

# Discovery of Novel Molecular Frameworks of Farnesoid X Receptor Modulators by Ensemble Machine Learning

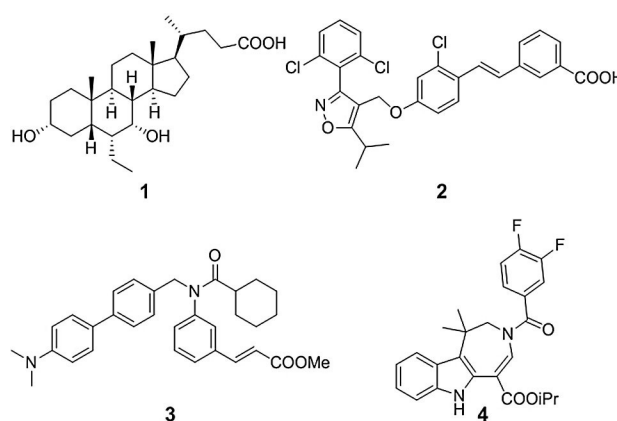
Daniel Merk<sup>+</sup>,<sup>\*,[a]</sup> Francesca Grisoni<sup>+</sup>,<sup>[a, b]</sup> Kay Schaller,<sup>[a]</sup> Lukas Friedrich,<sup>[a]</sup> and Gisbert Schneider<sup>\*,[a]</sup>

The bile acid activated transcription factor farnesoid X receptor (FXR) has revealed therapeutic potential as a molecular drug target for the treatment of hepatic and metabolic disorders. Despite strong efforts in FXR ligand development, the structural diversity among the known FXR modulators is limited. Only four molecular frameworks account for more than 50% of the FXR modulators annotated in ChEMBL. Here, we leverage machine learning methods to expand the chemical space of FXR-targeting small molecules by employing an ensemble of three complementary machine learning approaches. A counter-propagation artificial neural network, a *k*-nearest neighbor learner,

and a three-dimensional pharmacophore descriptor were combined to retrieve novel FXR ligands from a collection of more than 3 million compounds. The ensemble machine learning model identified six new FXR modulators among ten top-ranked candidates. These active hits comprise both FXR activators and antagonists with micromolar potencies. With four novel FXR ligand scaffolds, these computationally identified bioactive compounds appreciably expand the chemical space of known FXR modulators and may serve as starting points for hit-to-lead expansion.

## 1. Introduction

Ligands of the bile acid activated transcription factor farnesoid X receptor (FXR) possess therapeutic potential for the treatment of hepatic and metabolic disorders. The first-in-class FXR agonist obeticholic acid (OCA, **1**; Scheme 1)<sup>[1]</sup> is approved for second-line treatment of the rare liver disorder primary biliary cholangitis<sup>[2]</sup> and is expected to gain further relevance in the treatment of nonalcoholic fatty liver (NAFL) and nonalcoholic steatohepatitis (NASH).<sup>[3]</sup> Clinical trials with **1** have shown promising efficacy and validated FXR as a drug target for NAFL/NASH as the hepatic manifestation of the metabolic syndrome.<sup>[4,5]</sup> Steroidal FXR agonist **1** is succeeded by several analogues of FXR agonist GW4064 (**2**)<sup>[6]</sup> in early stages of clinical trials, whereas other FXR modulators, such as **3**<sup>[7]</sup> and **4**,<sup>[8]</sup>



**Scheme 1.** FXR agonists **1–4**: Obeticholic acid<sup>[1]</sup> (OCA, **1**;  $EC_{50}$  = 0.1  $\mu$ M), GW4064<sup>[6]</sup> (**2**,  $EC_{50}$  = 0.065  $\mu$ M), Fexaramine<sup>[7]</sup> (**3**,  $EC_{50}$  = 0.025  $\mu$ M), WAY-362450<sup>[8]</sup> (XL335, **4**;  $EC_{50}$  = 0.004  $\mu$ M).

failed in (pre)clinical development. Our analysis of the ChEMBL23<sup>[9]</sup> compound database revealed a limited scaffold diversity of known modulators, and merely four frameworks accounted for more than 50% of all annotated FXR modulators (median effective concentration/median inhibitory concentration,  $EC_{50}/IC_{50}$  < 50  $\mu$ M, 1134 compounds), with framework I (GW4064 derivatives and structural analogues) contained in approximately one third of the ligands (Figure 1). Here, we present a machine learning approach for virtual screening of large compound collections, which in a prospective application led to the identification of four new FXR ligand scaffolds with a success rate of 60%.

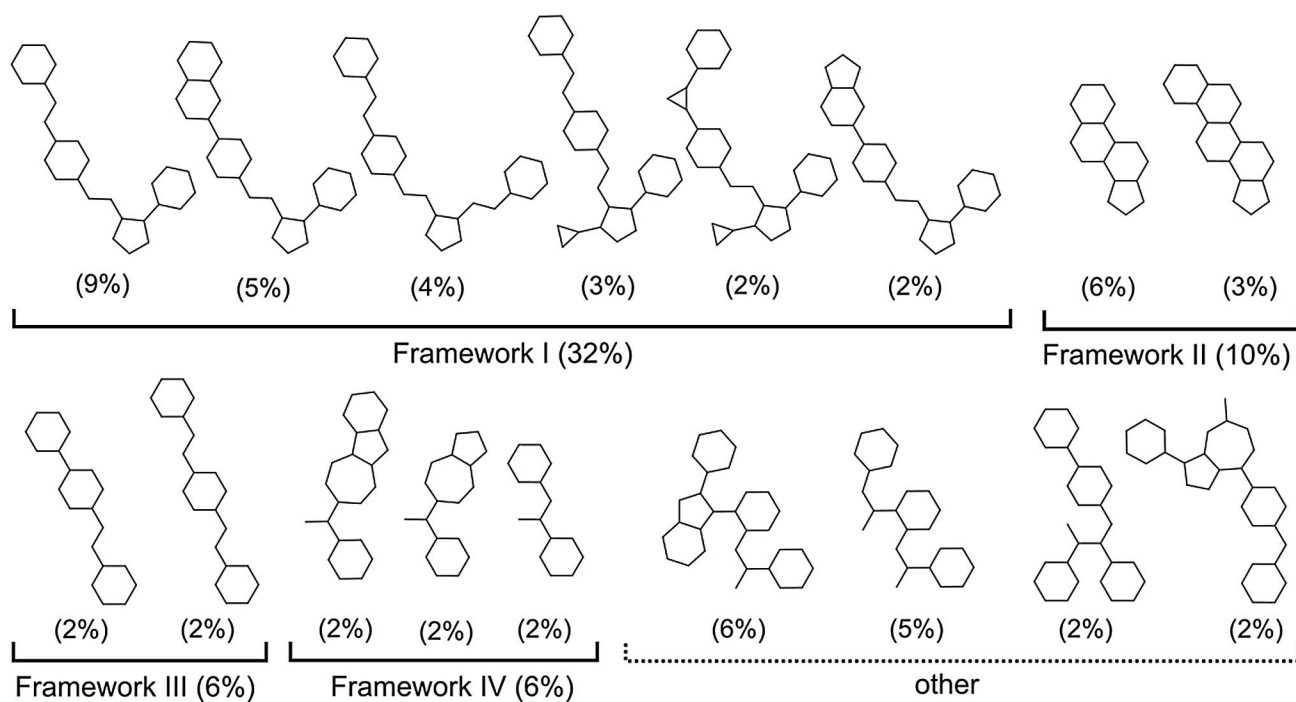
[a] Dr. D. Merk,<sup>+</sup> Dr. F. Grisoni,<sup>+</sup> K. Schaller, L. Friedrich, Prof. Dr. G. Schneider  
Department of Chemistry and Applied Biosciences  
Swiss Federal Institute of Technology (ETH) Zurich  
Vladimir-Prelog-Weg 4, 8093 Zurich (Switzerland)  
E-mail: daniel.merk@pharma.ethz.ch  
gisbert.schneider@pharma.ethz.ch

