI6, JGI7, PEOE_VSA6, and SdssC. For OR2W1, there were only two consensus features: JGI5 and JGI6, among the three different methods of feature selection. Therefore, the features predicted by three different methods were iteratively applied to finally select the five features for OR2W1 for training the model (Supplementary Figure S4). This selection of the features was based on the combination of features giving the maximum accuracy. The detailed description of each selected feature is mentioned in Supplementary Table S3. The selected combination of five features gave the maximum accuracy for all classification models. RF, SVM and NB classifiers were hyper-parameterized and trained using the selected features. The SVM classifier showed comparable performance to RF classifier for the OR1A1 dataset, while for the OR2W1 dataset, the RF classifier outperforms the other two classifiers (Figure 3).

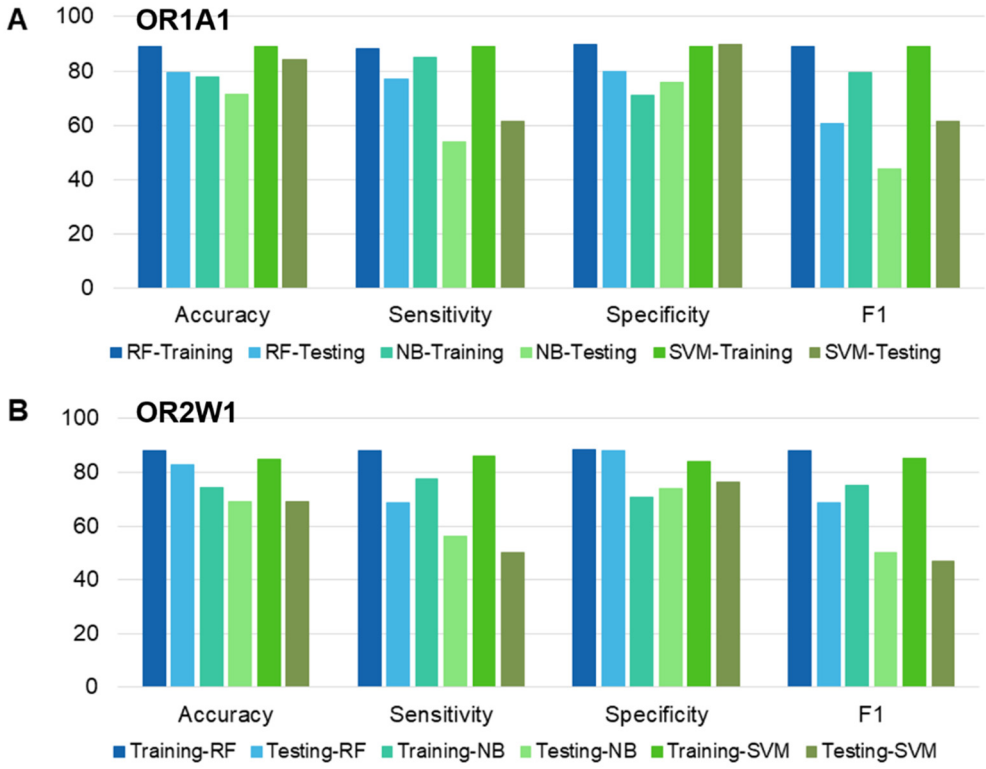

Figure 3. Performance comparison of different classifiers based on accuracy, sensitivity, specificity, and F1 score for (**A**) OR1A1 (**B**) OR2W1. Training and testing datasets for each OR have been compared, using RF: random forest, NB: Naïve Bayes and SVM: Support vector machine.

The performance of each classifier was validated by 10-fold CV, and testing data. Additionally, the classifiers (or models) were evaluated on the basis of accuracy, specificity, sensitivity, and F1 metrics. Additionally, we carried out two statistical tests to compare the performance of the classifiers (Supplementary Table S4). The *p*-value of each classifier was significant ($<2 \times 10^{-16}$) for the training dataset for both ORs while for testing dataset, the RF classifier was close second to the SVM classifier for OR1A1 and scored best for OR2W1-test set. The kappa values for RF classifiers outperformed the other two classifiers for both training and testing sets of OR1A1 and of OR2W1. Based on the classifier evaluation scores,

RF was selected as the predictive model for screening the compounds downloaded from five different databases.

3.3. Putative Ligand Screening through Machine Learning Based Classification

We downloaded 22,938,816 compounds from five different online databases (ZINC, HMDB, ChEBI, COD, and OdorDB). Since the classifiers were computationally trained on specific molecular descriptors for agonists and non-agonists of OR1A1 and OR2W1, they can only classify compounds belonging to similar chemical space. To identify similar chemically spaced compounds, we applied a Tanimoto index value of 0.85 to the entire downloaded dataset. The compound search space was thus reduced to 35,415 for OR1A1 and 27,127 for OR2W1 (Supplementary Figure S5). The trained RF classifier ranked these chemically similar compounds as agonists and non-agonists, separately, for each receptor. Based on classification probabilities, compounds were ranked as agonists. As the sensitivity and F1 score values of the RF classifier for testing data were below 0.80, we set up a threshold value of prediction probability as 1.0, in order to limit false positive predictions. Generally, the class membership probability has a threshold value of 0.5 [80]. With this threshold value, 67 compounds (three from ChEBI, one from COD, two from HMDB, three from OdorDB, and 58 from ZINC) were predicted as OR1A1 agonists. Independently, 83 compounds (three from ChEBI, none from COD and OdorDB, and 80 from ZINC) were predicted as OR2W1 agonists.

