Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis

Author(s):
Button, Alexander; Merk, Daniel; Hiss, Jan A.; Schneider, Gisbert

Publication date:
2019-07

Permanent link:
https://doi.org/10.3929/ethz-b-000377338

Originally published in:
Nature Machine Intelligence 1, https://doi.org/10.1038/s42256-019-0067-7

Funding acknowledgement:
182176 - De novo molecular design by deep learning (SNF)

nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-019-0067-7

Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis

Alexander Button, Daniel Merk, Jan A. Hiss and Gisbert Schneider*

Chemical creativity in the design of new synthetic chemical entities (NCEs) with drug-like properties has been the domain of medicinal chemists. Here, we explore the capability of a chemistry-savvy machine intelligence to generate synthetically accessible molecules. DINGOS (design of innovative NCEs generated by optimization strategies) is a virtual assembly method that combines a rule-based approach with a machine learning model trained on successful synthetic routes described in chemical patent literature. This unique combination enables a balance between ligand-similarity-based generation of innovative compounds by scaffold hopping and the forward-synthetic feasibility of the designs. In a prospective proof-of-concept application, DINGOS successfully produced sets of de novo designs for four approved drugs that were in agreement with the desired structural and physicochemical properties. Target prediction indicated more than 50% of the designs to be biologically active. Four selected computer-generated compounds were successfully synthesized in accordance with the synthetic route proposed by DINGOS. The results of this study demonstrate the capability of machine learning models to capture implicit chemical knowledge from chemical reaction data and suggest feasible syntheses of new chemical matter.

Early drug discovery aims to identify new chemical entities (NCEs) as synthetically accessible and pharmacologically active lead compounds that can be developed into drug candidates[1]. Computer-assisted de novo drug design supports this discovery process, utilizing a broad variety of techniques[2]. However, organic synthesis remains a rate-limiting factor in drug discovery projects[3]. Many de novo design methods therefore rely on predefined compound design rules and building block libraries to enable the automated generation of synthesizable chemical structures[4,5]. Recently, certain machine learning approaches that do not require explicitly coded chemical transformation rules but rely on implicit chemical knowledge representation have been proposed[6–9]. Despite these promising developments, the often challenging synthetic accessibility of the computationally generated molecules remains a limitation of both rule-based and the contemporary 'artificial intelligence'-based structure generators. To enable efficient de novo design of synthetically accessible NCEs, we developed the DINGOS (design of innovative NCEs generated by optimization strategies) approach, which unites certain aspects of rule-based and machine learning-based structure generators (Fig. 1). DINGOS assembles drug-like NCEs through modular and synthetically feasible design schemes. It utilizes a rule-based growing strategy, in which sets of chemical building blocks are combined to form virtual products by employing chemical transformation rules. For the selection of the optimal fusion partners, DINGOS employs a data-driven neural network model trained on the reactions from the US patent and trade office (USPTO) chemical database[10]. In contrast to employing computer-assisted retrosynthetic analysis for virtual product scoring[11–14], this machine learning concept allows us to consider the synthetic feasibility of each step already in the molecule assembly process, which results in a forward-oriented molecular construction process that emulates the approach of a synthetic chemist.

In four prospective applications, de novo designs were automatically generated with DINGOS to mimic structurally diverse approved drugs ('templates'). The structural similarity between the molecular designs and the respective template served as the fitness function driving the algorithm towards the selection of synthesizable NCEs resembling the template. Each run of DINGOS resulted in a population of 300 designs with structural and physicochemical properties similar to the template. An example structure from each design run was successfully synthesized following the reaction schemes suggested by the software, corroborating the design hypothesis.

Results and discussion

DINGOS assembles NCEs through the application of virtual synthetic transformations. These transformation schemes encode all desired chemical logic, thus encouraging synthetic feasibility in the de novo designs. For starting molecule and virtual reaction partner selection, the technique implements a template-based ligand scoring method, meaning that virtual compounds are generated based on their similarity towards a target ligand of interest (template). Thereby, DINGOS integrates the combinatorial approach with machine learning to provide a directed molecular assembly method where compounds are assembled from small molecular components, with the choice of fragments being made by a trained machine intelligence.

Algorithmic concept. DINGOS utilizes an iterative single-step molecule assembly method (Fig. 2). Building blocks are virtually combined to form new molecules, which are then evaluated according to their similarity to the template compound. To quantify similarity, molecules are represented as molecular descriptor vectors that encode certain structural aspects of a molecule, with the distance between descriptor vectors serving as the measure of similarity. The products of such building block combinations serve as the new starting points for the single-step assemblies (molecule growing steps). This process is continued until either the product converges

ETH Zurich, Department of Chemistry and Applied Biosciences, Zurich, Switzerland. *e-mail: gisbert@ethz.ch

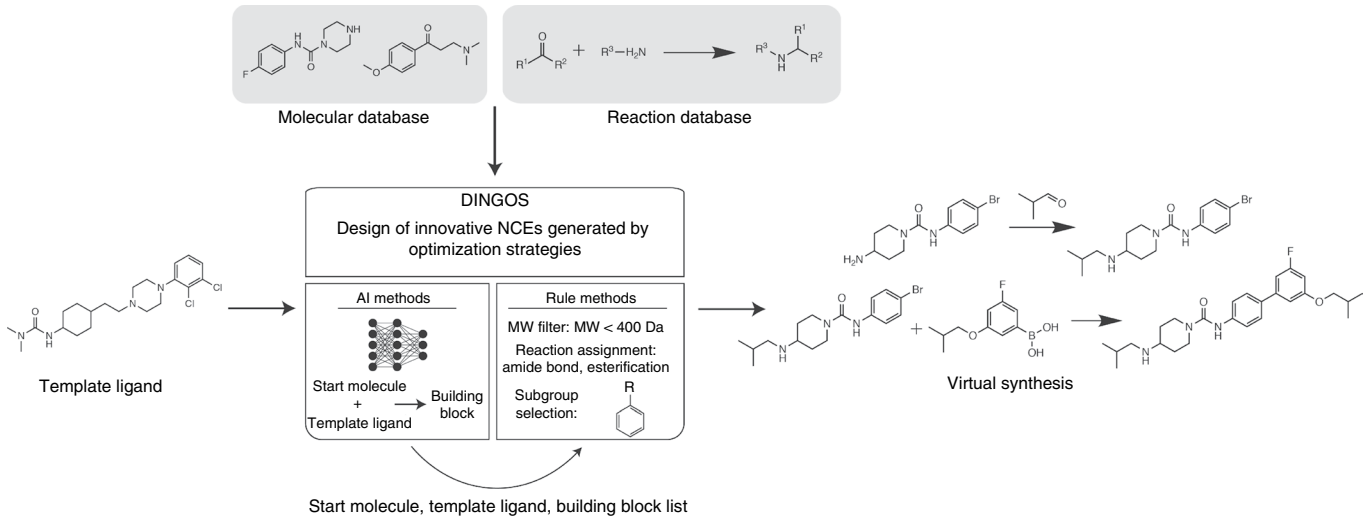

Fig. 1 | Overview of the DINGOS software. DINGOS produces new chemical entities that are structurally similar to a provided template molecule through a hybrid of machine learning and rule-based methodologies. The machine learning model is used for the recommendation of molecular building blocks for virtual chemical synthesis. It was trained to recommend synthons that would produce products that are structurally related to the provided template under the given descriptor representation. The building blocks are then reacted using suitable chemical transformation rules to produce a set of intermediate products. This virtual assembly procedure is repeated until the intermediate products satisfy one or more of the user-defined stop criteria. The products produced are evaluated according to user-defined metrics, and the virtual products best satisfying these criteria are selected as the final molecule designs.