[b] Dr. F. Grisoni<sup>+</sup>  
Department of Earth and Environmental Sciences  
University of Milano-Bicocca  
Piazza della Scienza 1, 20126 Milano (Italy)

[<sup>+</sup>] These authors contributed equally to this work

Supporting Information and the ORCID identification number(s) for the author(s) of this article can be found under:  
<https://doi.org/10.1002/open.201800156>.

© 2018 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.



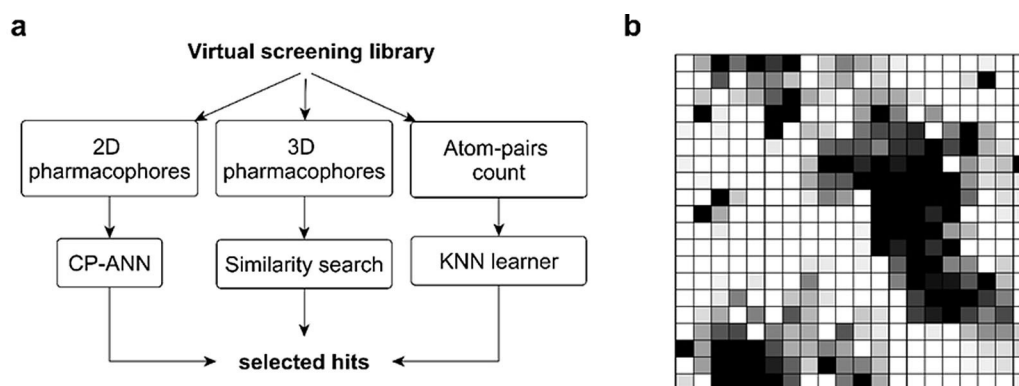
**Figure 1.** Most prevalent molecular graph frameworks of FXR modulators from ChEMBL23 ( $EC_{50}/IC_{50} < 50 \mu\text{M}$ , 1134 compounds) accounting for approximately 60% of all annotated FXR agonists and antagonists. The four most relevant ligand frameworks (Roman numerals), accounting for more than 50% of the known FXR ligands, are exemplified by representative scaffolds (frequency  $\geq 2\%$ ).

## 2. Results and Discussion

We implemented an ensemble machine learning approach as a computational strategy for virtual FXR ligand screening (Figure 2a). Ensemble approaches merge complementary computational concepts, each grasping partial information of the training data, to potentially reach better predictions than the individual models.<sup>[10]</sup> We implemented and combined three distinct computational strategies:

- 1) Counter-propagation artificial neural network<sup>[11]</sup> (CP-ANN) set up for Chemically Advanced Template Search<sup>[12]</sup> (CATS2)

descriptors and trained on a collection of 896 compounds tested on FXR, which were retrieved and curated from ChEMBL23.<sup>[9]</sup> CP-ANN is a modeling technique that combines aspects of supervised and unsupervised learning, for which the classifier self-organizes on the basis of the structural features and experimental responses of the chemicals. This algorithm allows the obtainment of a map (Figure 2b), on which the training compounds are clustered according to the similarity of both their experimental properties (i.e. FXR modulation) and their structural features (i.e. CATS2 descriptors). CATS2 descriptors<sup>[12]</sup> are based on the occurrence of pharmacophore feature pairs (lipophilic, aromatic, hydro-



**Figure 2.** Ensemble machine learning for the discovery of novel FXR modulators. a) Overview of the ensemble approach. The approach is based on three different models of the chemical space, namely, a counter-propagation artificial neural network (CP-ANN), trained on topological pharmacophores (CATS2); a similarity-based ranking, utilizing 3D pharmacophore distributions of 17 actives as templates; and a  $k$ -nearest neighbor (kNN) learner trained on atom-pair counts (AtomPair binary fingerprints). b) Depiction of the self-organizing map generated by CP-ANN training. Each square represents one neuron of the CP-ANN map. Coloring indicates the FXR-activity likeliness of the neuron. Compounds assigned to “high-excitation” neurons (depicted in black) are classified as FXR modulators. The training procedure successfully clustered active compounds in the same or in neighboring regions of the map.

