COMMUNICATIONS CHEMISTRY

ARTICLE

DOI: 10.1038/s42004-018-0068-1 OPEN

Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators

Daniel Merk [1], Francesca Grisoni [1,2], Lukas Friedrich [1] & Gisbert Schneider [1]

Instances of artificial intelligence equip medicinal chemistry with innovative tools for molecular design and lead discovery. Here we describe a deep recurrent neural network for de novo design of new chemical entities that are inspired by pharmacologically active natural products. Natural product characteristics are incorporated into a deep neural network that has been trained on synthetic low molecular weight compounds. This machine-learning model successfully generates readily synthesizable mimetics of the natural product templates. Synthesis and in vitro pharmacological characterization of four de novo designed mimetics of retinoid X receptor modulating natural products confirms isofunctional activity of two computer-generated molecules. These results positively advocate generative neural networks for natural-product-inspired drug discovery, reveal both opportunities and certain limitations of the current approach, and point to potential future developments.

[1] Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology (ETH), Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland. [2] Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1, IT-20126 Milan, Italy. Correspondence and requests for materials should be addressed to D.M. (email: daniel.merk@pharma.ethz.ch) or to G.S. (email: gisbert.schneider@pharma.ethz.ch)

Drug discovery progressively employs natural products with defined bioactivities as starting points for medicinal-chemistry-driven molecule optimization[1,2]. Bioactive natural products are considered potentially superior to small molecules of purely synthetic origin in terms of their unique geometries and scaffolds[3]. However, the synthesis of structurally intricate natural products may be challenging, rendering structure-activity relationship studies elaborate tasks. The computer-assisted de novo design of natural product mimetics offers a viable strategy to reduce synthetic efforts and obtain natural-product-inspired bioactive small molecules. This strategy has already led to the discovery of various innovative natural product mimetics[4]. Nevertheless, the current computational de novo design methods for generating natural-product-inspired small molecules suffer from several limitations, in particular unsatisfactory scoring of biological activities[5,6].

Contemporary applications of domain-specific artificial intelligence (AI) are currently permeating early drug discovery, including de novo design[5,7]. Recently, we have successfully employed generative AI for the computer-based design of fatty acid mimetics[8] to obtain novel chemotypes of retinoid X receptor (RXR)[9] and peroxisome proliferator-activated receptor (PPAR)[10] modulators[11]. Specifically, we developed a deep recurrent neural network (RNN) model with long short-term memory (LSTM) cells[12] for de novo ligand generation. This ligand-based design approach requires only a small set of known bioactive template structures to implicitly capture relevant structural features for the target(s) of interest. In this present study, we prospectively expand the application of the method to the computational design of novel bioactive mimetics of natural products with RXR modulating activities.

Results

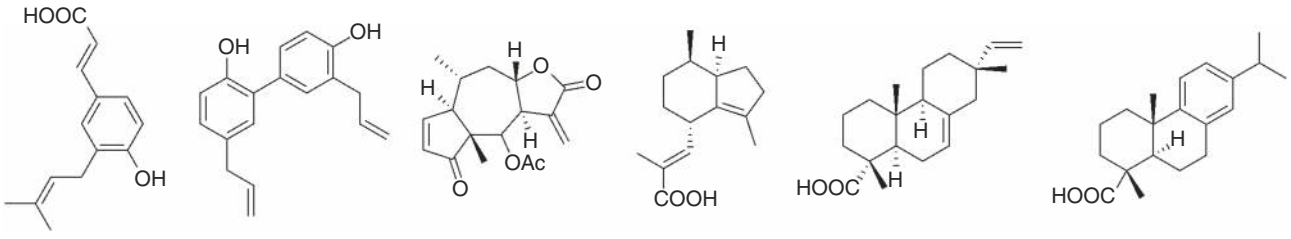
Tuning a generative AI-model on natural products. The deep learning model employed for this study was initially trained to capture the constitution of 541,555 bioactive small molecules retrieved from ChEMBL (K_D, K_i, EC_{50}, $IC_{50} < 1$ μM) represented as simplified molecular input line entry system (SMILES) strings[13]. A key feature of this approach is fine-tuning by transfer learning to bias the de novo molecule generation towards the desired bioactivities of the templates[11]. This fine-tuning step was employed to train the model on designing isofunctional natural product mimetics. Known RXR ligands often suffer from poor receptor subtype selectivity and excessive lipophilicity[4,14–16], and future RXR targeting drug discovery might considerably benefit from natural-product-inspired leads to overcome these liabilities.

Three natural products (Fig. 1, Table 1), namely drupanin (**1**)[17], honokiol (**2**)[18] and bigelovin (**3**)[19] have been known to activate RXRs at micromolar concentrations. We have recently discovered RXR agonistic activity for valerenic acid (**4**), isopimaric acid (**5**) and dehydroabietic acid (**6**) by computer-assisted screening[4]. Notably, valerenic acid (**4**) was found to possess a remarkable and unique preference for the RXRβ subtype. These natural products served as templates for de novo design.

As a first approach to AI-designed natural product mimetics targeting RXR, we fine-tuned our generative model on valerenic acid (**4**) by transfer learning, and then used the biased model to generate 1000 designs as SMILES strings. Out of these 1000 SMILES strings, 25% were chemically valid and 14% were unique (Fig. 2a). The 135 unique and valid designs contained a large proportion (22%) of close structural analogues of the template valerenic acid (**4**), only differing in single methyl groups or ring size. In addition, the model produced numerous instable structures (36%; e.g., carbonic acid monoesters, anhydrides, imines, acetals, antiaromatic structures), fragment-like compounds with molecular weights below 150 g/mol (21%) as well as unsubstituted linear fatty acids and hydrocarbons (14%). None of the remaining nine compounds (7% of the designs) was deemed a potential RXR agonist by SPiDER target prediction software[20]. We concluded from this preliminary experiment that fine-tuning of the generative AI model with a single template structure was insufficient to obtain synthetically accessible and stable isofunctional mimetics.

With the aim to increase the number and quality of the designed compounds, we expanded the set of templates for transfer learning to three templates, namely valerenic acid (**4**), drupanin (**1**) and honokiol (**2**). 1000 de novo SMILES strings were sampled, 79% of which were valid and 49% were unique (Fig. 2b). The fractions of close analogues of the templates (18%), instable structures (15%) and compounds with a molecular weight below 150 g/mol (4%) dropped considerably compared to the results obtained from the first experiment using a single template. However, the new model revealed a tendency to sample long alkyl chains, leading to a large fraction (90 designs, 19%) of unsubstituted fatty acids or pure hydrocarbons. In the remaining 215 compounds (44%), a large fraction of linear alkyl chains ($\geq C_6$) was observed (66 designs, 13%). SPiDER predicted 26 of the 215 stable and synthetically accessible designs as potential RXR agonists with p values < 0.1 (Fig. 3a). These positively predicted RXR modulators were further prioritized using the weighted atom localization and entity shape (WHALES) descriptors, which capture partial charges and 3D molecular shape patterns in a holistic way[21]. Previously, this protocol was successfully employed to identify novel RXR ligands[4,11,14]. Design **7** was ranked in top position according to its WHALES similarity to 12 known RXR binders (for details, see the experimental section), and was selected for synthesis and biological characterization.