3.4. Homology Model and Molecular Docking Analysis

To identify the top candidates for experimental validation, we conducted induced-fit docking of the highly probable compounds based on 3D homology models of OR1A1 and OR2W1. We built the homology models of OR2W1 (UniProtID: Q9Y3N9) and OR1A1 (UniProtID: Q9P1Q5) using the approach described previously [50], with the X-ray crystallographic structure of bovine rhodopsin (PDB ID: 1U19) [51] as a template, and Bio-GATS TM alignments [30]. The experimental structure of an insect OR has recently been resolved at 3.3 Å resolution [9] with an inverted topology to human ORs. The sequence identity (SI) and query coverage (QC) between the experimentally determined structure (PDBID: 7LID) and our target sequences is extremely low (SI for 7LID-OR1A1 pair: 7.0%, QC: 50% and SI for 7LID-OR2W1 pair: 7.16%, QC: 58%). Due to the low resolution of the insect OR structure and extremely low sequence identity and low query coverage with the human ORs under investigation, we did not proceed with the 7LID template. The structures for predicted ligands and benzophenone (experimentally known ligand for both receptors) were downloaded from PubChem [81] and optimized using the ICM package [56] and docked to the homology models for OR1A1 and OR2W1. The binding site for each receptor was selected as a consensus site considering the experimental mutagenesis sites for the other ORs (OR1A1, OR1A2, OR1G1, OR2AG1, OR2M3, OR5AN1, OR7D4 and OR51E2), based on the alignment of OR1A1, OR2W1 and the OR sequences with available mutagenesis data shown in Supplementary Figure S6. Positions G108$^{3.35}$, S109$^{3.36}$, C112$^{3.39}$, N155$^{4.56}$, I206$^{5.46}$ and Y252$^{6.48}$ of the predicted binding pocket are consistent with the available OR mutagenesis data for OR2W1 (numbering in superscripts are the respective Ballesteros-Weinstein residue numbers [82]). Also, positions G108$^{3.35}$, N109$^{3.36}$, S112$^{3.39}$, I205$^{5.46}$, Y251$^{6.48}$, Y258$^{6.55}$ and T277$^{7.42}$ of the OR1A1 binding pocket are consistent with mutagenesis data available for OR1A1. All predicted ligands and benzophenone were docked to their respective receptor model, with 100 conformations were generated for each predicted ligand and the control (benzophenone). The ICM docking score of benzophenone was used as a threshold to select the docked compounds with equivalent scores. The conformation that fits within the binding pocket and has an ICM docking score around that of benzophenone, was selected. This strategy reduced the predictions to 23 compounds for OR1A1 and 10 compounds for OR2W1, with ICM scores nearest to that of benzophenone (Tables 1 and 2). Of these, four compounds for OR1A1 and two compounds for

OR2W1 that were not reported in the Bushdid et al. [14] study were randomly selected, and experimentally tested using functional in vitro assays.

Table 1. Highly probable OR1A1 agonists based on docking scores. Control in italics; experimentally test compounds underlined.

PubChem_CID	Compound Name	Database	Chemical Nature	ICM Docking Score
3102	*Benzophenone*	*Control*	*Ketone*	*−12.8345*
10465547	[(*Z*)-Pent-3-enyl] 2-aminobenzoate	ZINC	Heterocyclic compound	−24.058033
70545042	Prop-2-enyl 3-iodobenzoate	ZINC	Heterocyclic compound	−20.113147
56806459	2-(4-Methylphenoxy)pentan-3-one	ZINC	Ketone	−19.297316
84603836	(3-Fluorophenyl)methyl 4-methylpentanoate	ZINC	Ester	−18.084389
101977	D-citronellol	ChEBI, HMDB	Terpene	−18.009
56828593	4-(3-Fluorophenyl)-3-methyl-4-oxobutanenitrile	ZINC	Heterocyclic compound	−17.436507
17973047	4-(1-Methylcyclopropyl)phenol	ZINC	Heterocyclic compound	−17.286264
30842889	Prop-2-enyl 2-(2,4-difluorophenyl)acetate	ZINC	Ester	−17.136729
22048986	6-Chloro-1-(3-fluorophenyl)hexan-1-one	ZINC	Ketone	−17.053948
11470552	Ethyl 2-(3-bromophenyl)acetate	ZINC	Ester	−16.812695
45085600	(5*S*)-5,6-dimethylhept-6-en-2-one	ZINC	Ketone	−16.733556
5352782	3-[(*E*)-But-1-enyl]pyridine	ZINC	Heterocyclic compound	−16.708306
7021479	Methyl 2-(2-methylphenyl)acetate	ZINC	Ester	−16.649985
59382573	(3-Methoxyphenyl)methyl butanoate	ZINC	Ester	−16.572913
84177	Ethyl 2-(4-chlorophenyl)acetate	ZINC	Ester	−16.278326
6368521	1-[(*E*)-2-Chloroethenyl]-4-methoxybenzene	ZINC	Heterocyclic compound	−16.266073
78901972	(3-Fluorophenyl)methyl 2-propylsulfanylacetate	ZINC	Ester	−16.165344
8842	Citronellol	OdorDB, ChEBI	Terpene	−14.6017
22311	Dipentene	OdorDB, COD	Terpene	−13.9064
1318	1,10-Phenanthroline	ChEBI	Heterocyclic compound	−13.7334
24473	Dihydrocarvone	OdorDB	Ketone	−11.3054
131752167	2,10-Bisaboladiene-1,4-diol	HMDB	Alcohol	−10.6224
78236	4-Nonanone	HMDB	Ketone	−10.5291

Table 2. Highly probable OR2W1 agonists based on docking scores. Control in italics; experimentally test compounds underlined.

PubChem_CID	Compound Name	Database	Chemical Nature	ICM Docking Score
3102	*Benzophenone*	*Control*	*Ketone*	*−11.8875*
13433021	4-Methyl-2-m-tolylpyridine	ZINC	Heterocyclic compound	−14.515694
2733871	2,4-Dimethyl-1-phenylpyrrole	ZINC	Heterocyclic compound	−14.150921
249799	1-Butoxy-4-phenylbenzene	ZINC	Heterocyclic compound	−13.419691
22562335	Methyl 3-(4-ethoxyphenyl)prop-2-ynoate	ZINC	Heterocyclic compound	−12.639271
16530415	(2,3,4,5,6-Pentafluorophenyl)methyl 2-hydroxy-3-methylbenzoate	ZINC	Heterocyclic compound	−11.02916

Table 2. *Cont.*

PubChem_CID	Compound Name	Database	Chemical Nature	ICM Docking Score
3847415	1-Ethenyl-4-[4-(4-ethenylphenoxy)butoxy]benzene	ZINC	Heterocyclic compound	−8.767672
12252872	Ethyl 4-hydroxy-3-prop-2-enylbenzoate	ZINC	Heterocyclic compound	−8.68684
231770	1,3-bis(4-Bromophenyl)prop-2-en-1-one	ZINC	Heterocyclic compound	−8.668044
7129	2-Ethoxynaphthalene	ZINC	Ether	−8.208685
60008260	Ethyl 2-amino-5-cyanobenzoate	ZINC	Heterocyclic compound	−8.011839

3.5. In Vitro Testing of Predicted Agonists Using Luciferase Assay

We tested the response of OR1A1 and OR2W1 to different concentrations of the candidate molecules using an in vitro luciferase assay (Figure 4A,B), following the verification of cell surface expression of these two ORs by flow cytometry (see Methods). We used Olfr539 and Olfr541 as positive and negative controls of OR cell surface expression, respectively [74]. In comparison to our controls, both ORs are relatively well trafficked to the cell surface (Supplementary Figure S7 and Supplementary Data File).