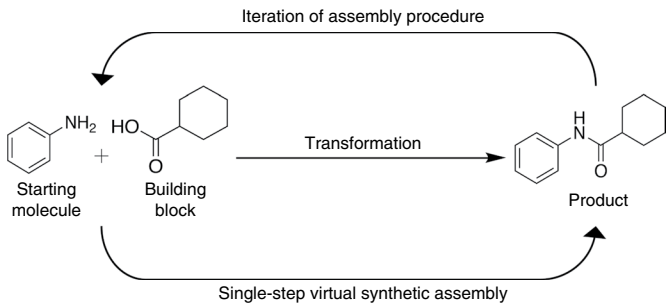

Fig. 2 | Representation of the single-step molecule assembly procedure. A starting molecule is selected based on its similarity to the target ligand of interest. The corresponding building block is then selected by the machine intelligence, and the product molecule is formed in silico. If none of the stop criteria are met, then the product molecule is used as the start molecule for another iteration. This procedure is repeated until either the score converges or one of the stop criteria is met.

to a locally optimal structure with minimal distance to the template (note that depending on the distance metric, multiple optimal structures may exist) or a stop criterion is met. The DINGOS algorithm can be broken down into four main steps (Fig. 3), as follows.

Step 1: Generation of the molecular building block library {mol}. The compound database can be any set of molecules, with the only requirement being that the molecular entries have a valid SMILES[15] representation. De novo drug design by fragment growing involves the assembly of complex molecules from smaller more simplistic components. For this rationale to be reflected within DINGOS, the molecular weight of the building blocks should be considerably smaller than that of the template drug. To ensure this, a molecular weight range is specified and molecules outside of this mass range are not considered during the assembly procedure. Additional filtering criteria based on molecular subgroups or properties can be applied.

Step 2: generation of the set of starting molecules {S}. The molecular descriptor of each member of the molecular building block library {mol} is calculated, and the distance between those building blocks and the template descriptor is evaluated. Building blocks are then sorted according to their increasing distance to the template. A subset of the M closest molecules {S} is then selected from this sorted set. Each element of {S} is used as the starting molecule for an individual assembly procedure. The selection of several unique starting points for each NCE encourages a high degree of structural diversity within the product set and is meant to promote designs with scaffolds that differ structurally from that of the target ligand for the purpose of chemical scaffold-hopping[16–18].

Step 3: construction of optimal intermediates and products P_{opt}. The ith product molecule P_i is formed from the ith element S_i of the start mol set {S}. Thereby, S_i and the template T serve as inputs for the machine learning model M, which takes the descriptor values of S_i and T and predicts a descriptor value. This predicted descriptor corresponds to the building block fingerprint $B^\star$ representing the ideal building block for transforming S_i to T, which has been learned by the model M during training. A distance calculation between $B^\star$ and {mol} is performed, and a subset {B} of the N most similar molecules is produced. All valid chemical transformations between S_i and {B} are applied, generating a set of intermediate products $\{P_1,..,P_k\}$ of size K. The element most similar to T is chosen as the optimal intermediate product P_{opt}. If none of the termination criteria is met (step 4), then P_{opt} is selected as the starting molecule for the next growing step ($S_i = P_{opt}$).

Step 4: termination. The growing of S_i is continued until at least one of the stop criteria is met. There are three conditions under which the construction is halted: (1) the molecular weight of the product exceeds the molecular weight limit, (2) the number of applied reaction steps exceeds that of the reaction step limit and (3) the distance of P_{opt} to the template T is greater than that of the starting molecule S_i. On halting the construction process, the current optimal product P_{opt} is saved as the ith final product ($P_{final} \equiv P_{opt}$) and P_{final} is added to the output product set {P}. In the event of criterion (3) being met, the starting molecule of the current step is saved as the final product ($P_{final} \equiv S_i$) instead of P_{opt}. The current P_{opt} is not considered for any further assembly steps, as it has been shown to be less similar to the target ligand than the starting molecule. Step 3 is then repeated for the next element of {S}, S_{i+1}.

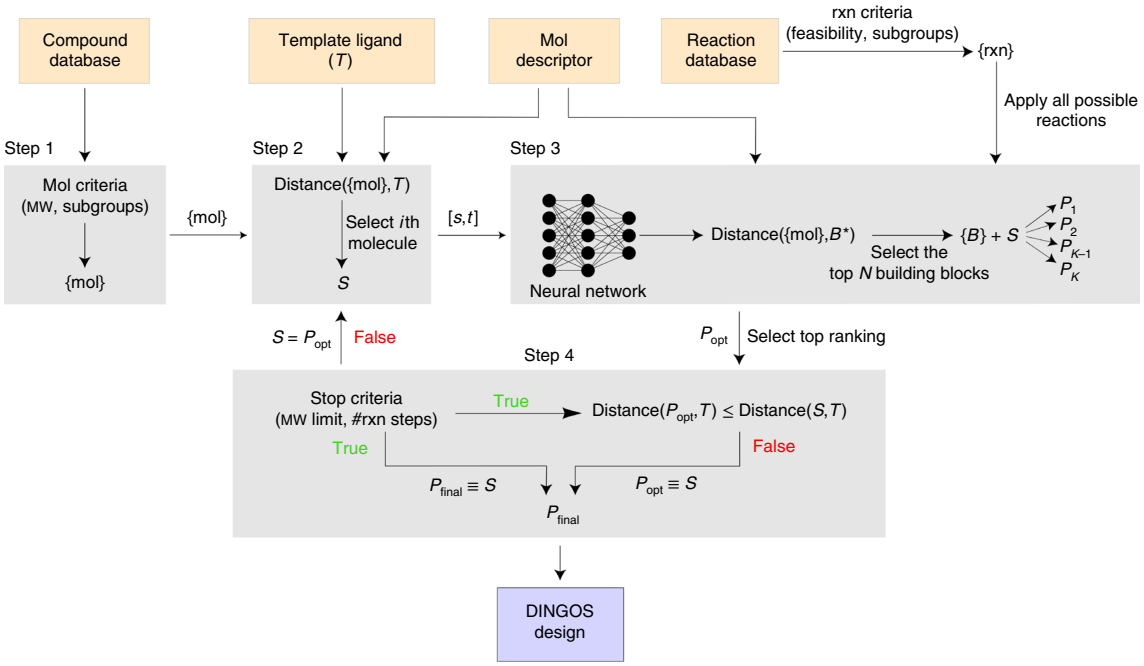

Fig. 3 | Flow chart summarizing the *i*th iteration of the DINGOS algorithm. Orange boxes, inputs; blue box, output. Step 1: the compound database is filtered according to a set of predefined criteria to obtain the set of building blocks {mol}. Step 2: this input molecule set {mol} is sorted according to its distance to the template molecule *T*. The most similar molecule *S* is selected as the starting molecule. The starting tuple [*S*,*T*] serves as the input for the trained machine learning model (here, a feedforward neural network). Step 3: the network predicts the descriptor value of the building block *B**, and the *N* most similar molecules form the building block set {mol}. The set of possible reactions {rxn} between {*B*} and *S* is selected from the reaction database and all reactions are performed, thus generating a set of intermediate products {*P*}. From the intermediate product set, the top-ranking molecule P_{opt} is selected. Step 4: if P_{opt} is more dissimilar to the template *T* than the starting molecule *S*, then the starting molecule is selected as the final product. If P_{opt} is more similar to *T* and none of the stop criteria are met, then P_{opt} is selected as the new start molecule for a further iteration of the DINGOS molecule growing algorithm. Otherwise, if the stop criteria are met, then P_{opt} is selected as the final product P_{final}.