After fine-tuning on one and three RXR modulating natural products, respectively, we further expanded the set of template compounds for transfer learning and included all six known natural products that activate RXR, namely valerenic acid (**4**), drupanin (**1**), honokiol (**2**), bigelovin (**3**), isopimaric acid (**5**) and dehydroabietic acid (**6**). Again, 1000 SMILES strings were sampled from this fine-tuned model. Although the fraction of



Drupanin (**1**) Honokiol (**2**) Bigelovin (**3**) Valerenic acid (**4**) Isopimaric acid (**5**) Dehydroabietic acid (**6**)

Fig. 1 Natural products with known modulatory activity on RXRs. Compounds **1**–**6** are known to activate RXRs at micromolar concentrations

valid chemical structures (79%) did not further increase compared to the three-template case, the portion of unique SMILES strings generated was markedly higher (74%, Fig. 2c). Moreover, the fractions of close template analogues (2%) and low molecular weight compounds (<1%) decreased noticeably. Despite the considerable number of instable (25%) and non-functionalized (18%) structures, 491 compounds (54%) appeared to be suitable for synthesis. Of these designs, 201 were predicted as RXR agonists with $p < 0.1$ by SPiDER (Fig. 3b) and ranked using the WHALES descriptors. The 50 best-ranked compounds were visually inspected for synthetic accessibility and building block availability, and the designs **8**–**10** were selected for synthesis and in vitro characterization.

Analysis of AI-designed compounds inspired by natural products. The compounds designed by the AI model possess a different scaffold distribution than the RXR binders from ChEMBL ($EC_{50}/IC_{50} < 50$ µM, $N = 521$)[14], and the NP templates utilized for fine-tuning (Fig. 4). This result highlights the ability of the AI approach to generate innovative molecular cores, thereby exploring novel regions of the chemical space.

The de novo designs were significantly superior compared to the compounds retrieved from ChEMBL in terms of their natural product likeness[22] but less natural-product-like than dictionary of natural products (DNP)[23] entries (Fig. 5; 30,000 randomly selected compounds each, $p < 0.001$, Kruskal-Wallis with post-hoc analysis and Bonferroni correction). These results show that the generative AI model successfully produced molecules that possess features of the synthetic ChEMBL compounds used for model training and the natural products used for transfer learning. The number of molecular targets predicted by SPiDER ($p < 0.05$) for the designed molecules correlates with the number of targets predicted for the respective templates and significantly differs between the sample sets (Fig. 6). Apparently, the generative model captures the bioactivity of the template(s) but a sufficient number of structurally distinct molecules that share a bioactivity is required for fine-tuning on a selected target activity.

Table 1 In vitro potencies of RXR modulating natural products

Natural product	EC_{50} RXRα	EC_{50} RXRβ	EC_{50} RXRγ
Drupanin (**1**)[17]	2.1 ± 0.1 µM	4.6 ± 0.3 µM	7.0 ± 0.3 µM
Honokiol (**2**)[18]	11.8 µM	—	—
Bigelovin (**3**)[19]	4.9 µM	—	—
Valerenic acid (**4**)[4]	27 ± 3 µM	5.2 ± 0.4 µM	43 ± 1 µM
Isopimaric acid (**5**)[4]	26 ± 1 µM	32 ± 1 µM	33 ± 1 µM
Dehydroabietic acid (**6**)[4]	42 ± 3 µM	42 ± 1 µM	42 ± 1 µM

Synthesis of AI-designed compounds. Designs **7**–**10** were prepared according to Fig. 7. Reaction of alkylamine **11** and alkylbromide **12** under microwave irradiation in DMF with triethylamine as base afforded aminoacid **7** in moderate yield. Design **8** was available in a four-step procedure from 6-hydroxyquinazolin-4(3*H*)-one (**13**), which was protected with *tert*-butyldimethylchlorosilane and subsequently reacted with 3-formyl-4-methylphenylboronic acid (**15**) to **16** under adapted Chan-Lam conditions[24], using copper(II)acetate as catalyst and triethylamine as ligand. Microwave irradiation of **16** in pyridine/piperidine in presence of malonic acid produced deprotected cinnamic acid derivative **17** which was treated with acryloyl chloride to afford **8**. Amide coupling of 4-aminophenol (**18**) and 4-iso-propyloxybenzoic acid (**19**) to **20** using EDC*HCl/4-DMAP and subsequent ester formation of phenol **20** and acryloyl chloride produced design **9**. Williamson ether formation between 3,4-dihydroxybenzaldehyde (**21**) and 1-bromodecan (**22**) to **23** followed by Knoevenagel condensation of benzaldehyde **23** with malonic acid in pyridine/piperidine under microwave irradiation yielded design **10**.

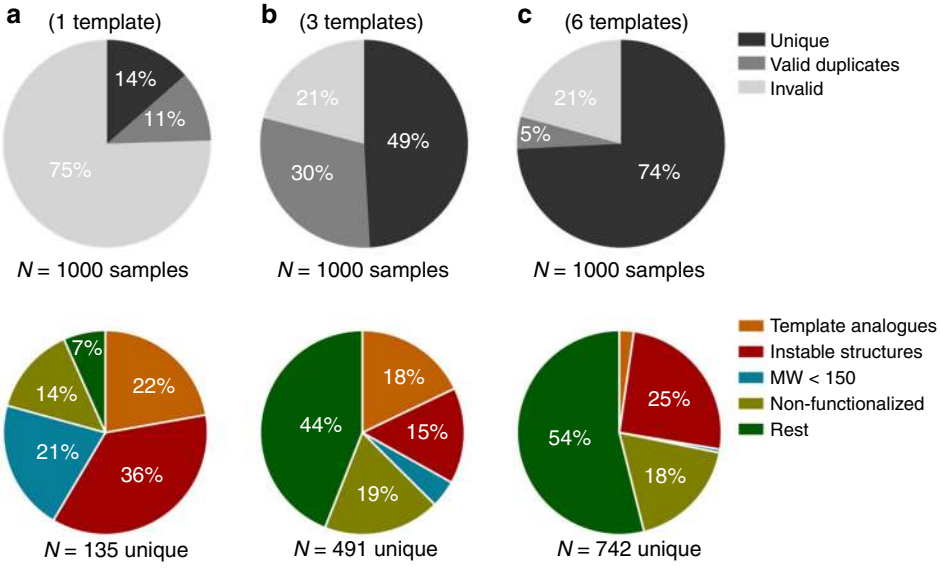


Fig. 2 Characteristics of the designs sampled by differentially fine-tuned AI-models: **a** With only one single natural product (**4**) as template, the model generates only a minor fraction of valid and unique designs. Moreover, close analogues of the template, instable structures, unfunctionalized fatty acids and very small molecules dominate the few unique and valid samples. **b** With an expanded collection of three templates (**1**, **2**, and **4**), the proportion of valid samples is markedly increased but again many duplicate designs are produced and template analogues as well as instable or unfunctionalized structures constitute relevant fractions. **c** With a set of six templates (**1**-**6**) sharing a certain bioactivity, the model generates a favorable proportion of valid structures and duplicates are reduced to a minor fraction. Moreover, close analogues of the individual templates were almost eliminated from the samples when six template structures were u-However, there still is a considerable number of non-functionalized or chemically instable designs

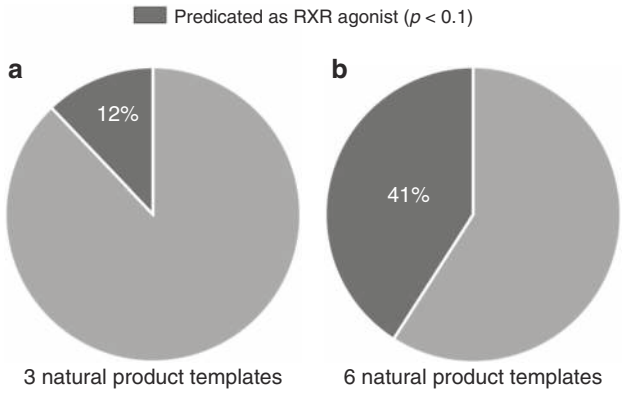


Fig. 3 Predicted activities of sample sets on RXRs: The valid and unique designs obtained from the AI model that was fine-tuned on three natural product templates (**1**, **2**, **4**) were predominantly not recognized by the target prediction software SPiDER as putative RXR modulators (12%, **a**. In contrast, the proportion of the designs predicted as RXR modulators by SPiDER increased considerably for the AI model that had been fine-tuned on six natural product templates (41%, **b**)

Biological characterization of AI-designed compounds. The computationally designed compounds **7**–**10** were characterized in vitro for RXR modulatory potency on all three RXR subtypes up to a concentration of 50 µM in specific hybrid reporter gene assays. These in vitro test systems employ chimera receptors composed of the human ligand binding domain of the nuclear receptor in question and the DNA binding domain of the Gal4 receptor from yeast. A Gal4-responsive firefly luciferase construct served as reporter gene and a constitutively expressed *Renilla* luciferase was used to normalize on transfection efficiency and observe test compound toxicity[25,26]. Designs **7**–**10** were tested for both agonistic and competitive antagonistic activity.

