

De novo Design

Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence**

Michael Moret[†], Moritz Helmstädter[†], Francesca Grisoni, Gisbert Schneider,* and Daniel Merk*

Abstract: Chemical language models enable *de novo* drug design without the requirement for explicit molecular construction rules. While such models have been applied to generate novel compounds with desired bioactivity, the actual prioritization and selection of the most promising computational designs remains challenging. Herein, we leveraged the probabilities learnt by chemical language models with the beam search algorithm as a model-intrinsic technique for automated molecule design and scoring. Prospective application of this method yielded novel inverse agonists of retinoic acid receptor-related orphan receptors (RORs). Each design was synthesizable in three reaction steps and presented low-micromolar to nanomolar potency towards ROR γ . This model-intrinsic sampling technique eliminates the strict need for external compound scoring functions, thereby further extending the applicability of generative artificial intelligence to data-driven drug discovery.

Introduction

Generative deep learning,^[1,2] that is, a class of machine learning models able to generate new data, can be applied to computationally design pharmacologically active compounds *de novo*.^[3–5] Deep learning-based molecular design algorithms can extract high-level molecular features from “raw” molecular representations,^[6–10] such as molecular graphs and the Simplified Molecular Input Line Entry System (SMILES, Figure 1 a),^[11] potentially allowing them to access unexplored regions of the chemical space.^[12] Previous studies showed that chemical language models (CLMs),^[13,14] in particular generative deep learning models trained on SMILES strings, can generate novel molecules with experimentally validated

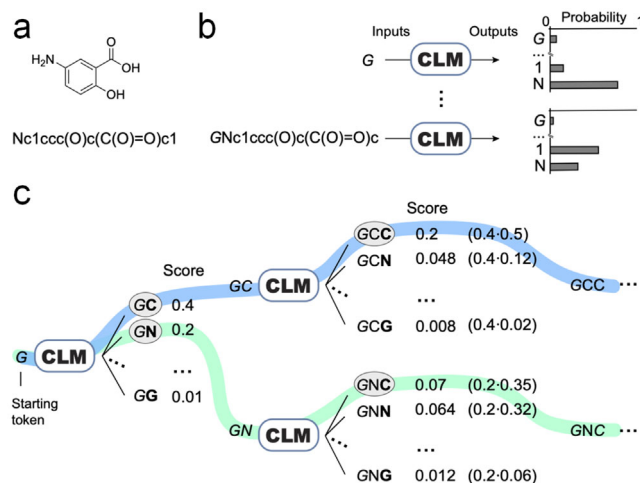


Figure 1. Molecule generation with a chemical language model (CLM) and beam search sampling. a) Kekulé structure of an example molecule and corresponding SMILES string. b) CLM training. The CLM learns to predict the probability of each SMILES string character (“token”) based on the previous tokens in the string. c) Beam search decoding of width two ($k=2$): The design algorithm keeps track of the two most likely SMILES strings (highlighted in color). In this example, the SMILES string generation proceeds from left to right.

bioactivity.^[9,15,16] CLMs have shown the ability to learn focused chemical features from small collections of template molecules by means of transfer learning, that is, a method to reuse previously learned knowledge on a new task for which the available data is scarce.^[15,17,18] Transfer learning is performed in two steps. In the first step, a model is trained

[*] M. Moret,^[†] Prof. Dr. F. Grisoni, Prof. Dr. G. Schneider
 ETH Zurich, Department of Chemistry and Applied Biosciences
 Vladimir-Prelog-Weg 4, 8093 Zurich (Switzerland)
 E-mail: gisbert@ethz.ch

M. Helmstädter,^[†] Prof. Dr. D. Merk
 Goethe University Frankfurt
 Institute of Pharmaceutical Chemistry
 Max-von-Laue-Strasse 9, 60438 Frankfurt (Germany)
 E-mail: merk@pharmchem.uni-frankfurt.de

Prof. Dr. F. Grisoni
 Eindhoven University of Technology
 Institute for Complex Molecular Systems
 Department of Biomedical Engineering
 Groene Loper 7, 5612AZ Eindhoven (Netherlands)

Prof. Dr. G. Schneider
 ETH Singapore SEC Ltd
 1 CREATE Way, #06-01 CREATE Tower
 Singapore 138602 (Singapore)

Prof. Dr. D. Merk
 LMU Munich, Department of Pharmacy
 Butenandtstrasse 7, 81377 Munich (Germany)

[†] These authors contributed equally to this work.

[**] A previous version of this manuscript has been deposited on a preprint server (<http://doi.org/10.26434/chemrxiv.14153408.v1>).

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under:
<https://doi.org/10.1002/anie.202104405>.