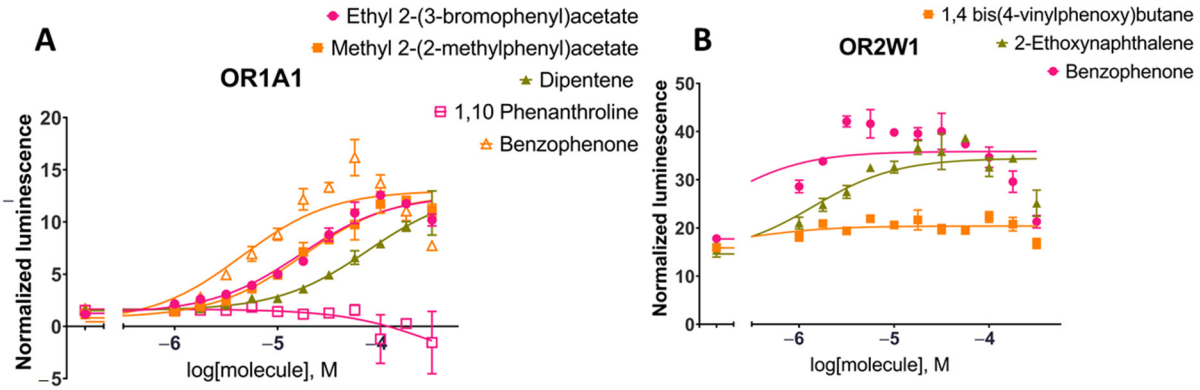

Figure 4. Dose-response curves for tested compounds against (**A**) OR1A1 and (**B**) OR2W1 for the luciferase assay (see Methods). The tested compounds were randomly selected from short-listed compounds after machine learning and molecular docking to evaluate the random forest model predictions. Cell surface expression of these two ORs from flow cytometry are shown in Figure S7.

Benzophenone was added to the set of tested molecules as a positive control for OR1A1 and OR2W1 activation as it is an agonist for both ORs [83]. Ethyl 2-(3-bromophenyl)acetate, methyl 2-(2-methylphenyl)acetate and 1,10-phenanthroline stimulations were tested for OR1A1 activation in dose-responses (Figure 4A). Both ethyl 2-(3-bromophenyl)acetate and methyl 2-(2-methylphenyl)acetate were able to activate OR1A1 and showed similar dose-response curves and EC_{50} values (EC_{50} (ethyl 2-(3-bromophenyl)acetate) = 11.5–24.4 μM; EC_{50} (methyl 2-(2-methylphenyl)acetate) = 12.2–25.0 μM, 95% CI). Dipentene (racemic limonene) was a weak activator, while 1,10-phenanthroline did not activate OR1A1 and was considered a non-agonist. 1,4-bis(4-Vinylphenoxy)butane and 2-ethoxynaphthalene stimulation were tested for OR2W1 (Figure 4B). 2-Ethoxynaphthalene activated OR2W1 in a dose-response manner with an EC_{50} of 6.59–3.05 μM (95% CI). 1,4-bis(4-Vinylphenoxy)butane was identified as a non-agonist of OR2W1.

3.6. Binding Mode of the Tested Ligands

The two activating agonists for OR1A1 and the activating agonist for OR2W1 were re-docked in their respective receptor's binding pocket, to analyse the receptor binding residues and the binding mode for these novel agonists. Interacting residues of the individual receptors are shown in Supplementary Figure S8. Ethyl 2-(3-bromophenyl)acetate

has a single hydrogen bond to S112$^{3.41}$ within the OR1A1 binding pocket, while the rest of the interactions are hydrophobic. Methyl 2-(2-methylphenyl)acetate shows hydrophobic interactions with OR1A1. 2-Ethoxynaphthalene also shows predominantly hydrophobic interactions with OR2W1.

4. Discussion

In the current study, we report the ML-based virtual screening workflow for agonist identification of two broadly tuned ectopic ORs: OR1A1 and OR2W1. Both receptors have physiological and pathophysiological implications. In an earlier study, SVM was applied to OR1A1 and OR2W1 to screen a test set of 258 compounds and resulted in the identification of novel agonists for both receptors, with a hit rate of 39 to 40% for these ORs [14]. In this present work, we further build hyper-parameterized RF and NB models along with SVM, selected the best performing RF model and achieved a hit rate of 75% for OR1A1 and 50% for OR2W1, respectively. The dataset, comprised of experimentally known agonists and non-agonists for both ORs, is highly imbalanced. Therefore, careful selection of features as well as the classification model is necessary. Further, the dataset for OR2W1 is extremely diverse, show high variance and low biased as compared to OR1A1 (Figure 2) which indicates that OR2W1 dataset is more prone to overfitting. The right balance between variance and bias is desired, to have an optimal ML model [84]. Therefore, we carefully selected the features to train the classifiers by applying filter-based, wrapper, and embedded methods. We have selected five features to suit the size of the data sets. Moreover, all three models were hyper-parameterized to avoid any overfitting that might occur due to decreased bias. As a result, we obtained models showing reasonably good classification accuracy, both on training and testing data, as shown in Figure 3. We then compared the performance of three well-established ML classifiers based on accuracy, sensitivity, specificity and F1 score. The hyper-parameterized RF classifier outperforms the other classifiers, SVM and NB, for both receptors with all values exceeding 0.85 for training data and thus capable of distinguishing agonists from non-agonists. However, sensitivity and F1 score for testing data was below 0.80. Therefore, we set up a threshold value of prediction probability to 1.0, in order to avoid excessive false positives. Generally, the class membership probability has a threshold value of only 0.5 [80], indicating that our selected threshold is much more stringent. Based on performance, we selected the RF classifier and used it to screen the huge test set of 22,938,816 compounds from five compound databases. Scanning similar spaced test set compounds with the RF classifier yielded 67 and 83 compounds ranked as agonists for OR1A1 and OR2W1, respectively. We docked these compounds into the binding pocket of the respective receptor structural models, to further validate our predictions. Compounds showing good binding affinity in docking runs were shortlisted and randomly selected for experimental testing through luciferase assays, to evaluate the validity of our approach. Of the four compounds tested for their responsiveness against OR1A1, ethyl 2-(3-bromophenyl)acetate and methyl 2-(2-methylphenyl)acetate are identified as high activity novel agonists for OR1A1. We also identified dipentene (racemic limonene), as an activating ligand for OR1A1. The two isomers of limonene that are *(S)*-(-)-limonene and (*R*)-limonene have already been identified as agonists for OR1A1 in multiple studies (Supplementary Table S1). 1,10-Phenanthroline did not activate the receptor and therefore, was regarded as a non-agonist for OR1A1. We also tested two compounds for their activity against OR2W1. 2-Ethoxynaphthalene turned to be the agonist, while 1,4 bis(4-vinylphenoxy) butane was regarded as a non-agonist. Our results are consistent with the observations of Bushdid et al. [14], where agonist and non-agonist spaces could not easily be differentiated on the basis of simple chemical descriptors. Evaluating the experimental affinity between ORs and odorant molecules could be helpful to establish a performant predictive model. Unfortunately, there are no such data available today for mammalian ORs.