Single-point evaluation at 50 µM revealed design **7** (obtained from model fine-tuning on three templates) and design **8** (fine-tuning on six templates) as inactive on RXRs, whereas the designs **9** and **10** obtained from the latter model were confirmed as RXR agonists (Table 2). Full dose-response characterization revealed double-digit micromolar potency on all three RXR subtypes for compound **9**, without apparent subtype preference and moderate transactivation efficacy. Design **10** possessed full agonistic activity on RXRα and RXRβ with low micromolar EC_{50} values, while its

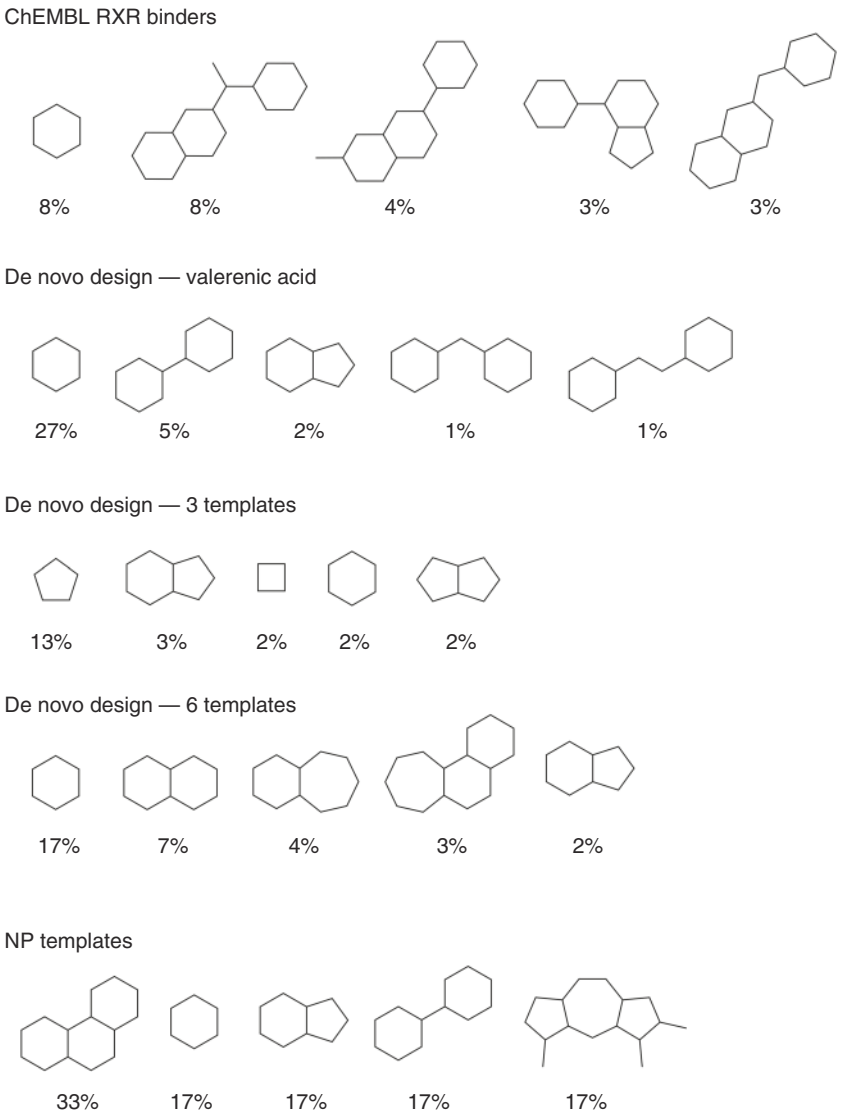


Fig. 4 Scaffold analysis: most frequently occurring graph scaffold amongst ChEMBL RXR binders (EC_{50}/IC_{50} < 50 µM, 521 compounds), NP templates and the designs generated by the LSTM machine learning model. Percentage indicates the frequency of occurrence of the graph scaffold among the considered set

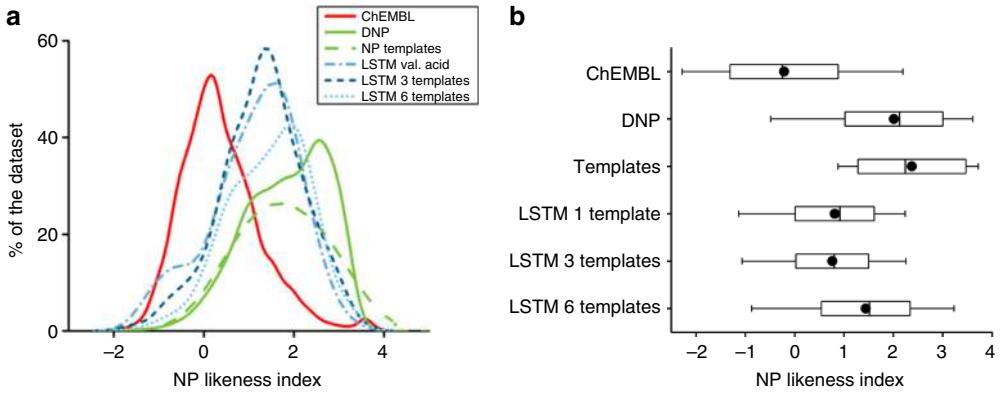


Fig. 5 Comparison between the distribution of the natural-product-likeness index[22] of ChEMBL compounds, DNP compounds, the natural product templates and the AI-generated designs. Distribution of compound sets (**a**) and simplified boxplot representation (**b**). The de novo designs have significantly superior natural-product-likeness than the ChEMBL compounds and are less natural-product-like than the natural products from the DNP ($p < 0.001$, Kruskal-Wallis with post hoc Bonferroni correction; 30,000 entries were randomly selected from ChEMBL and DNP, respectively). Boxplots indicate standard deviation (box), mean (circle), median (solid line) and 1st/99th percentile (whiskers)

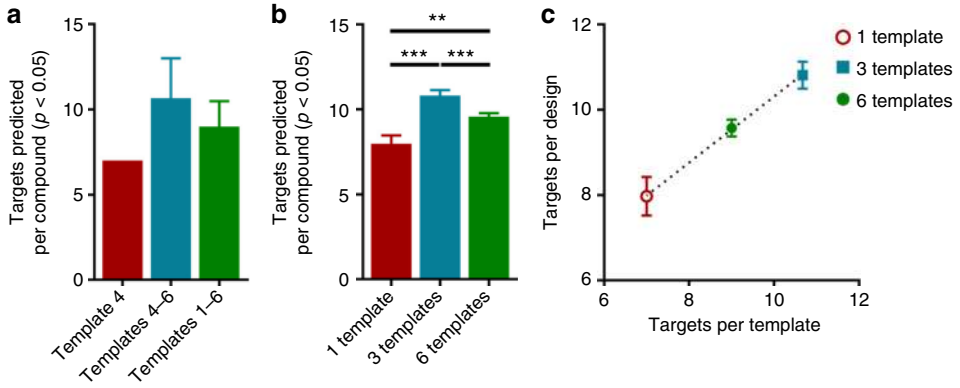


Fig. 6 Distributions of predicted targets for the natural product templates and the computer-generated designs: The average number of targets predicted by SPiDER ($p < 0.05$) for the individual designs (**b**) resembles the predicted target spectrum of the respective template sets (**a**) used for fine-tuning indicating a correlation (**c**) between template and design target spectrum. Data represents *mean* ± SEM; two-sided t-test: $^{**}p < 0.01$, $^{***}p < 0.001$

potency on the RXRγ subtype was markedly weaker. Thus, design **10** possesses distinctive subtype preference for RXRα and RXRβ.