Machine learning model. To assess the capabilities of the DINOGS method, a test predictive model was produced. This model was intended to be naïve, so as to emphasize the influence and capabilities of the DINGOS algorithm, rather than to focus solely on the machine learning component. A multilayered perceptron model (MLP) was used for the building block descriptor prediction. Here, the 167 public MACCS ('molecular access system') substructure keys were used as molecular descriptors (binary fingerprints)[19]. MACCS keys encode local chemical structure (subgraphs) and were selected to capture reaction centres and functional groups, without considering the overall molecular shape or connectivity. The MLP was trained on the US Patent and Trademark Office (USPTO) reaction dataset. After preprocessing (see Methods), the dataset was composed of 897,286 reactions. The MLP was trained to predict the correct fingerprint of each building block from the corresponding start and product molecules' fingerprints. The respective building block was then selected from the molecular database, based on its fingerprint similarity to the predicted, virtually optimal building block fingerprint. Fingerprint similarity was computed as the Hamming distance of the binary fingerprint, that is, the fraction of incorrectly predicted bit positions. This incorporation of reaction-based information into the building block prediction procedure aimed to bias the building block not only to be structurally related to the template, but also to be potentially synthetically compatible with the starting molecule.

DINGOS sorts the available building blocks according to the predicted building block fingerprint. From this sorted set, the top *M* (here, $M=20$) molecules are selected as potentially viable building blocks. All common reactions between the starting molecule and each of these selected building blocks are performed to produce a set of virtual products. The similarity of each of these products to the template is used to rank them. The best-ranking product is selected as the intermediate product and used for further assembly steps.

For MLP training, the Adam optimizer[20] was used with binary cross-entropy as the loss function. The network was trained for 50 epochs with a batch size of 256 at a learning rate of 0.001 without decay. The sigmoid function was chosen for the activation. A network of 334 fan-out input neurons (size of the input space, 2×167 MACCS keys), 167 output neurons and a single hidden layer with 334 neurons was selected by hyperparameter optimization (Supplementary Fig. 3). The network had an average loss of 0.0988 ± 0.0002 (mean ± s.d.) for the training set and 0.1029 ± 0.0006 for the validation set.

Prospective validation. A prospective application was performed to validate DINGOS for practical applications. The goal was to interrogate the functionality of the design objective, namely generating molecules that are (1) synthetically feasible and (2) adhere to a given design hypothesis. A set of suitable reference ligands were selected and de novo design populations were generated using DINGOS with the above described settings. These populations were then analysed in silico, and a selection of designs were chemically synthesized and biochemically tested.

Case study. Four Food and Drug Administration (FDA)-approved compounds (alectinib (**1**), FDA approval, 2015; cariprazine (**2**), FDA approval, 2015; osimertinib (**3**), FDA approval, 2015; pimavanserin (**4**), FDA approval, 2016) were selected as the template compounds. These compounds represent a set of drug molecules

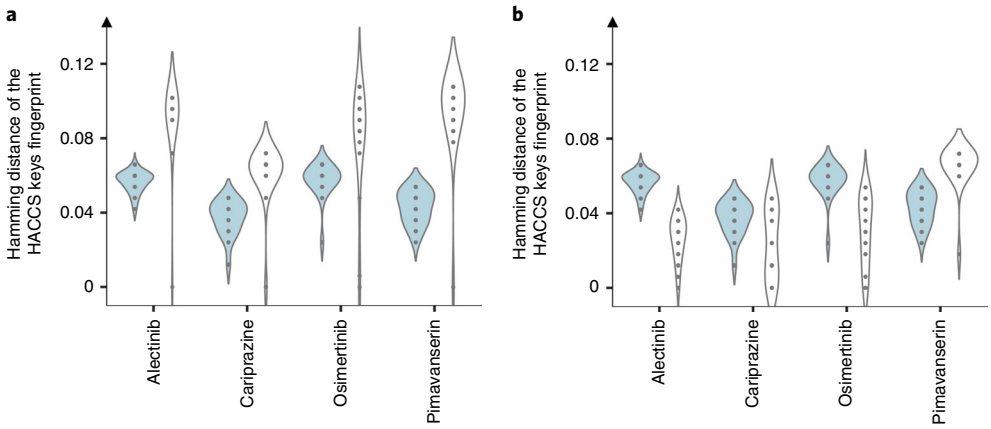

Fig. 4 | Distance comparison of the DINGOS, ChEMBL bioactive and construction sets. a,**b**, Distance comparison of the top 20 ranking DINGOS designs (light blue) against the top 20 ranking molecules of the construction set (white) (**a**) and the ChEMBL database (white) (**b**). Comparisons were made for the four reference compounds (alectinib, cariprazine, osimertinib, pimavanserin) using the MACCS keys Hamming distance as the distance metric. The DINGOS designs were shown to be more similar to the template than the top 20 most similar compounds from the construction set. In comparison with the ChEMBL database, only the pimavanserin DINGOS designs were shown to have a lower median distance. Each plot in the figure represents the estimated probability distribution of the distance values with dots representing the explicit data points. These distributions are mirrored to aid in the readability of the plots.

that are both diverse in structure as well as in their associated biological activity. Each drug was selected as the template compound for an independent de novo design run by DINGOS. During the de novo assembly, the building block set was restricted to molecular weights less than 400 g mol^{-1}. A product limit of 300 compounds and a product molecular weight limit of 600 g mol^{-1} was set, as this represented an upper limit of the templates' molecular weights. The number of reaction steps was set to a maximum of four, and the number of building blocks considered at each assembly step was 20. DINGOS was run successfully, producing four populations of 300 de novo designed molecules per run. All parameters were kept consistent across each run, and all calculations were performed on a single CPU within 1 h.

Distance distributions. A key goal of DINGOS is to produce sets of molecules that are consistent with the design hypothesis. This hypothesis is explicitly implemented into DINGOS by the choice of molecular representation and similarity metric. For this present proof-of-concept study, the MACCS keys fingerprint was chosen, as it is a local structural fingerprint with a relatively low dimensionality, thus reducing the amount of data required for successful training, and it reflects chemical thinking in terms of structural elements and functional groups. The distance values of the de novo generated populations, relative to their respective reference compounds, were calculated, and the populations were ranked according to distance. To determine whether or not DINGOS produced designs with improved similarity towards the target ligands, the distances were compared to those of the initial building block set ('construction set'). It was also of interest how these designs compared to the ChEMBL database (ChEMBL22, 2016)[21,22] as this represents a set of bioactive drug-like compounds. Neither the construction set nor the ChEMBL database is a focused library intended to be structurally related to the target ligands considered here; hence, for a fair comparison to be made, only the top 20 most similar compounds were selected for comparison. A comparative plot of the distance distributions is shown in Fig. 4.