© 2021 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

on a large amount of data that relate to the task to be performed (“pre-training”). In the case of CLMs, this is usually done using large collections of molecules (e.g., in the order of 200 000 to 1 000 000^[9,16,17]). Pre-training enables the generative model to capture a) the SMILES “syntax” (i.e., how alphanumeric characters should be assembled to generate strings that correspond to valid molecules, Figure 1) and b) the properties of the pre-training dataset, such as physicochemical features and synthesizability of the molecules in the dataset. In the second step, the pre-trained CLM is further trained (“fine-tuned”) with a smaller set of task-specific molecules.^[13,19,20] During this transfer learning process, the CLM is biased towards the chemical space of interest, that is, molecules with desired biological and physicochemical properties. This ability to learn in a low-data regime (“few-shot” learning^[21,22]) renders CLMs particularly useful for application to biological targets for which only few ligands are known. The fully trained CLM can be used to generate new molecules in the form of SMILES strings. Such data generation is performed by predicting one character of a SMILES string (“token”) at a time, based on all the previous tokens. Importantly, this process does not require handcrafted molecule design rules, as CLMs learn solely from the SMILES strings used for training.

Previous prospective applications of CLMs for de novo molecule generation used the so-called “temperature sampling” to generate large virtual molecular libraries.^[9,13,15] Temperature sampling allows to sample new SMILES strings by adding tokens to the (growing) string according to the probabilities learned by the CLM, wherein the most likely token at a given position will be sampled more often (Figure 1b). However, the generated SMILES strings might not always be “chemically meaningful” (invalid strings), or they might not match the feature distribution of the training data because of the random component of temperature sampling. Therefore, additional methods are usually needed to select the most promising designs from the virtual molecular libraries, e.g., based on the similarity to known bioactive molecules, external activity prediction, or reward functions.^[9,13,15,23] Here, we use the beam search algorithm as a model-intrinsic alternative to temperature sampling. This method enables the CLM to simultaneously generate and prioritize the molecular designs in an automated fashion, without employing additional selection methods.^[24,25] Beam search scoring was successfully validated in a prospective application aiming to generate new retinoic acid-related orphan receptor (ROR)^[26] ligands from scratch.

RORs were chosen as molecular targets because these receptor proteins are an attractive but not extensively studied family of potential drug targets. They constitute a family of ligand-activated transcription factors that mainly act as monomers and are involved in the circadian control of energy homeostasis^[27,28] and immune system regulation,^[29,30] among other functions. RORs hold promising pharmacological potential for various indications, specifically for autoimmune diseases.^[29,30] No ROR ligand has reached drug approval to date, partially owing to compound-related issues such as poor aqueous solubility, lack of selectivity, and clinical safety concerns.^[29,31,32]

Results and Discussion

Chemical Language Model and Beam Search Sampling for De Novo Design

We explored the beam search algorithm^[33] to generate molecules from a CLM as a potential alternative to temperature sampling combined with an external ranking method. Given the probabilities learnt by a CLM, a vast number of SMILES strings could in theory be sampled. As it is computationally not feasible to sample all outputs, a heuristic method such as beam search can be used to find the likely outputs. Here, our underlying hypothesis was that the probability for generating a certain SMILES string correlates with the quality of the corresponding molecule regarding the implicit design objective as represented in the fine-tuning set (e.g., desired bioactivity, physicochemical properties). During molecule generation by beam search sampling, the algorithm progressively adds tokens to a SMILES string while keeping track of the *k* most likely SMILES string(s). To add a new token, the algorithm computes the conditional probability of each possible token given the tokens in the existing string and defines the *k* most likely tokens to extend the string (Figure 1c). The set of *k* most likely selections is based on a scoring function (“beam search score”), which is computed as the product of the probabilities of each token (Figure 1c). This process is repeated until the SMILES string is completed (i.e., the “end-of-string” token is added) or a predefined maximal string length is reached. Thus, beam search can be used to generate highly probable molecules, as computed by (i) the underlying model and (ii) the beam search score. The beam search score allows to rank the de novo designs according to the probability of their SMILES tokens.

As a framework to probe beam search sampling, we employed a recently published CLM based on a recurrent neural network with long short-term memory cells (LSTM), which are suited for sequence modeling.^[34] The CLM was trained with the SMILES strings of 365 063 molecules from ChEMBL^[35] to iteratively predict the next token of each SMILES string given the preceding tokens (Figure 1b). The training procedure was carried out over ten epochs, meaning that each molecule used for training was seen by the CLM ten times. This pre-trained CLM was then fine-tuned using sets of known ROR ligands (Figure S1, Table S1), to obtain a bias towards the design objective, namely the generation of new molecules with bioactivity on RORs, by transfer learning. Open-source code for the CLM and the beam search algorithm, and the data used in this study are available at https://github.com/ETHmodlab/molecular_design_with_beam_search.