Model evaluation and summary. The experimental results confirm the suitability of the generative AI model for the de novo design of natural product mimetics. A certain number of bioactive templates appears to be required for model fine-tuning to obtain synthetically accessible bioactive mimetics. After fine-tuning on valerenic acid (**4**) as the sole template, the model primarily generated chemically invalid SMILES, and the small fraction of unique and valid entries was dominated by instable structures, close analogues of the template and very small compounds. Thus, fine-tuning on a single natural product template failed to achieve the objective of generating bioactive natural product mimetics. With three templates for fine-tuning, the model performance improved and the percentages of instable structures, close analogues of the templates and very small compounds dropped markedly. However, computational prediction of RXR modulation (SPiDER software) suggested only 12% of the synthetically accessible samples as potential RXR modulators with a p value < 0.1 (10% false-positive estimation, Fig. 3). Notably, not a single design obtained from this model was predicted as RXR agonist with $p < 0.05$. Ranking of the samples using holistic WHALES descriptors, which previously proved useful in the identification of RXR ligands[4], revealed design **7** as the highest ranked candidate for RXR modulation, but synthesis and in vitro characterization failed to confirm activity. The poor activity prediction and the experimentally confirmed inactivity of design **7** suggest that transfer learning with three natural products was insufficient to tune the network model towards designing isofunctional natural product mimetics.

With a set of six natural products sharing a bioactivity on RXRs as templates for fine-tuning, the model not only produced a high proportion of valid, stable and innovative structures but also captured the bioactivity of the templates as indicated by almost 50% of the stable samples being positively predicted as RXR agonists by SPiDER (Fig. 3). This estimation was confirmed experimentally, as two out of three designs selected from the top-50 ranking samples according to the WHALES descriptors possessed RXR agonistic activity in the same potency range as the natural product templates. Thus, with sufficient data available for the crucial target-focused fine-tuning step, the model was able to autonomously generate isofunctional mimetics of the given templates, while conserving their bioactivity on the shared biological target.

To characterize the novelty of the selected designs, we utilized four benchmark fingerprint descriptors for virtual screening[27] (AtomPairs fingerprints[28], Morgan fingerprints[29], RDKit fingerprints, MACCS keys[30]) to determine the structural similarity to

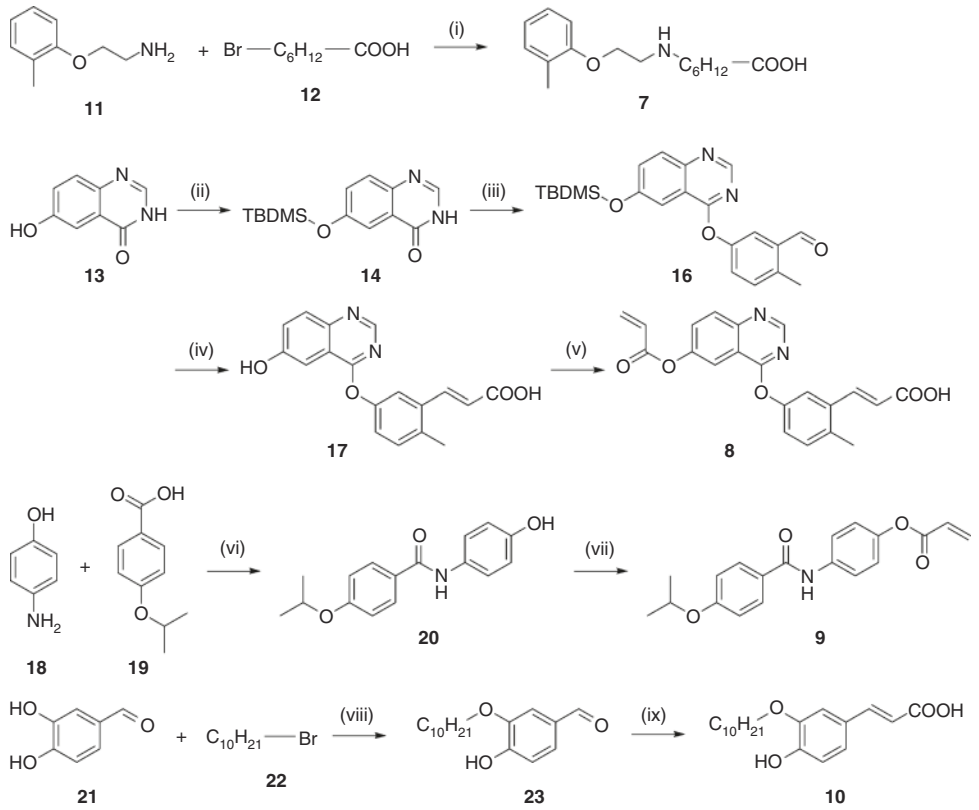


Fig. 7 Synthesis of designs **7**–**10**. Reagents and conditions: (i) DMF, NEt_3, μw, 80 °C, 2 h, 18%; (ii) TBDMS-Cl, DMF, NEt_3, room temperature, 24 h, 74%; (iii) 3-formyl-4-methylphenylboronic acid (**15**), $Cu(OAc)_2$, 4 Å-molecular sieve, CH_2Cl_2, NEt_3, room temperature, 4 h, 94%; (iv) malonic acid, pyridine/piperidine, μw, 100 °C, 30 min., 96%; (v) acryloyl chloride, $CHCl_3$/DMF, NEt_3, room temperature, 2 h, 23%; (vi) EDC*HCl, 4-DMAP, $CHCl_3$, reflux, 16 h, 79%; (vii) acryloyl chloride, THF, pyridine, r.t., 2 h, 66%; (viii) DMF, K_2CO_3, room temperature, 4 h, 39%; (ix) malonic acid, pyridine/piperidine, μw, 100 °C, 30 min., 51%

Table 2 In vitro activity of designs 7–10 on RXRs

ID	EC_{50} (fold activation)		
	RXRα	RXRβ	RXRγ
7	Inactive at 50 µM	Inactive at 50 µM	Inactive at 50 µM
8	Inactive at 50 µM	Inactive at 50 µM	Inactive at 50 µM
9	29 ± 5 µM (10 ± 1)	27 ± 1 µM (11 ± 1)	19.1 ± 0.1 µM (6.1 ± 0.1)
10	16.9 ± 0.6 µM (66 ± 4)	15.7 ± 0.8 µM (59 ± 5)	>50 µM

Data represents mean ± SEM of at least two independent experiments in duplicates.

known binders annotated in ChEMBL (EC_{50}, IC_{50} < 50 µM). The maximum and the average Jaccard-Tanimoto similarity index, which ranges from 0 to 1 (greater values indicate higher molecular similarity), was computed between each designed compound and the known RXR ligands annotated in ChEMBL (Table 3). Designs **7**–**10** revealed a low similarity to known RXR actives, especially in terms of the presence of branched atom-centered fragments (as encoded by Morgan fingerprints), suggesting structural novelty.

Given the limited availability of active templates (sparse data situation), this machine learning approach seems to reach its limit of applicability. The number of known actives required for successfully introducing the desired target-bias into the model may depend on the structural complexity of the actives and their structural difference to the ChEMBL library that was employed for training the basic AI model. With sufficient data, particularly bioactive natural products sharing a common molecular target, this generative model was proven suitable to generate isofunctional de novo natural-product-mimetics. These designs differ significantly from the ChEMBL training data in terms of greater natural-product-likeness. This approach, therefore, holds potential for de novo molecular design not only of bioactive new chemical entities but also in tuning them towards natural product-like properties.