- gen-bond acceptor, and hydrogen-bond donor atoms) at topological distances up to ten bonds and are specifically developed for scaffold hopping.<sup>[13]</sup> In previous studies, CATS2 enabled the identification of novel modulators of another nuclear receptor (retinoid X receptor).<sup>[14,15]</sup>
- 2) Similarity of 3D pharmacophore feature distributions (LIQUID)<sup>[16]</sup> based on 17 selected FXR agonists as templates. LIQUID is a similarity method that captures the spatial distribution of potential pharmacophore points (lipophilic, aromatic, positively and negatively charged, hydrogen-bond acceptor, and hydrogen-bond donor atoms) as Gaussians, which allows the generation of a probabilistic 3D pharmacophore model of the bioactive template compound(s). For similarity searching, the LIQUID model is represented as a descriptor vector. LIQUID performs alignment-free similarity searching of compound libraries by utilizing the pairwise Euclidean distance between the descriptor vectors of the model and the screening compounds. The LIQUID descriptor vector was computed separately for each of the 17 selected FXR agonists.
  - 3) *k*-nearest neighbor learner (kNN) trained on atom-pair distributions<sup>[17]</sup> at given topological distances by using the training data assembled from ChEMBL23.<sup>[9]</sup> kNN is a similarity-based machine learning algorithm that utilizes the information of portions of the chemical space (i.e. the *k* most similar molecules, "neighbors") to predict the activity of the query as the most frequently observed activity of its neighbors. For each molecule, a binary vector of 1024 bit was generated to capture the presence of all pairs of atoms at increasing topological distance (AtomPair molecular fingerprint). The kNN classifier was then trained on the binary representations to capture patterns present in the training data.

Each modeling method was selected as the result of a retrospective optimization procedure, in which we analyzed the performance of seven distinct molecular descriptions capturing 2D and 3D pharmacophore distributions (CATS2 and LIQUID, respectively), radial fragments (Morgan and FeatMorgan binary fingerprints),<sup>[18]</sup> atom pairs (AtomPair fingerprints), topological and physicochemical properties (MOE2D descriptors),<sup>[19]</sup> and molecular shape and partial charge distribution (WHALES).<sup>[20]</sup> The aggregation of predictors allows diverse molecular features responsible for the bioactivity to be taken into consideration, which thereby increases the overall predictive confidence. The ensemble of the three methods was utilized to screen a library of 3 million commercially available compounds that were compiled from four vendor catalogues. Each compound was scored individually by each method. The 500 top-scoring compounds on each ranked list were pooled and sorted according to the sum of their reciprocal ranks.

The ten top-ranking compounds, namely, **5–14**, from the consensus list were ordered and characterized in a specific Gal4 hybrid reporter gene assay for FXR activation.<sup>[21]</sup> This test system was based on a chimeric receptor composed of the human FXR ligand binding domain fused to the DNA binding domain of the Gal4 receptor from yeast. A Gal4-responsive fire-

fly luciferase and a constitutively expressed *Renilla* luciferase served as the reporter gene and as the internal control, respectively. The effects of **5–14** were obtained at a concentration of 10  $\mu\text{M}$  on FXR-Gal4 alone and in competition with **2** (1  $\mu\text{M}$ ) to detect antagonistic effects. In addition, the assay was repeated in the absence of the hybrid receptor construct as a control experiment to exclude unspecific effects. Full dose–response curves were recorded for the active compounds (Table 1).

Compounds **5**, **6**, **8**, and **11** were confirmed as FXR activators with low-micromolar  $\text{EC}_{50}$  values (ranging between 6  $\mu\text{M}$  and 14  $\mu\text{M}$ ) and modest (6–11-fold) activation efficacy. Compounds **7** and **9** revealed antagonistic potency ( $\text{IC}_{50}$  ranging between 32  $\mu\text{M}$  and > 50  $\mu\text{M}$ ) in competition with reference agonist **2** (1  $\mu\text{M}$ ). All FXR activating hits have a fatty-acid mimetic molecular architecture<sup>[22]</sup> and comprise a benzoic acid moiety. Compounds **5**, **6**, and **11** are structurally related and share a 5*H*-thiazolo[3,2-*a*]pyrimidin-3(2*H*)-one system as their central scaffold but differ in the substitution pattern of this residue. Compound **10**, despite sharing the same ring system as **5**, **6**, and **11**, lacked activity on FXR. This observation seemed to be due to the spacious naphthyl moiety of **10**, which is forced into a dihedral conformation by the neighboring substituents and may clash with the target protein. The molecular architecture of **8** is distinct from that of **5**, **6**, and **11**, and it thus constitutes another novel FXR activator scaffold. Compound **7** was identified as an FXR antagonist, albeit with modest potency ( $\text{IC}_{50} = 32 \pm 5 \mu\text{M}$ ) in competition with reference FXR agonist **2** (1  $\mu\text{M}$ ). This antagonist also repressed intrinsic baseline FXR activity in the reporter gene assay. Compound **8** shares structural similarity with several recently reported partial FXR agonists<sup>[23]</sup> but comprises more bulky substituents and has a geometry that is slightly different to that of these known ligands. The antagonistic potency of **8** was very weak ( $\text{IC}_{50} > 50 \mu\text{M}$ ) and could not be exactly quantified.

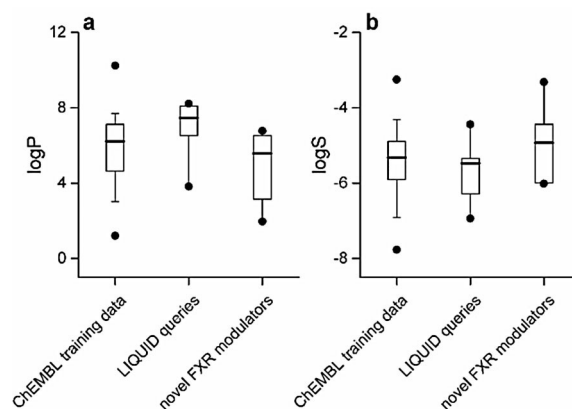
With low-micromolar potency and innovative scaffolds, new FXR ligands **5–9** and **11** may be considered for optimization towards more potent FXR modulators. Importantly, none of these active hits possesses a molecular framework that is annotated as FXR modulator in ChEMBL24 ( $\text{EC}_{50}/\text{IC}_{50} < 50 \mu\text{M}$ , 1134 compounds) or that is present in the patented data of SureChEMBL (v.2018).<sup>[24]</sup> The lipophilicity and solubility of these novel FXR modulators are predicted to be higher than those of the active molecules utilized for model development and similarity searching (Figure 3).