We further compared the potency of identified agonists with the previously identified potent agonists for OR1A1 and OR2W1 as illustrated in the supporting information of

Bushdid et al. [14]. Agonists identified in this study show a triggered response of >80% of the control (benzophenone) for OR1A1, while the agonist identified for OR2W1 shows 90% triggered response of the same control (benzophenone) (Figure 5). Although (−)-carvone is a stronger control for OR1A1 as compared to benzophenone, the novel agonists identified in this study show comparable triggered response to (−)-carvone (>75%). 2-ethoxynephthalene, identified as an agonist for OR2W1 in the current study is two times more potent than the strongest agonist identified for OR2W1 in the Bushdid et al. [14] study. Identification of highly potent agonists demonstrates the efficacy of our classification model.

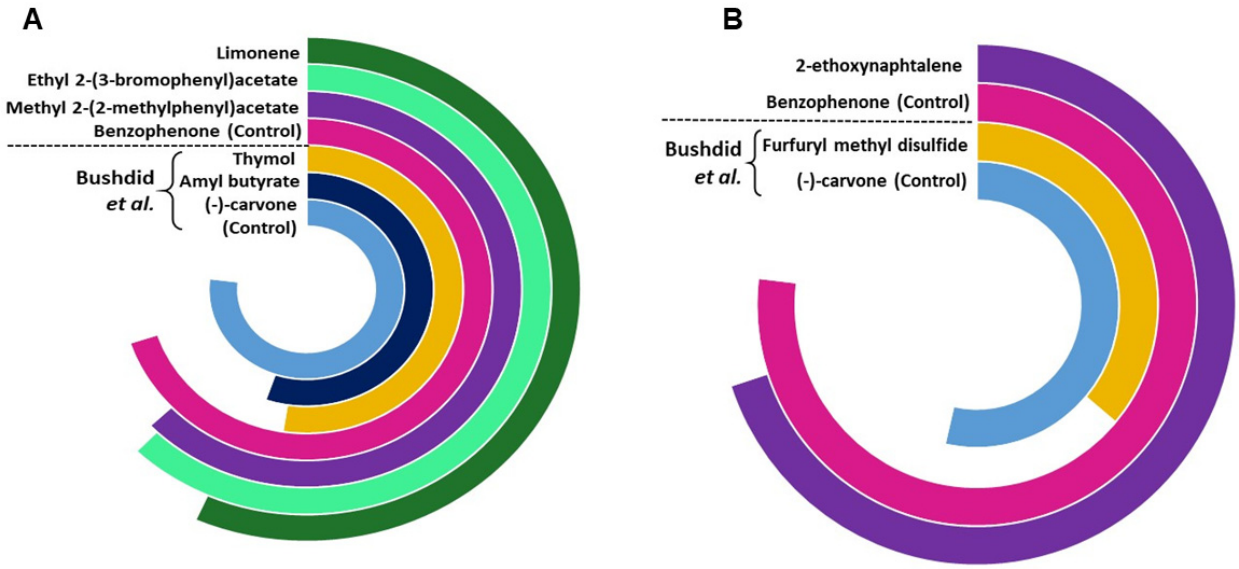

Figure 5. The comparison of potency of agonists and control identified in the current study compared with those reported by Bushdid et al. [14], for (**A**) OR1A1 and (**B**) OR2W1. The agonists reported in the current study are more potent than the agonists previously reported by Bushdid et al. [14] for both ORs.

We re-analysed the putative binding sites of novel agonists with the binding pocket of respective ORs. It is being reported by multiple studies that ligand binding niche for many ORs comprised of TM3, TM5, TM6 and TM7 [59],[85]. The putative interacting sites of all three novel agonists lie within the proposed ligand binding cradle of ORs (Supplementary Figure S9). Also, the residues G108$^{3.36}$, N109$^{3.37}$, S112$^{3.40}$, I205$^{5.46}$, Y251$^{6.48}$, Y258$^{6.55}$, T277$^{7.42}$ of OR1A1 have already been validated experimentally to be part of ligand binding pocket through site directed mutagenesis [20,21]. OR2W1 does not have any site directed mutagenesis data available yet, but the putative agonist binding residue positions within the receptor have been recognised as important in defining ligand binding cradle for ORs (Supplementary Figure S6). Position 3.36, 3.37, 5.46 (G108, S109, I206 in OR2W1) are part of ligand binding pocket in OR1A1 and OR1A2 [20]. Position 3.40 (C112 in OR2W1) is an important binding cradle position for OR1A1, OR1A2 [15], OR1G1 [23], and OR51E2 [24] while 6.48 is important for ligand binding in OR1A1. The position 7.42 is crucial for ligand binding in OR1A1 and OR7D4 [86].

In summary, we have identified two high activity agonists for OR1A1 and one high activity agonist for OR2W1 through binary classification based on RF model. The data driven approaches like ML coupled with in vitro approaches are well suited for linking odorants to their respective ORs. The proposed workflow is generic and applicable to other broadly tuned olfactory receptors including OR52D1 and a PSGR i.e., OR51E2 for discovering further high affinity ligands. Unfortunately, the majority of the ORs are either narrowly tuned or orphans so ML methods cannot be applied for discovering agonists for these ORs. Moreover, ML models can only classify the compounds that overlap the chemical space of already known compounds and are limited by their applicability domain.

Other methods such as pharmacophore-based virtual screening and structure based virtual screening might be helpful in identifying structurally different agonists.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms222111546/s1 and https://www.mdpi.com/article/10.3390/ijms222111546/s2. References [14,20,21,31,82,87–93] are also cited in the Supplementary Materials.