Discussion

AI methods bear potential for early drug discovery and computer-assisted medicinal chemistry. In de novo molecular design, generative machine learning methods, particularly generative neural networks, have been shown competent to autonomously design new chemical entities with inherited bioactivities from the given templates[11]. Here, we have demonstrated that small sets of templates can be sufficient for model fine-tuning on certain target spectra. However, as for virtually all methods in computational de novo molecular design, the generative neural networks employed for this task suffer from the need to score and rank the designs. A suitable and validated predicted bioactivity is crucial for meaningful compound selection from the set of new molecules generated. Although the results of this present study indicate that AI-driven de novo design with sufficient data can provide exceptionally high proportions of actives, the model output appears restricted to the quality of its input, and retrieving samples with bioactivities that exceed the potency of the templates seems unlikely. Despite these apparent limitations, however, the results of this study corroborate the ability of the method to generate synthetically accessible small molecule designs that populate uncharted regions of chemical space at the interface of bioactive natural products and druglike compounds.

Table 3 Similarity of de novo designs 7–10 to ChEMBL bioactives (EC_{50}, IC_{50} < 50 μM)

	AtomPairs		RDKit		Morgan		MACCS	
ID	**average**	**max**	**average**	**max**	**average**	**max**	**average**	**max**
7	0.26	0.37	0.25	0.34	0.14	0.25	0.31	0.52
8	0.39	0.49	0.49	0.63	0.16	0.23	0.35	0.59
9	0.32	0.42	0.35	0.41	0.15	0.24	0.32	0.52
10	0.28	0.43	0.32	0.41	0.17	0.33	0.36	0.70

The Jaccard-Tanimoto similarity index was computed for four types of molecular fingerprints to quantify structural molecular similarity. Both the average and maximum similarity values to the ChEMBL RXR binders are reported for each de novo design.

In contrast to several of the previously published rule-based de novo design approaches[5–7], the deep learning concept presented here is not specifically meant to design mimetics of a single template. The technique requires a set of several template structures that share a common biological target. If the model is fine-tuned on a single template it will sample almost exclusively this template structure and structurally close compounds, which is the effect of the neural network minimizing its error function. When the exact template structure is generated, the error reaches its minimum. Therefore, one cannot expect this method to generate mimetics of a single reference compound without artificially increasing the tolerance of the structure generator. Judging from the preliminary results obtained here, even with three template compounds, the generative model has not reached its full potential.

The attractiveness of this new de novo design approach lies in its ability to generate isofunctional new chemical entities (NCEs) to a set of bioactive small molecules. When natural products are used for the fine-tuning step that introduces the target focus, the model is capable of generating NCEs that populate the chemical space at the border of synthetic bioactive molecules (ChEMBL) and natural products. With these characteristics, generative AI for de novo molecular design has the potential to play a key role at the interface of computer-assisted drug discovery and natural-product-inspired medicinal chemistry.

Methods

Data preparation. Salts and stereochemistry information were removed, and compound structures were represented in their neutral state. Molecular structures were represented as simplified molecular input line entry system (SMILES) strings and converted to canonical SMILES with RDKit (Open-source cheminformatics; http://www.rdkit.org).

Generative machine learning model. All scripts were written in Python (Version 3.6), using RDKit (www.rdkit.org), Tensorflow (v1.2, www.tensorflow.org) and Keras (v2.0, https://keras.io) packages. The generative long short-term memory deep learning model trained on bioactive molecules from the ChEMBL database (ChEMBL22, pAffinity > 6) was used as previously published[12]. The model was re-trained (fine-tuning step) with datasets containing valerenic acid (set 1), valerenic acid, drupanin and honokiol (set 2) or valerenic acid, drupanin, honokiol, bigelovin, isopimaric acid and dehydroabietic acid (set 3). For this fine-tuning step, the model was trained for five epochs. 1000 SMILES strings were sampled from the fine-tuned models with a softmax temperature of 0.75 (see ref. [12] for technical details).

Similarity searching with holistic molecular descriptors. The similarity between the unique and valid molecules generated by the generative model and the sets of known RXR actives was calculated using weighted holistic atom localization and entity shape (WHALES) descriptors[21]. Molecular geometry was optimized using the MMFF94[31] force field with 1000 iterations and 10 starting conformers for each compound with RDKit; the minimum energy conformation was chosen for descriptor calculation. WHALES 3D descriptors were computed with freely-available software (https://github.com/grisoniFr/whales_descriptors), using Gasteiger-Marsili[32] partial charges as weighting scheme. RXR query structures of binders were retrieved from ChEMBL as the 12 most potent annotated ligands according to EC_{50}/K_i. For each dataset, every compound in turn was used as a query to perform similarity ranking on the basis of their Euclidean distance on Gaussian-normalized WHALES descriptor values. The results of the individual virtual screenings on each compound were merged according to the sum of their reciprocal ranks[33]. WHALES reference compounds can be found in Supplementary Data 1.

Self-organizing map consensus for target prediction. The bioactivities of all unique and valid molecules generated by the generative model were predicted with SPiDER software[20]. CATS2 descriptors[34] and the two-dimensional MOE descriptors (The Chemical Computing Group, Montreal, Canada; MOE2016.08; MOE descriptors KNIME node; forcefield: MMFF94*) were calculated for all generated molecules. The SPiDER results were filtered for compounds predicted to be active on RXR with $p < 0.1$. In addition, the number of targets with a predicted activity ($p < 0.05$) was retrieved for all templates and designs (Fig. 5).

Scaffold and similarity analysis. Molecular and graph scaffolds were computed with "RDKit Find Murcko Scaffolds" node in KNIME[35] (v. 3.6.1). Benchmark fingerprints were computed with the "RDKit Fingeprints" node in KNIME[35] v 3.6.1, with default settings (AtomPairs: NumBits = 1024, MinPathLength = 1, MaxPathLength = 30, UseChirality = False, RootedFingerprint = False; RDKit: NumBits = 1024, MinPathLength = 1, MaxPathLength = 7, UseChirality = False, RootedFingerprint = False; Morgan: NumBits = 1024; Radius = 2; UseChirality = False; MACCS: UseChirality = False).

Hybrid reporter gene assays for RXRα/β/γ activation. Gal4 hybrid reporter gene assays were performed as described previously[25,26]. *Plasmids*: The Gal4-fusion receptor plasmids pFA-CMV-hRXRα-LBD[26], pFA-CMV-hRXRβ-LBD[26], and pFA-CMV-hRXRγ-LBD[26] coding for the hinge region and ligand binding domain (LBD) of the canonical isoform of the respective nuclear receptor have been reported previously. pFR-Luc (Stratagene) was used as reporter plasmid and pRL-SV40 (Promega) for normalization of transfection efficiency and cell growth. Assay procedure: HEK293T cells were grown in DMEM high glucose, supplemented with 10% FCS, sodium pyruvate (1 mM), penicillin (100 U/ml) and streptomycin (100 μg/ml) at 37 °C and 5% CO_2. The day before transfection, HEK293T cells were seeded in 96-well plates ($2.5 \cdot 10^4$ cells/well). Before transfection, medium was changed to Opti-MEM without supplements. Transient transfection was carried out using Lipofectamine LTX reagent (Invitrogen) according to the manufacturer's protocol with pFR-Luc (Stratagene), pRL-SV40 (Promega) and pFA-CMV-hRXR-LBD. 5 h after transfection, medium was changed to Opti-MEM supplemented with penicillin (100 U/ml), streptomycin (100 μg/ml), now additionally containing 0.1% DMSO and the respective test compound or 0.1% DMSO alone as untreated control or bexarotene (1 μM) and 0.1% DMSO as positive control. Each concentration was tested in duplicates and each experiment was repeated independently at least two times. Following overnight (12-14 h) incubation with the test compounds, cells were assayed for luciferase activity 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 division of firefly luciferase data by renilla luciferase data and multiplying the value by 1000 resulting 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 untreated control. All hybrid assays were validated with reference agonist bexarotene which yielded EC_{50} values in agreement with literature.