The median distance values of all of the de novo design populations were lower than those of the construction set (Supplementary Table 1). This observation indicates that the algorithm was able to generate compounds that improved upon the initial building block set's similarity to the template compounds. Three of the de novo populations (alectinib, cariprazine, osimertinib) had a median value that was greater than that of ChEMBL. This result is not surprising, considering that the ChEMBL database comprises known bioactive and drug compounds and the top 20 of these were employed for comparison; however, for the pimavanserin set, the median distance value was lower than that of ChEMBL, indicating that for this reference ligand, DINGOS was able to generate compounds that were more similar to pimavanserin than those found in ChEMBL. Overall, DINGOS was shown to be capable of producing structurally focused new molecules through the assembly of less similar building blocks.

Scaffold analysis of the computer-generated molecule designs. DINGOS was designed to produce compounds that are structurally related under a given metric; however, it was also of interest that the compounds have a high degree of inherent scaffold diversity. This is of concern, as we wish to avoid producing near identical structures with only minor structural variations. As our objective function was similarity based, this situation would probably lead to high-performance scores that were not representative of a true novel design. This design objective was encouraged by constructing ('growing') each generated product from a unique starting molecule. To quantify this diversity, the percentage of unique Murcko scaffolds (atom scaffolds) in the DINGOS products was evaluated. The diversity of the DINGOS designs was compared with that of the top 300 distance-ranked ChEMBL and construction set compounds (Table 1). For all four of the template ligands, DINGOS produced compounds of a comparable or greater scaffold diversity than that of the ChEMBL and construction sets.

Physicochemical properties and drug-likeness of the computer-generated designs. The underlying assumption of the DINGOS design hypothesis is rooted in the chemical similarity principle, which states that molecules with a similar structure tend to have similar properties[23]. To investigate the overall drug-like nature of the design populations, we also looked into their pharmacophore similarity. To determine this, Lipinski's original recommendations for orally bioavailable compounds ('rule of 5' properties) were evaluated[24]. To aid in estimating the potential drug-likeness of the designed compounds, the physicochemical properties were also compared to those of a compiled set of bioactive compounds taken from ChEMBL (see Methods). This analysis revealed that both the DINGOS designs and the bioactive compounds agreed well with the

Table 1 | Percentage of unique Murcko scaffolds from the DINGOS, ChEMBL and construction set populations

	DINGOS designs (%)	ChEMBL dataset (%)	Construction set (%)
Alectinib	88	73	63
Cariprazine	43	45	25
Osimertinib	87	61	85
Pimavanserin	59	57	36

All Murcko scaffolds were calculated in RDKit with the GetScaffoldForMol() function.

values of the template compounds, thus confirming the molecular construction process as viable for generating appropriate new scaffolds with the desired properties (Supplementary Fig. 5).

Target prediction. Template-based de novo design ultimately intends to produce compounds that share the biological activity of the template. To virtually evaluate this in the compounds produced, we performed target prediction with the software SPiDER[25], which has repeatedly shown accurate predictions of biological targets. Because SPiDER relies on the CATS[26] topological pharmacophore representation and on MOE[27] physicochemical descriptors, this target prediction intentionally represents a clearly different approach to the MACCS keys-based ranking in our application of DINGOS. For all four sets, more than 50% of the generated molecules had a predicted bioactivity on the molecular target of their template drug. Of particular note is the case of pimavanserin, in which 270 of the 300 de novo designs were predicted to be active. The relatively high proportion, above 50%, of unique scaffolds found in the pimavanserin set indicates that this reported high proportion is not due solely to the reproduction of one privileged scaffold, but rather the overall structural characteristics of the generated molecules.

Synthesis of DINGOS designs. To be practically applicable, DINGOS needs to produce synthetically feasible designs. To investigate this, the de novo populations were reduced to a set of the 10 top-ranking molecules under the distance metric for each template. This enriched set of molecules was further refined by filtering based on the SPiDER activity predictions. All compounds not predicted active were removed. From these final sets, one design from each was selected for synthesis. The designs were chosen based on the availability of the corresponding building blocks. The structures of these designs, along with the corresponding distance ranks and syntheses, are shown in Fig. 5. Compound **5** was obtained by reductive amination of **9** and **10** but turned out very sensitive to light and water and was too instable for further characterization. Reductive amination of **11** and **12** yielded design **6**. Amide coupling of **13** with **14** generated design **7** and reductive amination of **15** and isobutyric aldehyde to **16** followed by Suzuki coupling with boronic acid **17** afforded design **8**.

In vitro pharmacological characterization. To investigate the validity of the stated design hypothesis, compounds **6**, **7** and **8** were evaluated in vitro for potential pharmacological activity on the targets of their respective template (**2**–**4**) at a concentration of 10 μM (Supplementary Fig. 7). Compound **6** revealed no agonistic or antagonistic activity on dopamine receptors D_{2S}, D_{2L} or D_3, and compound **7** was inactive on human epidermal growth factor receptor (EGFR) kinase. Pimavanserin mimetic **8**, however, antagonized 5-HT_{2B} activation by serotonin and the concentration–response characterization suggested dose-dependent partial antagonism.

Conclusions

The DINGOS method incorporates both structural and reaction information into its design considerations and relies on predefined rules as well as machine intelligence-based selection of structural building blocks. In the DINGOS procedure, virtual synthesis is facilitated through a set of predefined reaction rules, while the machine learning method is trained on existing reaction data, with the reactants and products represented as molecular fingerprints. The fingerprints of the start and product molecules serve as inputs, and the model produces a predicted fingerprint, representing the optimal building block molecule to convert the starting fragment into a close mimetic of the template measured by their similarity concerning a molecular descriptor. The building block molecule is then selected from a library of commercial compounds based on their distance to the predicted target fingerprint. The fact that the machine learning method is trained on structure-based reaction information simultaneously promotes the formation of both structurally similar and synthetically feasible designs. The fingerprints used for building block selection are based on a chosen molecular representations (descriptor). This representation can be freely customized to direct the designs towards desired properties of interest. Thus, the customizability of the modular components in DINGOS is an advantageous feature of this technique. Both the reaction and building block set, as well as the machine learning model, can be changed or modified within the algorithm. This modular design allows for a greater degree of control over the chemotypes of designs produced and can tailor the approach towards specific project demands. For example, assembly procedures can be based entirely on in-house available resources, with reactions restricted to only those most amenable for simplified or fully automated synthesis.The advantage of the stepwise assembly procedure of DINGOS over conventional rule-based de novo drug design methods is that each virtual synthetic step incorporates the structural optimization directly and relies on a machine learning model that in turn relies on the wealth of data available from patent data (USPTO chemical reactions). This combination of pre-encoded rules and data-driven, machine intelligence-based building block selection allows for the optimization to occur simultaneously with the compound generation, preventing the need for further structural modifications. The similarity of the output molecules from DINGOS relies on similarity-based building block selection and prediction of the most suitable reactions for building block fusion following the assumption that compounds with similar synthesis also have high similarity themselves.