To evaluate potential ligand binding poses and interactions with the nuclear receptor, all active hits were subjected to molecular docking by using the GOLD<sup>[25]</sup> algorithm with flexible fit of the receptor. FXR activators **5**, **6**, **8**, and **11** (Figure 4) were docked into an activated FXR-LBD complex (PDB-ID: 3FXV<sup>[26]</sup> with the agonist GW4064 *N*-oxide bound), whereas a model of the inactivated FXR complex (PDB-ID: 4OIV<sup>[27]</sup> with the antagonist NDB bound) was used for antagonists **7** and **9**.

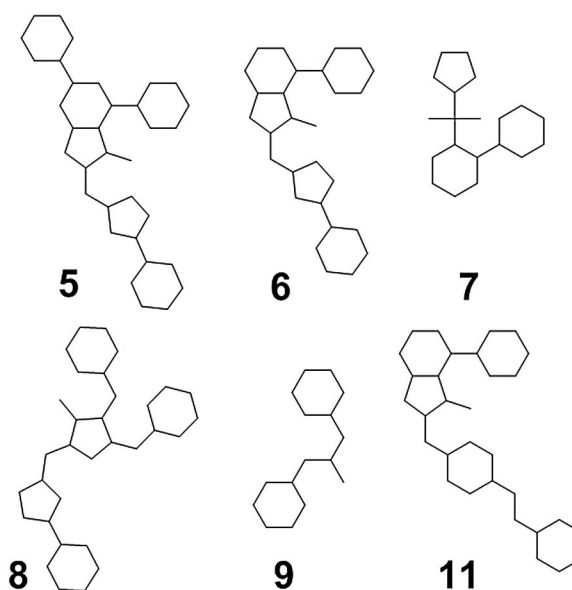
Due to the lipophilic binding site of FXR, few polar interactions were observable for the new FXR ligands. The computed binding poses suggest that FXR agonists **5**, **6**, and **11** form a canonical neutralizing interaction with Arg335 and entirely fill the spacious FXR ligand-binding site (Figure 5a,b). The dock-

**Table 1.** In vitro activity of computationally selected compounds 5–14 in a specific FXR-Gal4 hybrid reporter gene assay.<sup>[a]</sup>

Compd	Structure	In vitro activity <sup>[b]</sup> (FXR)
5		EC <sub>50</sub> = 6.3 ± 0.2 μM (7.7 ± 0.2-fold act.)
6		EC <sub>50</sub> = 7.4 ± 0.5 μM (5.8 ± 0.2-fold act.)
7		IC <sub>50</sub> = 32 ± 5 μM
8		EC <sub>50</sub> = 7.1 ± 0.6 μM (6.3 ± 0.1-fold act.)
9		IC <sub>50</sub> > 50 μM
10		inactive
11		EC <sub>50</sub> = 14 ± 2 μM (11 ± 2-fold act.)
12		inactive
13		inactive



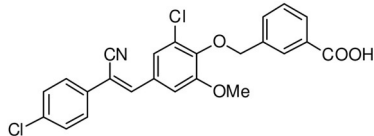
**Figure 3.** Comparison of the molecular properties of the novel modulators compared to the ChEMBL training and query molecules. Distribution of the a) lipophilicity (atomic log *P*, A log *P*) and b) aqueous solubility (atomic log *S*, A log *S*) values of the novel actives compared to the training molecules (CP-ANN and kNN) and the query compounds of the LIQUID-based similarity search. The boxes represent the 1st and 3rd quartiles, the median (solid line), the 10th and 90th percentiles (whiskers), and the minimum/maximum values (dots). The novel modulators are less lipophilic and potentially more soluble than the training/template compounds.



**Figure 4.** Graph scaffolds<sup>[28]</sup> of the novel FXR modulators identified by ensemble machine learning. The novel modulators possess six distinct and novel scaffolds (those of 5, 6, and 11 being closely related), which belong to four ligand frameworks.

ing solutions also suggest an interaction between the chlorine atom of 5 with Tyr373, similar to the co-crystallized ligand that interacts with Tyr373 through its pyridine *N*-oxide moiety.

FXR agonist 8 also interacts with Arg335 but its suggested binding mode notably differs from that of the co-crystallized agonist and modulators 5, 6, and 11 (Figure 5c). Due to the Y-shaped structure of agonist 8, the dichlorophenylpyrrol moiety is bound close to helix 12,

Table 1. (Continued)		
Compd	Structure	In vitro activity <sup>[b]</sup> (FXR)
14		inactive
[a] Results are the mean ± SEM ( $n=2$ for inactives, $n \geq 3$ for actives); antagonistic compounds <b>7</b> and <b>9</b> were characterized in competition with reference agonist <b>2</b> at a fixed concentration of $1 \mu\text{M}$ . [b] Inactive: no statistically significant reporter transactivation or repression at $10 \mu\text{M}$ .		