General chemical methods. All chemicals and solvents were reagent grade and used without further purification, unless specified otherwise. All reactions were conducted in oven-dried glassware under argon-atmosphere and in absolute solvents. NMR spectra were recorded on a Bruker AV 400 spectrometer (Bruker Corporation, Billerica, MA, USA). Chemical shifts (δ) are reported in ppm relative to TMS as reference; approximate coupling constants (J) are shown in Hertz (Hz). Mass spectra were obtained on an Advion expression CMS (Advion, Ithaka, NY, USA) equipped with an Advion plate express TLC extractor (Advion) using

electrospray ionization (ESI). High-resolution mass spectra were recorded on a Bruker maXis ESI-Qq-TOF-MS instrument (Bruker). Melting points were determined on a Büchi M-560 (Büchi Labortechnik, Flawil, Switzerland). Compound purity was analyzed by HPLC on a VWR LaChrom ULTRA HPLC (VWR, Radnor, PA, USA) equipped with a MN EC150/3 NUCLEODUR C18 HTec 5 µ column (Machery-Nagel, Düren, Germany) using a gradient (H_2O/MeCN 95:5 + 0.1% formic acid isocratic for 5 min to H_2O/MeCN 5:95 + 0.1% formic acid after additional 25 min and H_2O/MeCN 5:95 + 0.1% formic acid isocratic for additional 5 min) at a flow rate of 0.5 ml/min and UV-detection at 245 nm and 280 nm. All final compounds for biological evaluation had a purity > 95% (area-under-the-curve for UV_{245} and UV_{280} peaks).

Synthesis of 7-((2-(o-Tolyloxy)ethyl)amino)heptanoic acid (7): 2-(2-Methylphenoxy)ethylamine (**11**, 76 mg, 0.50 mmol, 1.00 eq) and 7-bromoheptanoic acid (**12**, 105 mg, 0.50 mmol, 1.00 eq) were dissolved in DMF (abs., 1.0 ml) and triethylamine (abs., 0.2 ml) was added. The mixture was stirred under microwave irradiation at 80 °C for 120 min. The solvents were then evaporated, and the crude product was purified by column chromatography using methylene chloride/methanol (9:1) as mobile phase. The product was then dissolved in methylene chloride and hydrochloric acid (4 M in dioxane, 0.25 ml) was added to precipitate the hydrochloride as colorless solid (28 mg, 18%). Mp (hydrochloride): >400 °C. ^{1}H NMR (400 MHz, D_2O) δ = 1.23–1.38 (m, 5H), 1.47–1.58 (m, 2H), 1.61–1.73 (m, 2H), 2.17 (s, 3H), 2.24–2.34 (m, 2H), 3.07–3.13 (m, 2H), 3.44–3.49 (m, 2H), 4.23–4.28 (m, 2H), 6.90–6.98 (m, 2H), 7.15–7.23 (m, 2H) ppm. ^{13}C NMR (101 MHz, D_2O) δ = 11.88, 25.15, 25.23, 27.56, 33.48, 47.63, 48.04, 52.07, 70.96, 112.09, 131.02, 131.23, 143.00, 155.27, 166.48 ppm. HRMS (ESI+): *m/z* 280.1907 calculated for $C_{16}H_{26}NO_3$, found 280.1907 ($[M + H]^+$).

Synthesis of 6-((*tert*-Butyldimethylsilyl)oxy)quinazolin-4(3*H*)-one (14): 6-Hydroxyquinazolin-4(3*H*)-one (**13**, 1.00 g, 6.17 mmol, 1.00 eq) was dissolved in DMF (abs., 20 ml) and triethylamine (2.0 ml), and TBDMS-Cl (1.20 g, 8.00 mmol, 1.30 eq) were slowly added. The resulting mixture was stirred at room temperature for 24 h. Aqueous hydrochloric acid (1 M, 50 ml) and ethyl acetate (50 ml) were then added, phases were separated, and the aqueous layer was extracted twice with ethyl acetate (2 × 50 ml). The combined organic layers were dried over magnesium sulfate and the residue was reduced to approx. 10 ml in vacuum. The crude product was precipitated by the addition of 50 ml water and recrystallized from hexane to yield the title compound as colorless solid (1.26 g, 74%). ^{1}H NMR (400 MHz, chloroform-*d*) δ = 0.20 (s, 6H), 0.95 (s, 9H), 7.25 (dd, *J* = 8.8, 2.8 Hz, 1H), 7.60 (d, *J* = 2.4 Hz, 1H), 7.61 (d, *J* = 3.4 Hz, 1H), 7.93 (s, 1H), 10.75 (s, 1H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = −4.27, 18.37, 25.77, 114.92, 123.74, 128.89, 129.50, 141.34, 143.76, 155.18, 182.38 ppm. MS (ESI+): *m/z* no molecular ion.

Synthesis of 5-((6-((*tert*-Butyldimethylsilyl)oxy)quinazolin-4-yl)oxy)-2-methylbenzaldehyde (16): **14** (550 mg, 2.00 mmol, 1.00 eq) was dissolved in methylene chloride (abs., 40 ml) and 3-formyl-4-methylphenylboronic acid (**15**, 600 mg, 3.00 mmol, 3.00 eq), molecular sieves (4 Å), triethylamine (2.08 ml, 3.04 g, 30.0 mmol, 15.00 eq) and copper(II)acetate (360 mg, 2.00 mmol, 1.00 eq) were added sequentially. The mixture was stirred at room temperature in an open flask for 4 h. Evaporated solvent was replaced every 60 min. The solvents were then evaporated in vacuum and the crude product was purified by column chromatography using methylene chloride/methanol (98:2) and hexane/ethyl acetate (2:1) as mobile phase. Recrystallization from methanol yielded the title compound as colorless solid (741 mg, 94%). ^{1}H NMR (400 MHz, chloroform-*d*) δ = 0.19 (s, 6H), 0.94 (s, 9H), 2.69 (s, 3H), 7.26 (dd, *J* = 8.7, 2.8 Hz, 1H), 7.39 (d, *J* = 8.1 Hz, 1H), 7.50 (dd, *J* = 8.1, 2.4 Hz, 1H), 7.62 (d, *J* = 8.7 Hz, 1H), 7.66 (d, *J* = 2.8 Hz, 1H), 7.80 (d, *J* = 2.4 Hz, 1H), 7.95 (s, 1H), 10.25 (s, 1H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = −4.42, 18.25, 19.21, 25.63, 115.53, 128.55, 129.25, 129.55, 131.94, 133.11, 135.03, 141.53, 143.26, 143.48, 155.50, 191.13 ppm. MS (ESI +): *m/z* no molecular ion.