DINGOS was successfully run for four approved drugs (alectinib (**1**), cariprazine (**2**), osimertinib (**3**) and pimavanserin (**4**)) in a prospective proof-of-concept study. The de novo designed molecules were found to be structurally similar to the reference compounds and in good agreement with their Lipinski rule-of-5 properties.

A de novo population of 300 molecules was generated for each template, and one compound from each de novo population was selected for synthesis and biological testing. Importantly, all four compounds were successfully synthesized with yields ranging from 33% to 76%, and all synthetic procedures were in accordance with those proposed by DINGOS. This preliminary result positively advocates the prospective applicability of the design algorithm.

Of the four compounds synthesized, three were tested in vitro on the respective biological targets of their templates (one compound, the alectinib de novo design, was unstable and hence not capable of being tested). In vitro testing of the three de novo designs showed modest activity on the intended target for one design (pimavanserin), while the other two designs were inactive. These results refute the stated design hypothesis that the MACCS keys similarity is sufficient to produce biologically active designs. Importantly, the synthetic capabilities of DINGOS allowed for the rapid and efficient evaluation of this design hypothesis. In future work, we aim to expand this study to different design problems and hypotheses.

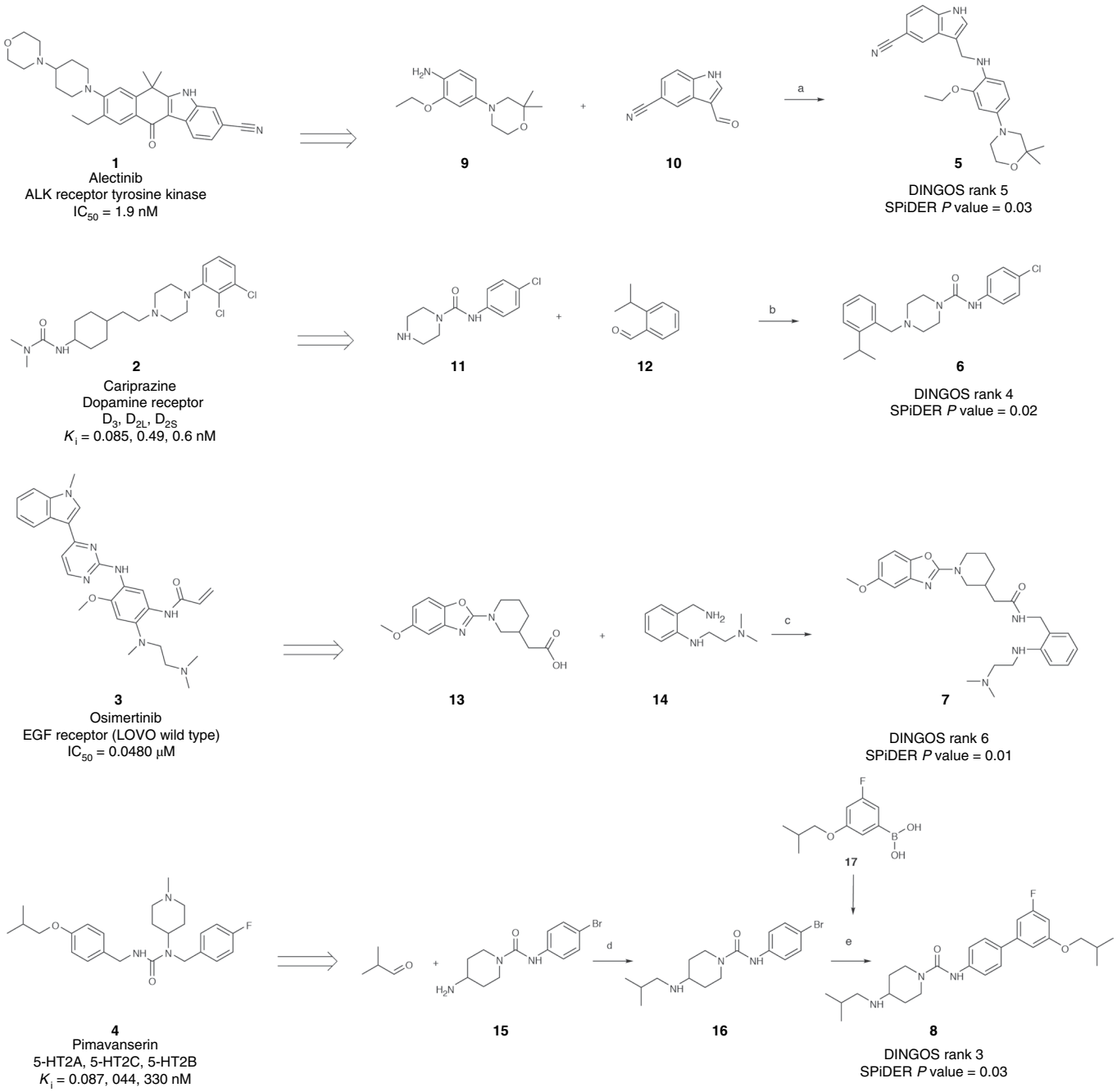

Fig. 5 | Selected de novo designs generated by DINGOS. Compounds were ranked according to the Hamming distance of the MACCS keys fingerprint of the respective template drug and filtered based on their SPiDER predicted activity against the intended targets. All of the selected de novo designs were within the top 10 distances and were predicted as active by SPiDER against the biological target of the respective template. All de novo designs were prepared in one to two synthetic steps from commercially available building blocks. Importantly, all synthetic strategies followed the procedures suggested by the software DINGOS. Synthetic reagents and conditions: [a]DCE (1,2-dichloroethane), $NaB(OAc)_3H$, room temperature, 5 h, product instable; [b]DCE, $NaB(OAc)_3H$, 50 °C, 48 h, 64%; [c]$CHCl_3$, EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide), 4-DMAP (4-(dimethylamino)pyridine), reflux, 2 h, 51%; [d]DCE, $NaB(OAc)_3H$, room temperature, 16 h, 76%; [e]dioxane/DMF, Cs_2CO_3, $Pd(PPh_3)_4$, reflux, 12 h, 33%. IC_{50}, half-maximum inhibitory concentration. K_i, inhibition constant; ALK, anaplastic lymphoma kinase; EGF, epidermal growth factor; LOVO, human LoVo cancer cells; 5-HT, 5-hydroxytrytamine (serotonin) receptor.

The production of large numbers of synthetically accessible drug mimetics provides an efficient means of producing novel bioactive compounds to ensure innovation in early drug discovery. DINGOS has the potential to accelerate existing drug design projects by providing an alternative or complement to traditional high-throughput screening methods and overcoming a key issue of many available computational de novo design techniques, namely limited synthetic accessibility of the designs. As DINGOS has proved itself capable of efficiently producing de novo NCEs, this approach also appears suitable for implementation in automated experimental procedures owing to its simplification of the design strategies and its modular architecture that enables easy updating. Thus, DINGOS shows potential for automating efficient and cost-effective de novo drug design. The modular method design also enables rapid optimization of its individual components, such as replacement of the MACCS keys metric, which might be a reason for the limited number of actives identified in this first prospective application of DINGOS[28]. In the future, DINGOS can be adapted to incorporate multiple

different metrics as a means of tackling complex multi-objective molecular design problems.