Application of Beam Search Sampling to Designing Inverse ROR γ Agonists

For prospective evaluation, we applied the beam search to the design of natural product-inspired ROR γ ligands. Learning from natural products as a traditional source of inspiration for drug discovery^[36,37] may hold several advantages over



learning from purely synthetic molecules, because of the overall higher structural diversity, greater three-dimensionality, and often superior selectivity of bioactive natural products.^[38,39] We aimed to obtain de novo designs possessing three properties: (i) natural product-inspired chemical structure, (ii) synthesizability, and (iii) bioactivity on ROR γ . Aiming to fulfil all three objectives during transfer learning, the previously pre-trained CLM on bioactive molecules from ChEMBL^[17] was fine-tuned on one synthetic and four natural product ROR γ modulators described in literature^[30] (Figure S1). From the fine-tuned model, beam search sampling was started after the fifth epoch of fine-tuning, to ensure that the CLM had sufficiently captured the molecular features of the small fine-tuning set.

All valid SMILES strings generated between epochs 5 and 16 (last fine-tuning epoch) were ranked by beam search scoring. The top five designs according to the beam search score (Figure 2a) were deemed synthetically inaccessible by medicinal chemists. This was further highlighted by the predictions of a machine learning algorithm for retrosynthetic analysis (IBM RXN)^[40] which did not find a synthetic route for any of these molecules. Thus, while the CLM captured natural product likeness, the model failed to meet the generic design criterion of synthesizability. These findings point to a benefit of beam search sampling in revealing the most likely CLM molecules to assess the success of fine-tuning in terms of the design objectives.

Aiming to improve upon these results, we performed a second experiment in which we applied a two-step fine-tuning strategy. First, the pre-trained model was fine-tuned for 20 epochs on 255 synthetic ROR γ ligands from the US patent subset of the Protein Data Bank^[41] (255 molecules, Table S1) to capture both bioactivity and synthesizability. Then, the model was fine-tuned with four natural product ROR γ modulators^[30] (Figure S1) for 16 epochs, aiming to bias

the model towards natural-product-likeness. Again, valid SMILES strings generated by beam search sampling between epochs 5 and 16 of the (second) fine-tuning step were considered. With this second approach, the top 5 sampled molecules (Figure 2b) were synthetically accessible according to IBM RXN,^[40] which could propose a synthetic route for each of them. Importantly, the computer-generated molecules possess certain natural product characteristics (Figure 3, Table S2), as indicated by a high fraction of sp³-hybridized carbon atoms (Fsp³). The top five designs have Fsp³ values ranging from 50% to 75%. These values are comparable to those computed for the MEGx natural product library (Analyticon Discovery GmbH, rel. 09-01-2018), and exceed the average Fsp³ value of the ChEMBL molecules used for pre-training ($51 \pm 30\%$ and $33 \pm 20\%$, respectively). These results suggested that the two-step fine-tuning procedure complied with the design objectives and the implemented two-step approach was chosen for prospective application.

We then compared the beam search designs obtained with the chosen computational strategy to known ROR γ modulators and to the fine-tuning molecules (Figure 3a,b). Despite favoring only some of the most likely tokens while generating new SMILES strings, and examining only a limited set of possibilities, the beam search sampling still allowed to explore the chemical space beyond the regions that are populated by the fine-tuning compounds (Figure 3a). Compared to the inverse ROR γ agonists annotated in ChEMBL ($IC_{50} < 1 \mu\text{M}$, Figure 2d), the beam search designs are structurally more diverse in terms of substructure fragments, as represented by Morgan fingerprints.^[42] Still, the designs possess a certain degree of similarity to the known active molecules in terms of their three-dimensional shape and partial charge distribution (as represented by the Weighted Holistic Atom Localization and Entity Shape [WHALES] descriptors^[43,44]). Apparently, the CLM, in addition to learning the SMILES “syntax”, also