Author Contributions: A.J. and S.R. designed the study. A.J. acquired the data, and performed the computational analysis. C.A.d.M. and H.M. conducted the experimental validation. A.J. and S.R. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health (NIH), grant number K99DC018333 (C.A.d.M) and R01DC016224 (H.M).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in Supplementary Files.

Acknowledgments: A.J. is grateful to Macquarie University for the award of an international Macquarie University Research Excellence Scholarship (iMQRES).

Conflicts of Interest: A.J., C.A.d.M. and S.R. have nothing to declare. H.M. has received royalties from Chemcom, research grants from Givaudan and consultancy fees from Kao.

References

1. Sriram, K.; Insel, P.A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Mol. Pharm.* **2018**, *93*, 251–258. [CrossRef]
2. Buck, L.; Axel, R. A Novel Multigene Family may Encode Odorant Receptors: A Molecular Basis for Odor Recognition. *Cell* **1991**, *65*, 175–187. [CrossRef]
3. Baker, M.S.; Ahn, S.B.; Mohamedali, A.; Islam, M.T.; Cantor, D.; Verhaert, P.; Fanayan, S.; Sharma, S.; Nice, E.C.; Connor, M.; et al. Accelerating the Search for the Missing Proteins in the Human Proteome. *Nat. Commun.* **2017**, *8*, 14271. [CrossRef] [PubMed]
4. Behrens, M.; Briand, L.; de March, C.A.; Matsunami, H.; Yamashita, A.; Meyerhof, W.; Weyand, S. Structure–Function Relationships of Olfactory and Taste Receptors. *Chem. Senses* **2018**, *43*, 81–87. [CrossRef]
5. Parmentier, M.; Libert, F.; Schurmans, S.; Schiffmann, S.; Lefort, A.; Eggerickx, D.; Ledent, C.; Mollereau, C.; Gérard, C.; Perret, J.; et al. Expression of Members of the Putative Olfactory Receptor Gene Family in Mammalian Germ Cells. *Nature* **1992**, *355*, 453–455. [CrossRef] [PubMed]
6. Flegel, C.; Manteniotis, S.; Osthold, S.; Hatt, H.; Gisselmann, G. Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing. *PLoS ONE* **2013**, *8*, e55368. [CrossRef]
7. Massberg, D.; Hatt, H. Human Olfactory Receptors: Novel Cellular Functions Outside of the Nose. *Physiol. Rev.* **2018**, *98*, 1739–1763. [CrossRef]
8. Lee, S.-J.; Depoortere, I.; Hatt, H. Therapeutic Potential of Ectopic Olfactory and Taste Receptors. *Nat. Rev. Drug Discov.* **2019**, *18*, 116–138. [CrossRef]
9. Del Mármol, J.; Yedlin, M.A.; Ruta, V. The Structural Basis of Odorant Recognition in Insect Olfactory Receptors. *Nature* **2021**, *597*, 126–131. [CrossRef] [PubMed]
10. Jabeen, A.; Mohamedali, A.; Ranganathan, S. *Looking for Missing Proteins, in Reference Module in Life Sciences*; Elsevier: Amsterdam, The Nederland, 2019. [CrossRef]
11. Bush, C.F.; Hall, R.A. Olfactory Receptor Trafficking to the Plasma Membrane. *Experientia* **2008**, *65*, 2289–2295. [CrossRef] [PubMed]
12. De March, C.A.; Ryu, S.; Sicard, G.; Moon, C.; Golebiowski, J. Structure–Odour Relationships Reviewed in the Postgenomic Era. *Flavour Fragr. J.* **2015**, *30*, 342–361. [CrossRef]
13. Abaffy, T.; Bain, J.R.; Muehlbauer, M.J.; Spasojevic, I.; Lodha, S.; Bruguera, E.; O'Neal, S.; Kim, S.Y.; Matsunami, H. A Testosterone Metabolite 19-Hydroxyandrostenedione Induces Neuroendocrine Trans-Differentiation of Prostate Cancer Cells via an Ectopic Olfactory Receptor. *Front. Oncol.* **2018**, *8*, 162. [CrossRef] [PubMed]
14. Bushdid, C.; de March, C.A.; Fiorucci, S.; Matsunami, H.; Golebiowski, J. Agonists of G-Protein-Coupled Odorant Receptors Are Predicted from Chemical Features. *J. Phys. Chem. Lett.* **2018**, *9*, 2235–2240. [CrossRef]
15. Yang, K.K.; Wu, Z.; Arnold, F.H. Machine-Learning-Guided Directed Evolution for Protein Engineering. *Nat. Methods* **2019**, *16*, 687–694. [CrossRef] [PubMed]
16. Le, N.Q.K.; Kha, Q.H.; Nguyen, V.H.; Chen, Y.-C.; Cheng, S.-J.; Chen, C.-Y. Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer. *Int. J. Mol. Sci.* **2021**, *22*, 9254. [CrossRef] [PubMed]