Recapitulating, DINGOS fuses rule-based de novo design methods with those of machine learning. This hybrid approach combines the advantages of both expert knowledge and implicit machine intelligence to offer a 'best-of-both-worlds' solution to the problem of molecular design. The empirical rule-based strategy ensures synthesizability and an overall simplification in the compound assembly, while the machine learning methods provide a directed approach to incorporating molecular descriptor information into the designs and limiting the output molecules to compounds with desirable similarity to the template. By producing design populations with high similarity to the given templates, DINGOS confirms our assumption that compound similarity in automated molecular design can be achieved by the similarity of their synthetic routes. The versatility and modular nature of DINGOS allows for the customizability of the method to a large number of potential drug design problems and for fully automated molecular design.

Methods

Chemical transformations. All in silico chemical manipulations were performed using the open-source cheminformatics software RDKit[29] (version 2017.03.3). Reactions were performed using the RunReactants() function. Conversion of the molecules into a canonical SMILES format was achieved with MolToSmiles() and generation of the MACCS keys molecular fingerprints with MACCSkeys(). The template molecules used for the de novo design were washed using the KNIME implementation of the MOE 'wash' method (Molecular Operating Environment, version 2011.10, The Chemical Computing Group).

Molecular building blocks. Suitable, commercially available chemical compounds were identified through Reaxys (www.reaxys.com, version 2018.3.14)[30]. All molecular structures were converted into a standardized canonical SMILES format using RDKit's molecule converter. Salts and minor components were removed as well as all incomplete and inaccurate structures. This yielded a dataset of 245,296 molecules, which were used as the construction set.

Virtual reaction database. The virtual reactions were hand-coded in the SMARTS format. This included mappings of the reactive centres as well as side chain and substituent restrictions. The entire set was composed of 64 reactions and contained both one- and two-component reactions. The relation between a given SMILES molecule and the set of virtual reactions was determined by applying the reactions explicitly. For two-component reactions, the molecule was paired with a generic reactant template corresponding to the other reactant molecule. To determine the molecule's position within a given chemical transformation sequence, the position of the molecule was interchanged with that of the generic reactant. The position resulting in a successfully generated product was recorded as the molecule's reactive position.

Training data. Entries extracted from the USPTO database were used as the training data for the neural network. The initial set was composed of 1.8 million cases. This dataset contained cases that fell outside the bounds of our considered problem set (peptides, large molecules and so on). To remove these cases, the product molecules were filtered based on molecular weight. An upper molecular weight limit of 400 g mol^{-1} was enforced on each of the starting reactants. This ensured that all products were formed from a combination of small molecular building blocks. The same sanitation procedure used for the construction set was applied to remove salts, minor components and erroneous cases. Reactions were filtered by number of reactants; a limit of two reactants per reaction was imposed. To extend the data set, examples were generated in which reactant positions were exchanged. This yielded a dataset of 897,286 examples.

In silico analysis. The molecule sets used for the in silico analysis were prepared with the same procedure used for the construction set. Entries from the ChEMBL dataset that did not have valid activity data were omitted. A molecular weight limit of 1,000 g mol^{-1} was enforced for the two datasets to ensure that only small molecule drug structures were considered. Four sets of compounds were extracted from ChEMBL, each sharing the biological targets of the four template compounds. Only compounds with inhibition constants (K_i values) less than 10 nM were considered.

Target prediction. SPiDER software utilizes self-organizing maps to evaluate the probability of a given molecule to be active on 251 predefined targets. Predictions are reported with *P*values, representing the probability of misclassification. For each population, the *P* value scores were calculated, and compounds with values less than 0.1 were considered as predicted active. SPiDER's predictive accuracy was tested against the four sets of bioactive compounds extracted from ChEMBL that showed a $K_i < 10$ nM against the four corresponding template targets. It was shown capable of predicting the correct targets with an accuracy ranging from 86 to 100% (Supplementary Table 2).

Flexible molecule alignment. Flexible alignment was performed with LigandScout software (version 4.2)[31]. All ligands were prepared in LigandScout. Chemical structures were ionized with the 'ionize acids/base' tool and then minimized with the MMF94 force field. Three-dimensional conformations were generated with the iCON best settings (200 conformations).

General chemistry. All chemicals and solvents were reagent grade and used without further purification, unless specified otherwise. All reactions were conducted in oven-dried glassware under an argon atmosphere and in absolute solvents. NMR spectra were recorded on a Bruker AV 400 spectrometer (Bruker Corporation). Chemical shifts (δ) are reported in ppm relative to TMS (tetramethylsilane) as reference; approximate coupling constants (*J*) are shown in Hz. Mass spectra were obtained on an Advion expression compact mass spectrometer (Advion) equipped with an Advion plate express thin-layer chromatography extractor (Advion) using electrospray ionization (ESI). High-resolution mass spectra were recorded on a Bruker maXis ESI-Qq-TOF-MS (electrospray ionization quadrupole time-of-flight mass spectrometry) instrument (Bruker). Compound purity was analysed by high-performance liquid chromatography (HPLC) on a VWR LaChrom ULTRA HPLC (VWR) system equipped with an MN EC150/3 NUCLEODUR C18 HTec 5 µm column (Machery-Nagel) 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 an additional 25 min and H_2O/MeCN 5:95 + 0.1% formic acid isocratic for an additional 5 min) at a flow rate of 0.5 ml min^{-1} and UV detection at 245 nm and 280 nm. All final compounds for biological evaluation had a purity of >95% (area-under-the-curve for UV_{245} and UV_{280} peaks).

Compound synthesis. For 3-(((4-(2,2-dimethylmorpholino)-2-ethoxyphenyl)amino)methyl)-1*H*-indole-5-carbonitrile (**5**), 4-(2,2-dimethylmorpholino)-2-ethoxyaniline (**9**, 125 mg, 0.50 mmol, 1.00 equiv.) and 3-formyl-1*H*-indole-5-carbonitrile (**10**, 85 mg, 0.50 mmol, 1.00 equiv.) were dissolved in dichloroethane (5 ml), a 4 Å molecular sieve and acetic acid (0.25 ml) were added and the mixture was stirred at room temperature for 60 min. Sodium triacetoxyborohydride (210 mg, 1.00 mmol, 2.00 equiv.) was then added and the mixture was stirred at room temperature for another 4 h. The mixture was filtered, water (25 ml) was added, the phases were separated, and the aqueous layer 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 purified by column chromatography using methylene chloride/methanol 98:2 as the mobile phase to obtain the title compound as yellow oil; MS(ESI+) *m/z* 405.4 ([M + H]$^+$). Compound **5** was sensitive to water and especially light, and was not stable enough for in vitro characterization.