Synthesis of (*E*)-3-(5-((6-Hydroxyquinazolin-4-yl)oxy)-2-methylphenyl) acrylic acid (17): **16** (395 mg, 1.00 mmol, 1.00 eq) was dissolved in pyridine (abs., 5.0 ml), malonic acid (105 mg, 1.00 mmol, 1.00 eq) and piperidine (0.5 ml) were added and the mixture was stirred at 100 °C under microwave irradiation for 30 min. After cooling to room temperature, 50 ml 10% aqueous sodium hydroxide solution were added, and the aqueous layer was washed with ethyl acetate (3 × 50 ml). The aqueous layer was then brought to pH 7 by the addition of 1 M aqueous hydrochloric acid and the precipitate was filtered off. The filter residue was washed with methanol (20 ml) and acetone (20 ml) to yield the title compound as colorless solid (309 mg, 96 %). ^{1}H NMR (400 MHz, DMSO-d_6) δ = 2.47 (s, 3H), 6.53 (d, *J* = 15.8 Hz, 1H), 7.38 (dd, *J* = 8.8, 2.8 Hz, 1H), 7.41–7.49 (m, 2H), 7.53 (d, *J* = 2.8 Hz, 1H), 7.62 (d, *J* = 8.8 Hz, 1H), 7.82 (d, *J* = 15.9 Hz, 1H), 7.89 (d, *J* = 1.9 Hz, 1H), 8.22 (s, 1H) ppm. ^{13}C NMR (101 MHz, DMSO-d_6) δ 19.42, 109.91, 122.04, 123.28, 124.51, 125.76, 129.11, 129.19, 131.83, 134.30, 136.56, 138.16, 140.47, 140.88, 144.63, 157.39, 167.82, 207.05 ppm. MS (ESI+): *m/z* 322.9 $(M + H)^+$.

Synthesis of (*E*)-3-(5-((6-(Acryloyloxy)quinazolin-4-yl)oxy)-2-methylphenyl) acrylic acid (8): **17** (65 mg, 0.20 mmol, 1.00 eq) was dissolved in chloroform (abs., 4.0 ml) and DMF (abs., 1.0 ml), triethylamine (0.10 ml) was added and acryloyl chloride (50 µl, 54 mg, 0.60 mmol, 3.00 eq) was slowly added under vigorous stirring. The resulting mixture was stirred at room temperature for 2 h. Methanol (10 ml) was added and the mixture was stirred for another 10 min. The solvents were then evaporated in vacuum and the crude product was purified by column chromatography using methylene chloride/methanol (95:5) as mobile phase. The product was crystallized from hexane/methylene chloride to yield the title compound as pale yellow solid (17 mg, 23%). Mp: 344–348 °C (decomposition). ^{1}H NMR (400 MHz, methanol-d_4) δ = 2.43 (s, 3H), 6.03 (dd, *J* = 10.4, 1.3 Hz, 1H), 6.33 (dd, *J* = 17.3, 10.4 Hz, 1H), 6.40 (d, *J* = 15.9 Hz, 1H), 6.55 (dd, *J* = 17.3, 1.3 Hz, 1H), 7.30–7.39 (m, 2H), 7.59 (dd, *J* = 8.8, 2.7 Hz, 1H), 7.70 (d, *J* = 2.1 Hz, 1H), 7.73 (d, *J* = 8.9 Hz, 1H), 7.89 (d, *J* = 15.9 Hz, 1H), 7.95 (d, *J* = 2.6 Hz, 1H), 8.22 (s, 1H) ppm. ^{13}C NMR (101 MHz, methanol-d_4) δ = 18.08, 118.40, 120.69, 124.83, 127.23, 128.10, 128.48, 128.76, 128.92, 131.53, 131.70, 131.99, 132.48, 132.64, 134.52, 137.38, 142.26, 146.86, 149.71, 162.25, 192.63 ppm. HRMS (ESI +): *m/z* 377.1132 calculated for $C_{21}H_{17}N_2O_5$ found 377.1127 ($[M + H]^+$).

Synthesis of *N*-(4-Hydroxyphenyl)-4-isopropoxybenzamide (20): 4-Aminophenol (**18**, 210 mg, 2.00 mmol, 1.00 eq), 4-isopropyloxybenzoic acid (**19**, 360 mg, 2.00 mmol, 1.00 eq) and 4-DMAP (245 mg, 2.00 mmol, 1.00 eq) were dissolved in $CHCl_3$ (abs., 20 ml) and EDC·HCl (575 mg, 3.00 mmol, 1.50 eq) was added. The mixture was stirred under reflux for 16 h. After cooling to room temperature, hydrochloric acid (1 M, 20 ml) and ethyl acetate (2 ml) were added, phases were separated, and the aqueous layer was extracted twice with ethyl acetate (2 × 20 ml). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using hexane/ethyl acetate (3:1) as mobile phase to yield the title compound as colorless solid (426 mg, 79%). ^{1}H NMR (400 MHz, DMSO-d_6) δ = 1.30 (d, *J* = 6.0 Hz, 6H), 4.73 (hept, *J* = 6.0 Hz, 1H), 6.66–6.78 (m, 2H), 6.96–7.06 (m, 2H), 7.45–7.57 (m, 2H), 7.85–7.95 (m, 2H), 9.24 (s, 1H), 9.84 (s, 1H) ppm. ^{13}C NMR (101 MHz, DMSO-d_6) δ = 22.20, 69.86, 115.37, 119.98, 122.70, 127.27, 129.86, 131.31, 153.96, 160.39, 164.85 ppm. MS (ESI-): *m/z* 270.2 ($[M-H]^-$).

Synthesis of 4-(4-Isopropoxybenzamido)phenyl acrylate (9): **20** (135 mg, 0.50 mmol, 1.00 eq) and was dissolved in THF (abs. 10 ml), pyridine (1 ml) was added and acryloyl chloride (60 µl, 68 mg, 0.75 mmol, 1.50 eq) was added dropwise. The mixture was stirred at room temperature for 2 h. Hydrochloric acid (1 M, 20 ml) and ethyl acetate (20 ml) were then added, phases were separated, and the aqueous layer was extracted twice with ethyl acetate (2 × 20 ml). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using hexane/ethyl acetate (5:1) as mobile phase to yield the title compound as colorless solid (108 mg, 66%). Mp: 172–174 °C. ^{1}H NMR (400 MHz, chloroform-*d*) δ = 1.30 (d, *J* = 6.1 Hz, 6H), 4.57 (hept, *J* = 6.1 Hz, 1H), 5.95 (dd, *J* = 10.5, 1.3 Hz, 1H), 6.25 (dd, *J* = 17.3, 10.4 Hz, 1H), 6.54 (dd, *J* = 17.3, 1.3 Hz, 1H), 6.85–6.92 (m, 2H), 7.03–7.11 (m, 2H), 7.54–7.63 (m, 2H), 7.68 (s, 1H), 7.71–7.78 (m, 2H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = 21.92, 70.14, 115.52, 121.03, 122.03, 126.47, 127.89, 128.89, 132.61, 135.87, 146.77, 161.03, 164.65, 165.16 ppm. HRMS (ESI +): *m/z* 326.1387 calculated for $C_{19}H_{20}NO_4$, found 326.1386 ($[M + H]^+$).

Synthesis of 3-(Decyloxy)-4-hydroxybenzaldehyde (23): 3,4-Dihydroxybenzaldehyde (**21**, 290 mg, 2.10 mmol, 1.05 eq) was dissolved in DMF (abs., 5 ml), potassium carbonate (290 mg, 2.10 mmol, 1.05 eq) and 1-bromodecan (**22**, 442 mg, 2.00 mmol, 1.00 eq) were added and the mixture was stirred at room temperature for 4 h. Hydrochloric acid (1 M, 25 ml) and ethyl acetate (25 ml) were added, phases were separated and the aqueous layer was extracted twice with ethyl acetate (2 × 25 ml). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was purified by column chromatography using ethyl acetate/hexane (1:3) as mobile phase to yield the title compound as a colorless (transparent) solid (216 mg, 39%). ^{1}H NMR (400 MHz, chloroform-*d*) δ = 0.81 (t, *J* = 6.8 Hz, 3H), 1.18–1.33 (m, 10H), 1.35–1.45 (m, 2H), 1.49–1.52 (m, 2H), 1.75–1.83 (m, 2H), 4.06 (q, *J* = 7.2 Hz, 2H), 5.69 (s, 1H), 6.88 (d, *J* = 8.3 Hz, 1H), 7.34 (dd, *J* = 8.3, 2.0 Hz, 1H), 7.37 (d, *J* = 1.8 Hz, 1H), 9.77 (s, 1H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = 14.11, 22.68, 25.93, 29.00, 29.30, 29.53, 31.88, 69.33, 110.87, 114.06, 124.47, 130.49, 146.20, 148.85, 191.01 ppm. MS (ESI+): *m/z* 279.3 ($[M + H]^+$).