17. Lam, L.H.T.; Le, N.; Van Tuan, L.; Ban, H.T.; Hung, T.N.K.; Nguyen, N.; Dang, L.H.; Le, N. Machine Learning Model for Identifying Antioxidant Proteins Using Features Calculated from Primary Sequences. *Biology* **2020**, *9*, 325. [CrossRef]
18. Wu, J.; Zhang, Q.; Wu, W.; Pang, T.; Hu, H.; Chan, W.K.B.; Ke, X.; Zhang, Y. WDL-RF: Predicting Bioactivities of Ligand Molecules Acting with G Protein-Coupled Receptors by Combining Weighted Deep Learning and Random Forest. *Bioinformatics* **2018**, *34*, 2271–2282. [CrossRef]
19. He, S.-B.; Hu, B.; Kuang, Z.-K.; Wang, D.; Kong, D.-X. Predicting Subtype Selectivity for Adenosine Receptor Ligands with Three-Dimensional Biologically Relevant Spectrum (BRS-3D). *Sci. Rep.* **2016**, *6*, 36595. [CrossRef]
20. Schmiedeberg, K.; Shirokova, E.; Weber, H.-P.; Schilling, B.; Meyerhof, W.; Krautwurst, D. Structural Determinants of Odorant Recognition by the Human Olfactory Receptors OR1A1 and OR1A2. *J. Struct. Biol.* **2007**, *159*, 400–412. [CrossRef]
21. Geithe, C.; Protze, J.; Kreuchwig, F.; Krause, G.; Krautwurst, D. Structural Determinants of a Conserved Enantiomer-Selective Carvone Binding Pocket in the Human Odorant Receptor OR1A1. *Cell. Mol. Life Sci.* **2017**, *74*, 4209–4229. [CrossRef]
22. Ahmed, L.; Zhang, Y.; Block, E.; Buehl, M.; Corr, M.J.; Cormanich, R.; Gundala, S.; Matsunami, H.; O'Hagan, D.; Özbil, M.; et al. Molecular Mechanism of Activation of Human Musk Receptors OR5AN1 and OR1A1 by (R)-Muscone and Diverse Other Musk-Smelling Compounds. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E3950–E3958. [CrossRef]
23. Launay, G.; Téletchéa, S.; Wade, F.; Pajot-Augy, E.; Gibrat, J.-F.; Sanz, G. Automatic Modeling of Mammalian Olfactory Receptors and Docking of Odorants. *Protein Eng. Des. Sel.* **2012**, *25*, 377–386. [CrossRef] [PubMed]
24. Wolf, S.; Jovancevic, N.; Gelis, L.; Pietsch, S.; Hatt, H.; Gerwert, K. Dynamical Binding Modes Determine Agonistic and Antagonistic Ligand Effects in the Prostate-Specific G-Protein Coupled Receptor (PSGR). *Sci. Rep.* **2017**, *7*, 16007. [CrossRef]
25. Jimenez, R.C.; Casajuana-Martin, N.; Recio, A.G.; Alcántara, L.; Pardo, L.; Campillo, M.; Gonzalez, A. The Mutational Landscape of Human Olfactory G Protein-Coupled Receptors. *BMC Biol.* **2021**, *19*, 21. [CrossRef] [PubMed]
26. Lim, V.J.Y.; Du, W.; Chen, Y.Z.; Fan, H. A Benchmarking Study on Virtual Ligand Screening Against Homology Models of Human GPCRs. *Proteins Struct. Funct. Bioinform.* **2018**, *86*, 978–989. [CrossRef] [PubMed]
27. Sharifi-Rad, J.; Salehi, B.; Varoni, E.M.; Sharopov, F.; Yousaf, Z.; Ayatollahi, S.A.; Kobarfard, F.; Sharifi-Rad, M.; Afdjei, M.H.; Sharifi-Rad, M.; et al. Plants of the Melaleuca Genus as Antimicrobial Agents: From Farm to Pharmacy. *Phytotherapy Res.* **2017**, *31*, 1475–1494. [CrossRef] [PubMed]
28. Jabeen, A.; Ranganathan, S. Applications of Machine Learning in GPCR Bioactive Ligand Discovery. *Curr. Opin. Struct. Biol.* **2019**, *55*, 66–76. [CrossRef]
29. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590–596. [CrossRef] [PubMed]
30. Jabeen, A.; Vijayram, R.; Ranganathan, S. BIO-GATS: A Tool for Automated GPCR Template Selection Through a Biophysical Approach for Homology Modeling. *Front. Mol. Biosci.* **2021**, *8*, 617176. [CrossRef]
31. Braun, T.; Voland, P.; Kunz, L.; Prinz, C.; Gratzl, M. Enterochromaffin Cells of the Human Gut: Sensors for Spices and Odorants. *Gastroenterology* **2007**, *132*, 1890–1901. [CrossRef]
32. Tan, Q.; Cui, J.; Huang, J.; Ding, Z.; Lin, H.; Niu, X.; Li, Z.; Wang, G.; Luo, Q.; Lu, S. Genomic Alteration During Metastasis of Lung Adenocarcinoma. *Cell. Physiol. Biochem.* **2016**, *38*, 469–486. [CrossRef]
33. Wu, C.; Jia, Y.; Hae, J.L.; Kim, Y.; Sekharan, S.; Batista, V.S.; Lee, S.-J. Act. OR1A1 Suppresses PPAR-Gamma Expr. By Inducing HES-1 Cult. Hepatocytes. *Int. J. Biochem. Cell Biol.* **2015**, *64*, 75–80. [CrossRef] [PubMed]
34. Umemura, S.; Mimaki, S.; Makinoshima, H.; Tada, S.; Ishii, G.; Ohmatsu, H.; Niho, S.; Yoh, K.; Matsumoto, S.; Takahashi, A.; et al. Therapeutic Priority of the PI3K/AKT/mTOR Pathway in Small Cell Lung Cancers as Revealed by a Comprehensive Genomic Analysis. *J. Thorac. Oncol.* **2014**, *9*, 1324–1331. [CrossRef] [PubMed]
35. Saito, N.; Yamano, E.; Ishii, A.; Tanaka, M.; Nakamura, J.; Watanabe, Y. Involvement of the Olfactory System in the Induction of Anti-Fatigue Effects by Odorants. *PLoS ONE* **2018**, *13*, e019526. [CrossRef] [PubMed]
36. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminform.* **2018**, *10*, 4. [CrossRef]
37. Schober, P.; Boer, C.; Schwarte, L.A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef] [PubMed]
38. Caballero-Vidal, G.; Bouysset, C.; Grunig, H.; Fiorucci, S.; Montagné, N.; Golebiowski, J.; Jacquin-Joly, E. Machine Learning Decodes Chemical Features to Identify Novel Agonists of a Moth Odorant Receptor. *Sci. Rep.* **2020**, *10*, 1655. [CrossRef]
39. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.R-project.org (accessed on 21 October 2021).
40. Chawla, N.V.; Bowyer, K.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
41. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz Information Miner: Version 2.0 and beyond. *AcM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31. [CrossRef]
42. Fawagreh, K.; Gaber, M.; Elyan, E. Random Forests: From Early Developments to Recent Advancements. *Syst. Sci. Control. Eng.* **2014**, *2*, 602–609. [CrossRef]
43. Tian, Y.; Shi, Y.; Liu, X. Recent Advances on Support Vector Machines Research. *Technol. Econ. Dev. Econ.* **2012**, *18*, 5–33. [CrossRef]
44. Lee, J.; Kumar, S.; Lee, S.-Y.; Park, S.J.; Kim, M.-H. Development of Predictive Models for Identifying Potential S100A9 Inhibitors Based on Machine Learning Methods. *Front. Chem.* **2019**, *7*, 779. [CrossRef]