For *N*-(4-chlorophenyl)-4-(2-isopropylbenzyl)piperazine-1-carboxamide (**6**), *N*-(4-chlorophenyl)-piperazine-1-carboxamide hydrochloride (**11**, 138 mg, 0.50 mmol, 1.00 equiv.) and 2-isopropylbenzaldehyde (**12**, 96 mg, 0.65 mmol, 1.30 equiv.) were dissolved in dichloroethane, 4 Å molecular sieve was added and the mixture was stirred at room temperature for 30 min. Sodium triacetoxyborohydride (211 mg, 1.00 mmol, 2.00 equiv.) was slowly added and the mixture was stirred at 50 °C for 48 h. The reaction mixture was then filtered, added to saturated sodium carbonate solution (25 ml), the phases were separated, and the aqueous layer was extracted 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 purified by column chromatography using methylene chloride/methanol (98:2) as the mobile phase to obtain the title compound as a colourless solid (119 mg, 64%). ^{1}H NMR (400 MHz, chloroform-*d*) δ = 1.16 (d, *J* = 6.9, 6H), 2.37–2.44 (m, 4H), 3.28 (hept., *J* = 6.9, 1H), 3.35–3.40 (m, 4H), 3.47 (s, 2H), 7.06 (td, *J* = 7.2, 1.6, 1H), 7.13–7.18 (m, 3H), 7.20–7.28 (m, 4H) ppm. ^{13}C NMR (101 MHz, chloroform-*d*) δ = 24.06, 28.51, 44.29, 52.62, 60.66, 121.08, 125.67, 127.85, 128.04, 128.86, 130.42, 134.21, 137.61, 148.60, 154.68 ppm. MS(ESI+) *m/z* 372.3 ([M + H]$^+$). HRMS(ESI+) *m/z* calculated 372.1837 for $C_{21}H_{27}ClN_3O$, found 372.1841 ([M + H]$^+$). HPLC, retention time: 2.740 min.

For *N*-(2-((2-(dimethylamino)ethyl)amino)benzyl)-2-(1-(5-methoxybenzo[*d*]oxazol-2-yl)piperidin-3-yl)acetamide (**7**), 2-(1-(5-methoxybenzo[*d*]oxazol-2-yl)piperidin-3-yl)acetic acid (**13**, 77 mg, 0.25 mmol, 1.00 equiv.), N^1-(2-(aminomethyl)phenyl)-N^2,N^2-dimethylethane-1,2-diamine (**14**, 58 mg, 0.30 mmol, 1.20 equiv.) and 4-DMAP (31 mg, 0.25 mmol, 1.00 equiv.) were dissolved in chloroform (abs., 10.0 ml) and EDC (47 mg, 53 µl, 0.30 mmol, 1.20 equiv.) was slowly added. The mixture was stirred under reflux for 2 h. After cooling to room temperature, 15 ml saturated sodium carbonate solution was added, phases were separated, and the aqueous layer was extracted with ethyl acetate (2 × 15 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 methylene chloride/methanol (95:5) and acetone/triethylamine (98:2) as mobile phases to obtain the title compound as reddish oil (59 mg, 51%). ^{1}H NMR (400 MHz, chloroform-*d*): δ = 1.34 (td, *J* = 8.8, 4.2, 1H), 1.50–1.69 (m, 2H), 1.81–1.88 (m, 1H), 2.01 (dd, *J* = 13.1, 5.5, 1H), 2.11 (d, *J* = 0.5, 1H), 2.11–2.18 (m, 1H), 2.22 (dd, *J* = 13.1, 7.8, 1H), 2.41 (s, 6H), 2.76 (t, *J* = 6.3, 2H), 3.15

(dd, J=13.2, 7.8, 1H), 3.27 (t, J=6.4, 2H), 3.28–3.35 (m, 1H), 3.71 (s, 3H), 3.74–3.81 (m, 1H), 3.85 (dd, J=13.2, 3.6, 1H), 4.28 (dd, J=14.6, 5.5, 1H), 4.38 (dd, J=14.6, 6.2, 1H), 6.47 (dd, J=8.7, 2.6, 1H), 6.56 (dd, J=8.1, 1.2, 1H), 6.62 (td, J=7.4, 1.1, 1H), 6.69 (d, J=2.5, 1H), 7.00 (d, J=8.7, 1H), 7.09–7.17 (m, 2H) ppm. ^{13}C NMR (101 MHz, chloroform-d): δ=23.56, 29.28, 30.42, 30.93, 32.84, 39.87, 40.60, 41.04, 44.81, 46.35, 50.44, 53.80, 55.96, 57.62, 101.11, 107.00, 108.54, 110.43, 116.92, 129.36, 130.80, 143.14, 143.97, 146.05, 157.01, 163.14 ppm. MS(ESI+) m/z 466.1 ([M+H]$^+$). HRMS(ESI+) m/z calculated 466.2813 for $C_{26}H_{36}N_5O_3$, found 466.2811 ([M+H]$^+$). HPLC, retention time: 15.650 min.

For *N*-(4-bromophenyl)-4-(isobutylamino)piperidine-1-carboxamide (**16**), 4-amino-*N*-(4-bromophenyl)-piperidine-1-carboxamide hydrochloride (**15**, 100 mg, 0.33 mmol, 1.00 equiv.) and isobutyric aldehyde (40 µl, 32 mg, 0.43 mmol, 1.30 equiv.) were dissolved in dichloroethane (5.0 ml), acetic acid (0.50 ml) and 4 Å molecular sieve were added and the mixture was stirred at room temperature for 30 min. Sodium triacetoxyborohydride (95 mg, 0.43 mmol, 1.30 equiv.) was slowly added and the mixture was stirred at room temperature for 16 h. The reaction mixture was filtered, added to saturated sodium carbonate solution (25 ml), the phases were separated and the aqueous layer was extracted 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 purified by column chromatography using methylene chloride/methanol (95:5) as the mobile phase to obtain the title compound as a colourless solid (89 mg, 76%). ^{1}H NMR (400 MHz, DMSO-d_6) δ=0.97 (d, J=6.7, 6H), 1.44–1.59 (m, 2H), 1.90–2.10 (m, 3H), 2.73–2.86 (m, 4H), 4.20 (d, J=13.6, 2H), 7.38–7.43 (m, 2H), 7.43–7.49 (m, 2H), 8.55 (s, 2H), 8.78 (s, 1H) ppm. ^{13}C NMR (101 MHz, DMSO-d_6) δ=20.59, 26.07, 28.33, 42.83, 51.44, 55.42, 121.82, 131.53, 140.45, 144.62, 155.03 ppm. MS(ESI+) m/z 354.2, 356.2 ([M+H]$^+$).