Synthesis of (*E*)-3-(3-(Decyloxy)-4-hydroxyphenyl)acrylic acid (10): **23** (139 mg, 0.50 mmol, 1.00 eq) and malonic acid (52 mg, 0.50 mmol, 1.00 eq) were dissolved in a mixture of pyridine (1.0 ml) and piperidine (0.10 ml). The mixture was stirred at 100 °C under microwave irradiation for 30 min. After cooling to room temperature, 10% aqueous hydrochloric acid (25 ml) were added, and the mixture was extracted three times with ethyl acetate (3 × 25 ml). The combined organic layers were dried over magnesium sulfate and the solvents were evaporated in vacuum. The crude product was recrystallized from hexane/ethyl acetate and water/acetone to yield the title compound as pale yellow solid (82 mg, 51%). Mp: 141–143 °C. ^{1}H NMR (400 MHz, chloroform-*d*) δ = 0.81 (t, *J* = 7.0 Hz, 3H), 1.13–1.33 (m, 12H), 1.34–1.43 (m, 2H), 1.76 (quin, *J* = 6.7 Hz, 2H), 4.01 (t, *J* = 6.6 Hz, 2H), 6.22 (d, *J* = 15.9 Hz, 1H), 6.77 (d, *J* = 8.4 Hz, 1H), 6.97 (dd, *J* = 8.4, 2.1 Hz, 1H), 7.09 (d, *J* = 2.1 Hz, 1H), 7.61 (d, *J* = 15.9 Hz, 1H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = 14.11, 22.67, 25.96, 29.08, 29.31, 29.54, 31.89, 69.11, 111.30, 113.14, 114.92, 122.21, 127.53, 146.01, 146.87, 148.32, 171.29 ppm. HRMS (ESI +): *m/z* 321.2060 calculated for $C_{19}H_{29}O_4$, found 321.2059 ($[M + H]^+$).

Data availability

The datasets and code used, generated or analyzed during the current study are available from the corresponding authors on reasonable request.

Received: 17 July 2018 Accepted: 2 October 2018
Published online: 22 October 2018

References

1. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).
2. Chen, Y., De Bruyn Kops, C. & Kirchmair, J. Data resources for the computer-guided discovery of bioactive natural products. *J. Chem. Inf. Model.* **57**, 2099–2111 (2017).
3. Schneider, P. & Schneider, G. Privileged structures revisited. *Angew. Chem., Int. Ed.* **56**, 7971–7974 (2017).
4. Merk, D., Grisoni, F., Friedrich, L., Gelzinyte, E. & Schneider, G. Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J. Med. Chem.* **61**, 5442–5447 (2018).
5. Schneider, P. & Schneider, G. De novo design at the edge of chaos. *J. Med. Chem.* **59**, 4077–4086 (2016).
6. Schneider, G., Funatsu, K., Okuno, Y. & Winkler, D. De novo drug design – ye olde scoring problem revisited. *Mol. Inf.* **36**, 1681031 (2017).
7. Hartenfeller, M. & Schneider, G. De novo drug design. *Methods Mol. Biol.* **672**, 299–323 (2010).
8. Proschak, E., Heitel, P., Kalinowsky, L. & Merk, D. Opportunities and challenges for fatty acid mimetics in drug discovery. *J. Med. Chem.* **60**, 5235–5266 (2017).
9. Germain, P. et al. International union of pharmacology. LXIII. Retinoid X receptors. *Pharmacol. Rev.* **58**, 760–772 (2006).
10. Michalik, L. et al. International union of pharmacology. LXI. Peroxisome proliferator-activated receptors. *Pharmacol. Rev.* **58**, 726–741 (2006).
11. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).
12. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inf.* **37**, 1700111 (2018).
13. Weininger, D. SMILES, a chemical language and information system: 1: introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
14. Merk, D., Grisoni, F., Friedrich, L., Gelzinyte, E. & Schneider, G. Scaffold hopping from synthetic RXR modulators by virtual screening and de novo design. *Med. Chem. Commun.* **9**, 1289–1292 (2018).
15. de Lera, A. R., Bourguet, W., Altucci, L. & Gronemeyer, H. Design of selective nuclear receptor modulators: RAR and RXR as a case study. *Nat. Rev. Drug Discov.* **6**, 811–820 (2007).
16. Vaz, B. & de Lera, Á. Advances in drug design with RXR modulators. *Expert Opin. Drug Discov.* **7**, 1003–1016 (2012).
17. Nakashima, K.-I., Murakami, T., Tanabe, H. & Inoue, M. Identification of a naturally occurring retinoid X receptor agonist from Brazilian green propolis. *Biochim. Biophys. Acta* **1840**, 3034–3041 (2014).
18. Kotani, H., Tanabe, H., Mizukami, H., Makishima, M. & Inoue, M. Identification of a naturally occurring rexinoid, honokiol, that activates the retinoid X receptor. *J. Nat. Prod.* **73**, 1332–1336 (2010).
19. Zhang, H. et al. Structure basis of bigelovin as a selective RXR agonist with a distinct binding mode. *J. Mol. Biol.* **407**, 13–20 (2011).
20. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl Acad. Sci. USA* **111**, 4067–4072 (2014).
21. Grisoni, F. et al. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **1**, 44 (2018).
22. Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *J. Chem. Inf. Model.* **48**, 68–74 (2008).
23. Dictionary of natural products. (Taylor & Francis Group and CRC Press: Boca Raton, FL, U.S. 2011).
24. Lam, P. Y. et al. New aryl/heteroaryl C-N bond cross-coupling reactions via arylboronic acid/cupric acetate arylation. *Tetrahedron Lett.* **39**, 2941–2944 (1998).
25. Schmidt, J. et al. A dual modulator of farnesoid X receptor and soluble epoxide hydrolase to counter nonalcoholic steatohepatitis. *J. Med. Chem.* **60**, 7703–7724 (2017).
26. Flesch, D. et al. Non-acidic farnesoid X receptor modulators. *J. Med. Chem.* **60**, 7199–7205 (2017).
27. Grisoni, F. et al. Matrix-based molecular descriptors for prospective virtual compound screening. *Mol. Inf.* **36**, 1600091 (2017).
28. Carhart, R. E., Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73 (1985).
29. Morgan, H. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
30. MACCS-II, MDL Information Systems Inc, San Leandro, CA, USA, 1987.
31. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**, 490–519 (1996).
32. Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **36**, 3219–3228 (1980).
33. Chen, B., Mueller, C. & Willett, P. Combination rules for group fusion in similarity-based virtual screening. *Mol. Inf.* **29**, 533–541 (2010).
34. Reutlinger, M. et al. Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Inf.* **32**, 133–138 (2013).
35. Berthold, M. R. et al. KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor Newsl.* **11**, 26–31 (2009).

Acknowledgements

This research was financially supported by the Swiss National Science Foundation (Grant IZSEZ0_177477). D.M. was supported by an ETH Zurich Postdoctoral Fellowship (Grant 16–2 FEL-07).

Author contributions

D.M. and L.F. fine-tuned and analyzed the model and sampled the computational designs; D.M. and F.G. computationally ranked the sampled designs and analyzed the chemical space of the sets; D.M. selected, synthesized and characterized the designs and tested the designs in vitro for RXR modulatory activity; D.M. and G.S. supervised the project and wrote the manuscript. All authors approved the final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s42004-018-0068-1.

Competing interests: G.S. declares a potential financial conflict of interest in his role as life-science industry consultant and cofounder of inSili.com GmbH, Zurich. The other authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.