45. Irwin, J.; Shoichet, B.K. ZINC–A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182. [CrossRef]
46. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Allison, P.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef]
47. Degtyarenko, K.; De Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2008**, *36*, D344–D350. [CrossRef] [PubMed]
48. Janfaza, S.; Nojavani, M.B.; Khorsand, B.; Nikkhah, M.; Zahiri, J. Cancer Odor Database (COD): A Critical Databank for Cancer Diagnosis Research. *Database* **2017**, *2017*, bax055. [CrossRef]
49. Marenco, L.; Wang, R.; McDougal, R.; Olender, T.; Twik, M.; Bruford, E.; Liu, X.; Zhang, J.; Lancet, D.; Shepherd, G.; et al. ORDB, HORDE, ODORactor and other On-Line Knowledge Resources of Olfactory Receptor-Odorant Interactions. *Database* **2016**, *2016*, baw132. [CrossRef] [PubMed]
50. Jabeen, A.; Mohamedali, A.; Ranganathan, S. Protocol for Protein Structure Modelling. In *Encyclopedia of Bioinformatics and Computational Biology*; Ranganathan, S., Nakai, K., Schonbach, C., Eds.; Academic Press: Oxford, UK, 2019; pp. 252–272. [CrossRef]
51. Okada, T.; Sugihara, M.; Bondar, A.-N.; Elstner, M.; Entel, P.; Buss, V. The Retinal Conformation and its Environment in Rhodopsin in Light of a New 2.2Å Crystal Structure. *J. Mol. Biol.* **2004**, *342*, 571–583. [CrossRef]
52. Cvicek, V.; Goddard, W.A., 3rd; Abrol, R. Structure-Based Sequence Alignment of the Transmembrane Domains of All Human GPCRs: Phylogenetic, Structural and Functional Implications. *PLoS Comput. Biol.* **2016**, *12*, e1004805. [CrossRef]
53. Webb, B.; Sali, A. Protein Structure Modeling with MODELLER. *Methods Mol. Biol.* **2017**, *1654*, 39–54. [CrossRef]
54. Williams, C.J.; Headd, J.J.; Moriarty, N.W.; Prisant, M.G.; Videau, L.L.; Deis, L.N.; Verma, V.; Keedy, D.A.; Hintze, B.; Chen, V.B.; et al. MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci.* **2018**, *27*, 293–315. [CrossRef]
55. Wang, Q.; Canutescu, A.A.; Dunbrack, R.L., Jr. SCWRL and MolIDE: Computer Programs for Side-Chain Conformation Prediction and Homology Modeling. *Nat. Protoc.* **2008**, *3*, 1832–1847. [CrossRef]
56. Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM: A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation. *J. Comput. Chem.* **1994**, *15*, 488–506. [CrossRef]
57. Saito, H.; Kubota, M.; Roberts, R.W.; Chi, Q.; Matsunami, H. RTP Family Members Induce Functional Expression of Mammalian Odorant Receptors. *Cell* **2004**, *119*, 679–691. [CrossRef] [PubMed]
58. Dahoun, T.; Grasso, L.; Vogel, H.; Pick, H. Recombinant Expression and Functional Characterization of Mouse Olfactory Receptor mOR256-17 in Mammalian Cells. *Biochemistry* **2011**, *50*, 7228–7235. [CrossRef]
59. Gelis, L.; Wolf, S.; Hatt, H.; Neuhaus, E.M.; Gerwert, K. Prediction of a Ligand-Binding Niche Within a Human Olfactory Receptor by Combining Site-Directed Mutagenesis with Dynamic Homology Modeling. *Angew. Chem. Int. Ed. Engl.* **2012**, *51*, 1274–1278. [CrossRef] [PubMed]
60. Busse, D.; Kudella, P.; Grüning, N.-M.; Gisselmann, G.; Ständer, S.; Luger, T.; Jacobsen, F.; Steinsträßer, L.; Paus, R.; Gkogkolou, P.; et al. A Synthetic Sandalwood Odorant Induces Wound-Healing Processes in Human Keratinocytes via the Olfactory Receptor OR2AT4. *J. Investig. Dermatol.* **2014**, *134*, 2823–2832. [CrossRef]
61. Maßberg, D.; Simon, A.; Häussinger, D.; Keitel, V.; Gisselmann, G.; Conrad, H.; Hatt, H. Monoterpene (−)-Citronellal Affects Hepatocarcinoma Cell Signaling Via an Olfactory Receptor. *Arch. Biochem. Biophys.* **2015**, *566*, 100–109. [CrossRef]
62. Thach, T.T.; Hong, Y.-J.; Lee, S.; Lee, S.-J. Molecular Determinants of the Olfactory Receptor Olfr544 Activation by Azelaic Acid. *Biochem. Biophys. Res. Commun.* **2017**, *485*, 241–248. [CrossRef]
63. Tong, T.; Ryu, S.E.; Min, Y.; de March, C.A.; Bushdid, C.; Golebiowski, J.; Moon, C.; Park, T. Olfactory Receptor 10J5 Responding to A-Cedrene Regulates Hepatic Steatosis via the cAMP-PKA Pathway. *Sci. Rep.* **2017**, *7*, 9471. [CrossRef]
64. Weber, L.; Schulz, W.; Philippou, S.; Eckardt, J.; Ubrig, B.; Hoffmann, M.J.; Tannapfel, A.; Kalbe, B.; Gisselmann, G.; Hatt, H. Characterization of the Olfactory Receptor OR10H1 in Human Urinary Bladder Cancer. *Front. Physiol.* **2018**, *9*, 456. [CrossRef]
65. Liu, M.T.; Ho, J.; Liu, J.K.; Purakait, R.; Morzan, U.N.; Ahmed, L.; Batista, V.S.; Matsunami, H.; Ryan, K. Carbon Chain Shape Selectivity by the Mouse Olfactory Receptor OR-I7. *Org. Biomol. Chem.* **2018**, *16*, 2541–2548. [CrossRef]
66. Choi, Y.; Shim, J.; Park, J.-H.; Kim, Y.-S.; Kim, M. Discovery of Orphan Olfactory Receptor 6M1 as a New Anticancer Target in MCF-7 Cells by a Combination of Surface Plasmon Resonance-Based and Cell-Based Systems. *Sensors* **2021**, *21*, 3468. [CrossRef] [PubMed]
67. Gat, U.; Nekrasova, E.; Lancet, D.; Natochin, M. Olfactory Receptor Proteins. Expression, Characterization and Partial Purification. *Eur. J. Biochem.* **1994**, *225*, 1157–1168. [CrossRef] [PubMed]
68. Katada, S.; Nakagawa, T.; Kataoka, H.; Touhara, K. Odorant Response Assays for a Heterologously Expressed Olfactory Receptor. *Biochem. Biophys. Res. Commun.* **2003**, *305*, 964–969. [CrossRef]
69. Bushdid, C.; de March, C.A.; Matsunami, H.; Golebiowski, J. Numerical Models and In vitro Assays to Study Odorant. *Receptors* **2018**, *1820*, 77–93. [CrossRef]
70. Zhuang, H.; Matsunami, H. Evaluating Cell-Surface Expression and Measuring Activation of Mammalian Odorant Receptors in Heterologous Cells. *Nat. Protoc.* **2008**, *3*, 1402–1413. [CrossRef]
71. Krautwurst, D.; Yau, K.-W.; Reed, R.R. Identification of Ligands for Olfactory Receptors by Functional Expression of a Receptor Library. *Cell* **1998**, *95*, 917–926. [CrossRef]