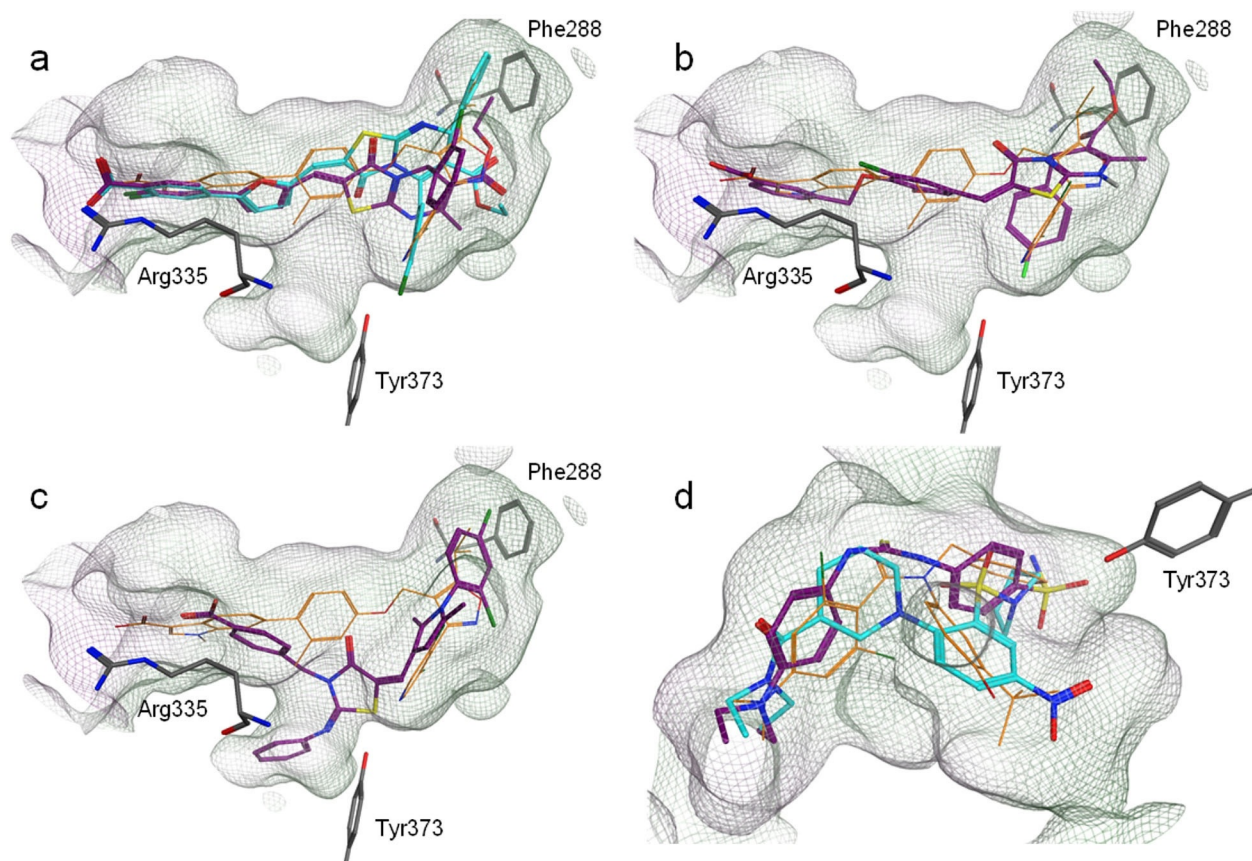
whereas the phenylimine residue protrudes to a subpocket of the ligand-binding site that is not explored by the co-crystallized ligand. Compared to GW4064 (**2**) and derivatives, which typically interact with His447 and Phe329 through their unique

phenylisoxazole hammerhead structure,<sup>[29]</sup> FXR ligands **5**, **6**, **8**, and **11** lack these specific contacts.

Similar to the co-crystallized antagonist NDB, no polar interactions with FXR were observed for antagonists **7** and **9** in the docking poses (Figure 5d). The dialkylaniline moieties of both antagonists aligned with the respective group of the crystallized ligand. The binding geometry of antagonist **7** was similar to that of NDB and occupied the entire ligand-binding site of FXR in the antagonistic conformation. The extended structure of **9**, in contrast, prevented a similar binding mode and failed to occupy the region in which the *tert*-butylphenol group of NDB is bound.

### 3. Conclusions

With the goal to expand the chemical space of FXR ligands, we implemented an ensemble of machine learning techniques to identify novel FXR modulators. From their consensus, ten



**Figure 5.** Molecular docking of the novel FXR ligands into the FXR ligand-binding domain (PDB-ID 3FXV<sup>[26]</sup> for agonistic compounds **5**, **6**, **8**, and **11**; PDB-ID 4OIV<sup>[27]</sup> for antagonists **7** and **9**). Ligand docking was performed in MOE<sup>[19]</sup> by using the GOLD<sup>[25]</sup> algorithm with flexible receptor. Pocket surfaces are colored according to their lipophilicity (green = lipophilic, red = hydrophilic). The co-crystallized ligands GW4064 pyridine *N*-oxide (PDB-ID 3FXV<sup>[26]</sup>) and NDB (PDB-ID 4OIV<sup>[27]</sup>) are shown in orange for comparison and amino-acid residues are shown in gray. a, b) FXR modulators **5** (a, light blue), **6** (a, purple), and **11** (b, purple) form the characteristic binding mode of FXR agonists and interact with Arg335 through a canonical neutralizing contact. Both compounds entirely occupy the large lipophilic binding site defined by the co-crystallized ligand. c) FXR modulator **8** (purple) forms a binding mode that differs from that of **5**, **6**, **11**, and the co-crystallized ligand. It also interacts with Arg335, but due to its Y-shaped geometry, it protrudes into an additional subpocket. d) FXR antagonists **7** (light blue) and **9** (purple) appear to form no polar contacts with the FXR ligand-binding site, which is also the case for the co-crystallized antagonist NDB. In this model, it is revealed that the binding mode of **7** is similar to that of NDB, whereas **9** fails to occupy one of the three arms of the binding site.

top-ranked compounds were characterized *in vitro*, which confirmed six candidates as new FXR ligands. Despite limited potency, the two novel FXR activator frameworks (5/6/11 and 8) and the two new antagonist frameworks (7 and 9) expand the chemical space of known FXR modulators and contribute to our knowledge of their structure–activity relationship. The results of this study validate ensemble machine learning for prospective hit finding in medicinal chemistry and chemical biology. The data-driven machine learning models implicitly captured pharmacologically relevant features of the known bioactives without the necessity to code chemical knowledge about the underlying structure–activity relationship into the programs.

## Experimental Section

### Computational Methods

#### Training Molecules and Pretreatment

A training library of 896 molecules with annotated FXR activity was compiled from the ChEMBL database v.23.<sup>[9]</sup> Compounds with  $EC_{50} < 10 \mu\text{M}$  were considered active. Records with conflicting experimental values were discarded. The final library consisted of 896 molecules used for model training and is provided in the Supporting Information. Additionally, 17 FXR ligands (Table 2) were selected as queries for ligand-based similarity searching with LIQUID. The training compounds and the search queries were standardized within MOE v.2016.08<sup>[19]</sup> to remove ion pairs, deprotonate strong acids, and protonate strong bases (pH 7). Molecular geometries were optimized by using the MMFF94<sup>[30]</sup> force field (RDKit v.2015.09.2), with 1000 iterations and 10 starting conformers for each compound; the lowest-energy conformer of each molecule was retained.