For *N*-(3′-fluoro-5′-isobutoxy-[1,1′-biphenyl]-4-yl)-4-(isobutylamino) piperidine-1-carboxamide (**8**), **16** (65 mg, 0.18 mmol, 1.00 equiv.), 3-fluoro-5-isobutyloxyphenylboronic acid (**17**, 78 mg, 0.37 mmol, 2.00 equiv.) and caesium carbonate (180 mg, 0.55 mmol, 3.00 equiv.) were dissolved in a mixture of dioxane (9.0 ml) and DMF (1.0 ml) and the mixture was stirred for 30 min at room temperature. Tetrakis(triphenylphosphine)palladium(0) (42 mg, 0.04 mmol, 0.20 equiv.) was then added and the mixture was stirred for 12 h under reflux. After cooling to room temperature, the reaction mixture was filtered, water (25 ml) was added and the mixture was extracted 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 purified by column chromatography using methylene chloride/methanol (9:1) as the mobile phase and recrystallized from chloroform/hexane to obtain the title compound as a colourless solid (26 mg, 33%). ^{1}H NMR (400 MHz, methanol-d_4) δ=0.95 (d, J=5.0, 6H), 0.97 (d, J=5.0, 6H), 1.52 (qd, J=12.5, 4.4, 2H), 1.88–2.02 (m, 3H), 2.06–2.12 (m, 2H), 2.83 (d, J=7.2, 2H), 2.85–2.92 (m, 1H), 3.25 (s, 1H), 3.70 (d, J=6.4, 2H), 4.25 (dt, J=13.8, 2.6, 2H), 4.48 (s, 1H), 6.53 (dt, J=10.8, 2.3, 1H), 6.79 (ddd, J=9.9, 2.3, 1.5, 1H), 6.84 (t, J=1.9, 1H), 7.33–7.38 (m, 2H), 7.42–7.46 (m, 2H) ppm. ^{13}C NMR (101 MHz, methanol-d_4) δ=12.38, 17.13, 18.11, 18.76, 18.93, 26.17, 27.96, 28.18, 42.42, 74.45, 87.72, 98.97, 104.83, 116.50, 116.97, 120.70, 126.77, 155.41, 168.24, 171.31 ppm. MS(ESI+) m/z 442.4 ([M+H]$^+$). HRMS(ESI+) m/z calculated 442.2864 for $C_{26}H_{37}FN_3O_2$, found 442.2865 ([M+H]$^+$). HPLC, rentention time: 18.367 min.

Biological evaluation. The synthesized designs **6**–**8** were characterized in vitro for their biological activity on the target of their respective templates (**2**–**4**). All compounds were tested in 10 µM concentration and all assays refer to the human protein of interest. All experiments were conducted as two independent repeats. Compound **6** was studied on dopamine receptors D_{2L}, D_{2S} and D_3. D_{2L} activation and antagonism were studied in a functional assay using membranes containing human recombinant D_{2L} receptors (expressed in Chinese hamster ovary (CHO) cells) wherein binding of radiolabelled [^{35}S]GTPγS was determined. Results represent relative activity compared to 1 mM dopamine. Activity on D_{2S} was assessed in a cell-based (HEK293 (human embryonic kidney) cells) impedance assay and a cellular (CHO cells) homogeneous time resolved fluorescence (HTRF) assay with cyclic adenosine monophosphate (cAMP) readout served for D_3 testing. The inhibitory potency of compound **7** on EGFR was studied on recombinant enzyme (expressed in insect cells) with poly-Glu-Tyr as substrate in the presence of radiolabelled [γ^{32}P]ATP (adenosine triphosphate). Substrate phosphorylation was quantified by scintillation measurements. The activity of compound **8** on serotonin receptors 5-HT_{2A}, 5-HT_{2B} and 5-HT_{2C} was determined in cellular functional assays (HEK293 cells for 5-HT_{2A} and 5-HT_{2C} and CHO cells for 5-HT_{2B}) with detection of IP1 by HTR fluorescence resonance energy transfer. Serotonin receptor activation and antagonism were assessed, and the results represent relative activity compared to 1 µM serotonin. Biological assays were performed by Eurofins (www.eurofins.com) on a fee-for-service basis.

Data availability

The trained machine learning model, CAS numbers of the training data and reaction SMARTS used in this Article are provded in the Code Ocean capsule https://doi.org/10.24433/CO.6930970.v1[32]. All molecules were preprocessed in accordance with the procedure stated in the Methods (see 'Molecular building blocks' section).

Code availability

The code for this Article, along with an accompanying computational environment, are available and executable online as a Code Ocean capsule: https://doi.org/10.24433/CO.6930970.v1[32].

Received: 3 January 2019; Accepted: 24 May 2019;
Published online: 1 July 2019

References

1. Shih, H.-P., Zhang, X. & Aronov, A. M. Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nat. Rev. Drug Discov.* **17**, 19–33 (2017).
2. Hartenfeller, M. & Schneider, G. Enabling future drug discovery by de novo design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 742–759 (2011).
3. Blakemore, D. C. et al. Organic synthesis provides opportunities to transform drug discovery. *Nat. Chem.* **10**, 383–394 (2018).
4. Schneider, P. & Schneider, G. De novo design at the edge of chaos. *J. Med. Chem.* **59**, 4077–4086 (2016).
5. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395 (2013).
6. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
7. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inform.* **37**, 1700153 (2018).
8. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
9. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **1**, 68 (2018).
10. Lowe, D. M. Chemical reactions from US patents (1976–Sep2016) (2017); https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873
11. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
12. Feng, F., Lai, L. & Pei, J. Computational chemical synthesis analysis and pathway design. *Front. Chem.* **6**, 199 (2018).
13. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
14. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
15. 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).
16. Grisoni, F. et al. Scaffold hopping from natural products to synthetic mimetics by holistic molecular similarity. *Commun. Chem.* **1**, 44 (2018).
17. 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).
18. Grisoni, F., Merk, D., Byrne, R. & Schneider, G. Scaffold-hopping from synthetic drugs by holistic molecular representation. *Sci. Rep.* **8**, 16469 (2018).
19. MACCS-II (MDL Information Systems, 1987).
20. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations, ICLR2015*, 1–13 (2015).
21. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
22. *ChEMBL Database* (EBI, 2017); https://www.ebi.ac.uk/chembl/
23. Johnson, M. A. & Maggiora, G. M. *Concepts and Applications of Molecular Similarity* (Wiley, 1990).
24. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
25. 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).
26. Reutlinger, M. et al. Chemically advanced template search (CATS) for scaffold-hopping and prospective target prediction for 'orphan' molecules. *Mol. Inform.* **32**, 133–138 (2013).
27. *Molecular Operating Environment (MOE)* (Chemical Computing Group, 2017).
28. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **8**, 1–14 (2016).
29. *RDKit: Open-source Cheminformatics* (RDKit); www.rdkit.org
30. *Reaxys* (Elsevier).

31. Wolber, G. & Langer, T. LigandScout: 3D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **45**, 160–169 (2005).
32. Button, A., Merk, A., Hiss, J. A. & Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Code Ocean* (2019); https://doi.org/10.24433/CO.6930970.v1

Acknowledgements

The authors thank L. Friedrich, C. Brunner, B. Huisman, X. Zhang and R. Byrne for stimulating discussions and technical support. D.M. was financially supported by an ETH Zurich Postdoctoral Fellowship (grant no. 16–2 FEL-07). This research was financially supported by the Swiss National Science Foundation (grant no. 205321_182176 to G.S.).

Author contributions

A.B. programmed the software and performed the computational experiments. A.B., J.A.H. and G.S. designed the algorithm and analysed the data. D.M. supervised the chemical part of the study and, together with A.B., synthesized the compounds. G.S. designed the study. All authors analysed the results and contributed to the manuscript.

Competing interests

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

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0067-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.S.

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