72. Zhuang, H.; Matsunami, H. Synergism of Accessory Factors in Functional Expression of Mammalian Odorant Receptors. *J. Biol. Chem.* **2007**, *282*, 15284–15293. [CrossRef]
73. Li, Y.R.; Matsunami, H. Activation State of the M3 Muscarinic Acetylcholine Receptor Modulates Mammalian Odorant Receptor Signaling. *Sci. Signal.* **2011**, *4*, ra1. [CrossRef]
74. Ikegami, K.; de March, C.A.; Nagai, M.H.; Ghosh, S.; Do, M.; Sharma, R.; Bruguera, E.S.; Lu, Y.E.; Fukutani, Y.; Vaidehi, N.; et al. Structural Instability and Divergence from Conserved Residues Underlie Intracellular Retention of Mammalian Odorant Receptors. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 2957–2967. [CrossRef]
75. Laird, D.W.; Molday, R.S. Evidence Against the Role of Rhodopsin in Rod Outer Segment Binding to RPE Cells. *Investig. Ophthalmol. Vis. Sci.* **1988**, *29*, 419–428.
76. Dey, S.; Matsunami, H. Calreticulin Chaperones Regulate Functional Expression of Vomeronasal Type 2 Pheromone Receptors. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16651–16656. [CrossRef]
77. Jović, A.; Brkić, K.; Bogunović, N. A Review of Feature Selection Methods with Applications. In Proceedings of the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 25–29 May 2015. [CrossRef]
78. Aboudi, N.E.; Benhlima, L. Review on Wrapper Feature Selection Approaches. In Proceedings of the 2016 International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco, 22–24 September 2016. [CrossRef]
79. Li, M.; Ling, C.; Xu, Q.; Gao, J. Classification of G-Protein Coupled Receptors Based on a Rich Generation of Convolutional Neural Network, N-Gram Transformation and Multiple Sequence Alignments. *Amino Acids* **2018**, *50*, 255–266. [CrossRef]
80. Yang, M.; Tao, B.; Chen, C.; Jia, W.; Sun, S.; Zhang, T.; Wang, X. Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors. *J. Chem. Inf. Model.* **2019**, *59*, 5002–5012. [CrossRef]
81. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, A.B.; Thiessen, A.P.; Yu, B.; et al. PubChem 2019 update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109. [CrossRef]
82. Ballesteros, J.A.; Weinstein, H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods Neurosci.* **1995**, *25*, 366–428. [CrossRef]
83. Saito, H.; Chi, Q.; Zhuang, H.; Matsunami, H.; Mainland, J.D. Odor Coding by a Mammalian Receptor Repertoire. *Sci. Signal.* **2009**, *2*, ra9. [CrossRef]
84. Rashidi, H.H.; Tran, N.K.; Betts, E.V.; Howell, L.P.; Green, R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad. Pathol.* **2019**, *6*, 2374289519873088. [CrossRef]
85. de March, C.A.; Yu, Y.; Ni, M.J.; Adipietro, K.A.; Matsunami, H.; Ma, M.; Golebiowski, J. Conserved Residues Control Activation of Mammalian G Protein-Coupled Odorant Receptors. *J. Am. Chem. Soc.* **2015**, *137*, 8611–8616. [CrossRef]
86. Keller, A.; Zhuang, H.; Chi, Q.; Vosshall, L.B.; Matsunami, H. Genetic Variation in a Human Odorant Receptor Alters Odour Perception. *Nature* **2007**, *449*, 468–472. [CrossRef]
87. Belloir, C.; Miller-Leseigneur, M.L.; Neiers, F.; Briand, L.; Le Bon, A.M. Biophysical and functional characterization of the human olfactory receptor OR1A1 expressed in a mammalian inducible cell line. *Protein Expr. Purif.* **2017**, *129*, 31–43. [CrossRef] [PubMed]
88. Mainland, J.D.; Keller, A.; Li, Y.R.; Zhou, T.; Trimmer, C.; Snyder, L.L.; Moberly, A.H.; Adipietro, K.A.; Liu, W.L.; Zhuang, H.; et al. The missense of smell: Functional variability in the human odorant receptor repertoire. *Nat. Neurosci.* **2014**, *17*, 114–120. [CrossRef]
89. Sato-Akuhara, N.; Horio, N.; Kato-Namba, A.; Yoshikawa, K.; Niimura, Y.; Ihara, S.; Shirasu, M.; Touhara, K. Ligand Specificity and Evolution of Mammalian Musk Odor Receptors: Effect of Single Receptor Deletion on Odor Detection. *J. Neurosci.* **2016**, *36*, 4482–4491. [CrossRef]
90. Adipietro, K.A.; Mainland, J.D.; Matsunami, H. Functional evolution of mammalian odorant receptors. *PLoS Genet.* **2012**, *8*, e1002821. [CrossRef] [PubMed]
91. Li, S.; Ahmed, L.; Zhang, R.; Pan, Y.; Matsunami, H.; Burger, J.L.; Block, E.; Batista, V.S.; Zhuang, H. Smelling Sulfur: Copper and Silver Regulate the Response of Human Odorant Receptor OR2T11 to Low-Molecular-Weight Thiols. *J. Am. Chem. Soc.* **2016**, *138*, 13281–13288. [CrossRef]
92. 92. Audouze, K.; Tromelin, A.; Le Bon, A.M.; Belloir, C.; Petersen, R.K.; Kristiansen, K.; Brunak, S.; Taboureau, O. Identification of odorant-receptor interactions by global mapping of the human odorome. *PLoS ONE* **2014**, *9*, e93037. [CrossRef]
93. McRae, J.F.; Mainland, J.D.; Jaeger, S.R.; Adipietro, K.A.; Matsunami, H.; Newcomb, R.D. Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the "grassy" smelling odor, *cis-3-hexen-1-ol*. *Chem. Senses* **2012**, *37*, 585–593. [CrossRef] [PubMed]