#### Descriptor Calculation

CATS2 descriptors were calculated with in-house software and default settings (max. topological distance = 10; scaling = “types”). Morgan and FeatMorgan<sup>[18]</sup> binary fingerprints were calculated with RDKit (bit = 1024, radius = 2). AtomPair binary fingerprints were calculated with RDKit with 1024 bit and path lengths be-

tween 1 and 30 bonds (default settings). MOE 2D descriptors were computed with MOE v.2016.08 and default options. WHALES<sup>[20]</sup> were calculated with open-access Python software<sup>[31]</sup> by using Gas-teiger-Marsili<sup>[32]</sup> partial charges as a weighting scheme. LIQUID descriptors were computed with in-house software with default settings. A log *P* and A log *S* were calculated with AlogPS2.1.<sup>[33]</sup>

#### Similarity Searching

The suitability of the descriptors was assessed by retrospective virtual screening. MOE and WHALES descriptors were tested with two types of normalization (Gaussian and MinMax) to rule out the introduction of any bias related to descriptor scaling. Either Jaccard–Tanimoto (Morgan, FeatMorgan, AtomPair fingerprints) or the Euclidean distance (LIQUID, CATS, WHALES, MOE2D) was employed to quantify molecular (dis)similarity. The 17 FXR ligands (Table 2) were selected as queries and used in turn to perform a virtual screening on the training library. For each run, the Enrichment Factor<sup>[34]</sup> of the top 1% ranked list ( $EF_{1\%}$ ) was computed. LIQUID was the best method based on the  $EF_{1\%}$  and, thus, was selected for the prospective similarity searching (LIQUID:  $EF_{1\%} = 1.26 \pm 0.06$ ).

#### Counter-Propagation Artificial Neural Networks (CP-ANN)

WHALES, AtomPair, LIQUID, and CATS descriptors were used to train several CP-ANN using a published MATLAB (v.2017b)<sup>[35]</sup> toolbox module.<sup>[36]</sup> The training compounds were randomly split into training (70%, 608 compounds) and test sets (30%, 261 compounds) by stratified sampling. The resulting training set was used to calibrate several CP-ANN models, with square topology, toroidal, and nontoroidal boundaries and varying numbers of training epochs (75, 100, 200, 300, 500) and neurons of the self-organizing map (10 × 10, 15 × 15, 20 × 20). The best models were selected on the basis of a fivefold crossvalidation (Venetian-blind sampling protocol) on the training set and were then validated on the test set. The best CP-ANN model was toroidal and calibrated on CATS by using 20 × 20 neurons per side and 100 training epochs (classification performance metrics<sup>[36]</sup> on the test set: sensitivity = 91%, specificity = 84%, precision = 93%).

**Table 2.** Query FXR ligands used for similarity searching

```
[O-]C(CC[C@@H](C)[C@@H]1[C@@]2(C)[C@@H](CC1)[C@@H]3[C@@H](O)C[C@H]4C[C@@H](O)CC[C@]4(C)[C@@H]3CC2)=O
O=C(C1=CC(C=CC(N(C)C)=C2)=C2C=C1)N(CC3CCCC3)C4=CC=CC(/C=C/C(OC)=O)=C4
O=C(C1=CC(C=CC(NC)=C2)=C2C=C1)N(CC3CCCC3)C4=CC=CC(/C=C/C(OC)=O)=C4
O=C(C1=CC(C=CC(NC)=C2)=C2C=C1)N(CC3=CC=C(C4=CC=C(N(C)C)C=C4)C=C3)C5=CC=CC(/C=C/C(OC)=O)=C5
O=C(C1=CC(C=CC(N(C)C)=C2)=C2C=C1)N(CC3=CC=C(C4=CC=C(N(C)C)C=C4)C=C3)C5=CC=CC(/C=C/C(OC)=O)=C5
O=C(N(C1=CC(/C=C/C(OC)=O)=CC=C1)CC2=CC=C(C3=CC=C(N(C)C)C=C3)C=C2)C4CCCC4
CC(C)C(ON=[C@]1[C@]2=C(C)C=CC=C2)C=C1COC3=CC=C(/C=C/C4=CC(C([O-])=O)=CC=C4)C(C)=C3
CC(C)C1=C([C@]1([C@]2=C(C)C=CC=C2)NO1)COC3=CC=C(C4=CC(C([O-])=O)=CN5)=C5C=C4)C=C3
C[C@]12[C@@H]([C@@H](C)[C@@H](O)[C@H]3[C@H]2CC[C@]4(C)[C@H]3CC[C@H]4[C@H](C)CC([O-])=O)C[C@H](O)CC1
CC1=CC(OCC2=C(C)ON=[C@]2[C@]3=C(C)C=CC=C3)C=C(C)C1/C=C/C4=CC(C([O-])=O)=CC=C4
C(C=C(OCC1=C(C)ON=[C@]1[C@]2=C(F)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
C(C=C(OCC1=C(C)ON=C1C2=C(OC(F)F)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
C(C=C(OCC1=C(C)ON=C1C2=C(C)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
C(C=C(OCC1=C(C)C)ON=[C@]1[C@]2=C(Br)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
C(C=C(OCC1=C(C)C)ON=[C@]1[C@]2=C(C)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
CC(C=C(OCC1=C(C)C)ON=[C@]1[C@]2=C(C)C=CC=C2)C=C3)=C3/C=C/C4=CC(C([O-])=O)=CC=C4
CC1=CC(OCC2=C(C)C)ON=[C@]2[C@]3=C(C)C=CC=C3)C=C(C)C1/C=C/C4=CC=C(C([O-])=O)C=C4
```

### Nearest-Neighbor Classifier

A nearest-neighbor classifier (kNN) was trained on the same set as the CP-ANN model by using AtomPairs descriptors, which resulted in the second-best method in the retrospective virtual screening ( $EF_{1\%} = 1.23 \pm 0.08$ ). The value of  $k$  ( $k=2$ ) was optimized in three-fold crossvalidation on the training set, which led to a Non-Error Rate<sup>[37]</sup> (NER) equal to 87% (sensitivity = 93%). Compounds predicted as active were ranked according to their Jaccard–Tanimoto distance to the 17 active templates and were scored according to the reciprocal sum of their ranks.

### Commercial Screening Library

A library of 3383942 compounds was assembled from commercially available synthetic compounds from Asinex (ASINEX Ltd., Moscow, Russia; Elite, Fragments, Gold & Platinum collections, downloaded May 2015), ChemBridge screening compound collection (ChemBridge corp., San Diego, CA, USA; downloaded June 2015), Enamine (Enamine LLC, Monmouth, NJ, USA; Advanced and HTS collections, downloaded May 2015), and Specs screening compounds (Specs, Zoetermeer, The Netherlands; downloaded June 2015).

### Prospective Virtual Screening

The commercial library was screened by combining three computational methods: 1) similarity search on AtomPair fingerprints, 2) similarity search on LIQUID descriptors, and 3) CP-ANN predictions (CATS descriptors). The top-500-scoring compounds for each method were combined and sorted according to the reciprocal sum of ranks. The top-ranking compounds containing an acidic functional group (carboxylic acid, aromatic amine, aromatic alcohol, sulfonamide) were visually screened, and ten compounds were selected.

### Molecular Docking

The crystal structures of the human farnesoid X receptor in agonistic (PDB ID: 3FXV)<sup>[26]</sup> and antagonistic (PDB ID: 4OIV)<sup>[27]</sup> conformations were prepared with QuickPrep in MOE2016.08<sup>[19]</sup> by protonating the molecular structure at pH 7, correcting structural issues (missing residues, incorrect hybridization), removing water molecules farther away than 4.5 Å from the receptor or ligand, and restraining the positions of receptor atoms (force constant = 10, buffer = 0.25 Å). The positions of all atoms farther away than 8 Å from the ligand were fixed. Protein and ligand structures were minimized by using the AMBER10:EHT force field (termination value =  $0.1 \text{ kcal} \times \text{mol}^{-1} \times \text{Å}^{-1}$ ). For compound **8**, this procedure yielded no satisfying result, and thus, **8** was prepared by a conformational search by using default parameters in MOE. Ligands were docked within the MOE2016.08 environment by using the integrated GOLD docking program as placement method.<sup>[25]</sup> The active site was defined by the ligand atoms of the co-crystallized ligands. The efficiency of the docking calculation was set to "Very Flexible" (200%) with otherwise default GOLD docking options. The GOLD-score fitness function was used as a scoring method. Thirty poses of each ligand were generated. The "Induced Fit" method was selected for subsequent refinement of the poses by using the standard parameters. The GBVI/WSA dG scoring was chosen as final scoring function. The five best-scoring final poses of each ligand were retained. As a control, redocking of the co-crystallized ligands resulted in poses aligning with the crystal bound structures with

low root-mean-square deviation (RMSD) values (agonist GW4064 pyridine *N*-oxide: RMSD = 0.3343; antagonist NDB: RMSD = 0.2196).

### In Vitro Biological Characterization

#### Hybrid Reporter Gene Assay for FXR

**Plasmids:** The Gal4-fusion receptor plasmid pFA-CMV-hFXR-LBD<sup>[38]</sup> coding for the hinge region and ligand-binding domain (LBD) of the canonical FXR isoform was reported previously. pFR-Luc (Stratagene) was used as reporter plasmid, and pRL-SV40 (Promega) was used for normalization of transfection efficiency and cell growth.

**Assay procedure:** HEK293T cells were grown in Dulbecco's modified Eagle's medium (DMEM) high glucose supplemented with 10% fetal calf serum (FCS), sodium pyruvate (1 mM), penicillin ( $100 \text{ U mL}^{-1}$ ), and streptomycin ( $100 \mu\text{g mL}^{-1}$ ) at 37 °C and 5% CO<sub>2</sub>. The day before transfection, HEK293T cells were seeded in 96-well plates ( $3 \times 10^4$  cells per well). Before transfection, medium was changed to Opti-MEM without supplements. Transient transfection was performed by using Lipofectamine LTX reagent (Invitrogen) according to the manufacturer's protocol with pFA-CMV-hFXR-LBD, pFR-Luc (Stratagene), and pRL-SV40 (Promega). After transfection for 5 h, the medium was changed to Opti-MEM supplemented with penicillin ( $100 \text{ U mL}^{-1}$ ) and streptomycin ( $100 \mu\text{g mL}^{-1}$ ) and additionally containing 0.1% DMSO and the respective test compound or 0.1% DMSO alone as untreated control. Each concentration was tested in triplicate, and each experiment was repeated independently at least three times. Following overnight (12–14 h) incubation with the test compounds, cells were assayed for luciferase activity by using Dual-Glo Luciferase Assay System (Promega) according to the manufacturer's protocol. Luminescence was measured with an Infinite M200 luminometer (Tecan Deutschland GmbH). Normalization of transfection efficiency and cell growth was done by dividing the firefly luciferase data by the *Renilla* luciferase data and multiplying the value by 1000, which resulted in relative light units (RLU). Fold activation was obtained by dividing the mean RLU of a test compound at a respective concentration by the mean RLU of the untreated control. The assay was validated with chenodeoxycholic acid (CDCA), **1**, and **2** as reference agonists, which yielded EC<sub>50</sub> values in agreement with the literature.

### Acknowledgements

Dr. Petra Schneider is thanked for compiling the screening compound library. This research was financially supported by the Swiss National Foundation (grant no. IZSEZO 177477). D.M. was supported by an ETH Zurich Postdoctoral Fellowship (grant no. 16-2 FEL-07).

### Conflict of Interest

G.S. declares a potential financial conflict of interest in his role as life-science industry consultant and cofounder of inSili.com GmbH, Zurich. No further competing interests are declared.

**Keywords:** drug design · drug discovery · neural networks · nuclear receptors · virtual screening

- [1] R. Pellicciari, S. Fiorucci, E. Camaioni, C. Clerici, G. Costantino, P. R. Maloney, A. Morelli, D. J. Parks, T. M. Willson, *J. Med. Chem.* **2002**, *45*, 3569–3572.
- [2] G. M. Hirschfield, A. Mason, V. Luketic, K. Lindor, S. C. Gordon, M. Mayo, K. V. Kowdley, C. Vincent, H. C. Bodhenheimer, A. Parés, M. Trauner, H.-U. Marschall, L. Adorini, C. Sciacca, T. Beecher-Jones, E. Castelleo, O. Böhm, D. Shapiro, *Gastroenterology* **2014**, *148*, 751–761.
- [3] L. Gellrich, D. Merk, *Nucl. Recept. Res.* **2017**, *4*, 101310.
- [4] B. A. Neuschwander-Tetri, R. Loomba, A. J. Sanyal, J. E. Lavine, M. L. Van Natta, M. F. Abdelmalek, N. Chalasani, S. Dasarathy, A. M. Diehl, B. Hameed, K. V. Kowdley, A. McCullough, N. Terrault, J. M. Clark, J. Tonascia, E. M. Brunt, D. E. Kleiner, E. Doo, *Lancet* **2014**, *385*, 956–965.
- [5] S. Mudaliar, R. R. Henry, A. J. Sanyal, L. Morrow, H. U. Marschall, M. Kipnes, L. Adorini, C. I. Sciacca, P. Clopton, E. Castelleo, P. Dillon, M. Pruzanski, D. Shapiro, *Gastroenterology* **2013**, *145*, 574–582.e1.
- [6] P. R. Maloney, D. J. Parks, C. D. Haffner, a. M. Fivush, G. Chandra, K. D. Plunket, K. L. Creech, L. B. Moore, J. G. Wilson, M. C. Lewis, S. A. Jones, T. M. Willson, *J. Med. Chem.* **2000**, *43*, 2971–2974.
- [7] S. M. Soisson, G. Parthasarathy, A. D. Adams, S. Sahoo, A. Sitlani, C. Sparrow, J. Cui, J. W. Becker, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5337–5342.
- [8] B. Flatt, R. Martin, T. L. Wang, P. Mahaney, B. Murphy, X. H. Gu, P. Foster, J. Li, P. Pircher, M. Petrowski, I. Schulman, S. Westin, J. Wrobel, G. Yan, E. Bischoff, C. Daige, R. Mohan, *J. Med. Chem.* **2009**, *52*, 904–907.
- [9] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- [10] L. Breiman, *Mach. Learn.* **1996**, *24*, 123–140.
- [11] R. Hecht-Nielsen, *Appl. Opt.* **1987**, *26*, 4979.
- [12] M. Reutlinger, C. P. Koch, D. Reker, N. Todoroff, P. Schneider, T. Rodrigues, G. Schneider, *Mol. Inf.* **2013**, *32*, 133–138.
- [13] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894–2896; *Angew. Chem.* **1999**, *111*, 3068–3070.
- [14] D. Merk, F. Grisoni, L. Friedrich, E. Gelzinyte, G. Schneider, *MedChemComm* **2018**, *9*, 1289–1292.
- [15] D. Merk, F. Grisoni, L. Friedrich, E. Gelzinyte, G. Schneider, *J. Med. Chem.* **2018**, *61*, 5442–5447.
- [16] Y. Tanrikulu, M. Nietert, U. Scheffer, E. Proschak, K. Grabowski, P. Schneider, M. Weidlich, M. Karas, M. Göbel, G. Schneider, *ChemBioChem* **2007**, *8*, 1932–1936.
- [17] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Model.* **1985**, *25*, 64–73.
- [18] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.
- [19] Chemical Computing Group, Molecular Operating Environment (MOE), Montreal, QC, Canada, **2016**.
- [20] F. Grisoni, D. Merk, V. Consonni, J. Hiss, S. Giani Tagliabue, R. Todeschini, G. Schneider, *Commun. Chem.* **2018**, *1*, 44.
- [21] M. Gabler, J. Kramer, J. Schmidt, J. Pollinger, J. Weber, A. Kaiser, F. Löhr, E. Proschak, M. Schubert-Zsilavec, D. Merk, *Sci. Rep.* **2018**, *8*, 6846.
- [22] E. Proschak, P. Heitel, L. Kalinowsky, D. Merk, *J. Med. Chem.* **2017**, *60*, 5235–5266.
- [23] J. Schmidt, M. Rotter, T. Weiser, S. Wittmann, L. Weizel, A. Kaiser, J. Heering, T. Goebel, C. Angioni, M. Wurglics, A. Paulke, G. Geisslinger, A. Kahnt, D. Steinhilber, E. Proschak, D. Merk, *J. Med. Chem.* **2017**, *60*, 7703–7724.
- [24] G. Papadatos, M. Davies, N. Dedman, J. Chambers, A. Gaulton, J. Siddle, R. Koks, S. A. Irvine, J. Pettersson, N. Goncharoff, A. Hersey, J. P. Overington, *Nucleic Acids Res.* **2016**, *44*, D1220–D1228.
- [25] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.
- [26] S. Feng, M. Yang, Z. Zhang, Z. Wang, D. Hong, H. Richter, G. M. Benson, K. Bleicher, U. Grether, R. E. Martin, J.-M. Plancher, B. Kuhn, M. G. Rudolph, L. Chen, *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2595–2598.
- [27] X. Xu, X. Xu, P. Liu, Z. Y. Zhu, J. Chen, H. A. Fu, L. L. Chen, L. H. Hu, X. Shen, *J. Biol. Chem.* **2015**, *290*, 19888–19899.
- [28] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.
- [29] D. Merk, D. Steinhilber, M. Schubert-Zsilavec, *Future Med. Chem.* **2012**, *4*, 1015–1036.
- [30] T. A. Halgren, *J. Comput. Chem.* **1996**, *17*, 490–519.
- [31] Python package for WHALES descriptors calculation, v.1.0, available at [https://github.com/grisoniFr/whales\\_descriptors.git](https://github.com/grisoniFr/whales_descriptors.git), **2018**.
- [32] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219–3228.
- [33] I. V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, A. S. Makarenko, V. Y. Tanchuk, V. V. Prokopenko, *J. Comput. Aided. Mol. Des.* **2005**, *19*, 453–463.
- [34] J. F. Truchon, C. I. Bayly, *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- [35] The MathWorks Inc., MATLAB 2017a, Natick, MA, **2017**.
- [36] D. Ballabio, V. Consonni, R. Todeschini, *Chemom. Intell. Lab. Syst.* **2009**, *98*, 115–122.
- [37] D. Ballabio, F. Grisoni, R. Todeschini, *Chemom. Intell. Lab. Syst.* **2018**, *174*, 33–44.
- [38] J. Schmidt, F.-M. Klingler, E. Proschak, D. Steinhilber, M. Schubert-Zsilavec, D. Merk, *Sci. Rep.* **2015**, *5*, 14782.

---

 Received: July 30, 2018