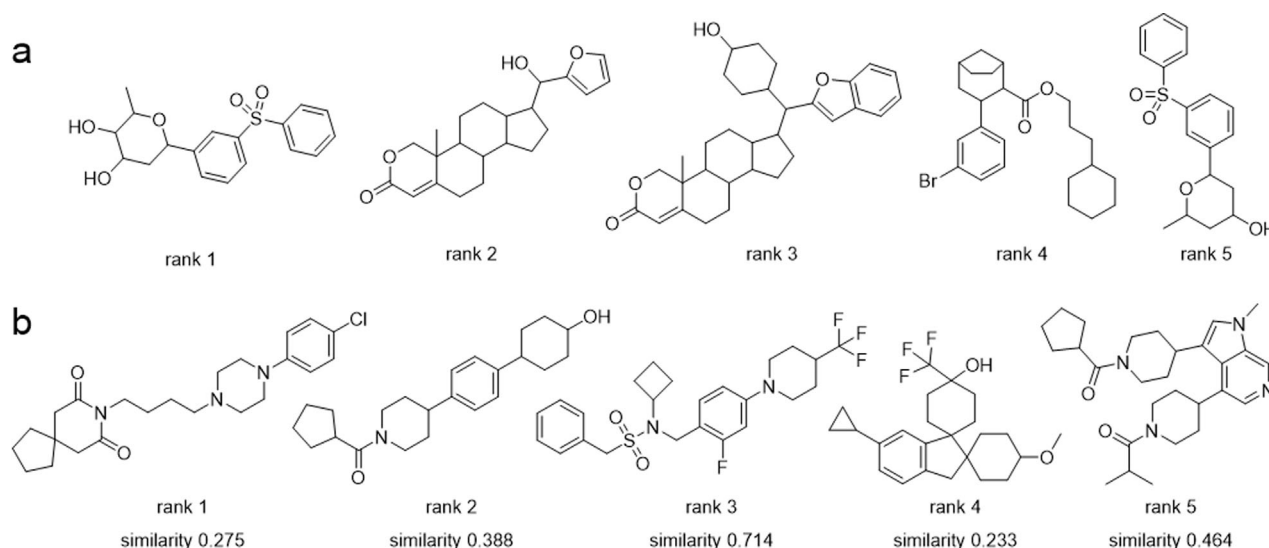


Figure 2. Top-ranked designs obtained by beam search sampling. a) Single fine-tuning, b) double fine-tuning. Ranks are based on the beam search score of the molecular designs. For the top-ranked molecules from the double fine-tuning experiment, the similarity values refer to the Tanimoto similarity computed on Morgan fingerprints (length = 1024, 2-bond radius) to the closest known active molecule annotated in ChEMBL with an IC_{50} value for ROR γ (structures are shown in Figure S2).

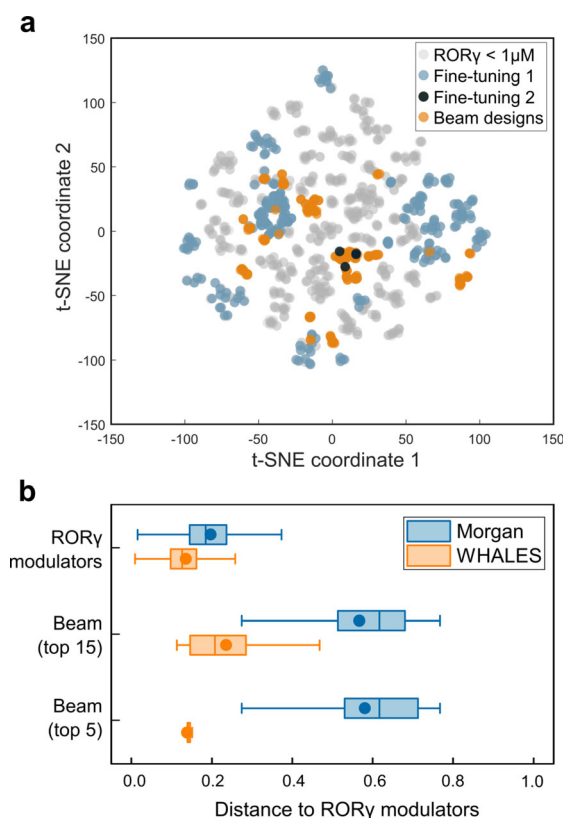


Figure 3. Characteristics of designs from the CLM with double fine-tuning. a) Stochastic neighbor embedding (t-SNE)^[45] projection of the compound sets based on Morgan fragment fingerprints (length = 1024, 2-bond radius, Tanimoto similarity). The location of the two-fine tuning sets, the RORγ modulators annotated in ChEMBL ($IC_{50} < 1 \mu\text{M}$, 1091 compounds), and the beam search designs are shown. b) Comparison of the sampled molecular designs with known RORγ modulators ($IC_{50} < 1 \mu\text{M}$) in terms of Morgan fragment fingerprints (“Morgan”) and three-dimensional shape and electrostatics descriptors (WHALES). The pairwise distance distribution among known RORγ modulators contained in ChEMBL is shown as a reference. For Morgan fingerprints, the Tanimoto distance is shown; for WHALES the range-scaled Euclidean distance is shown. “Beam (15)” and “Beam (5)” indicate the top 15 and top 5 designs, respectively. Boxplots indicate 25th, 50th, and 75th percentiles (lines), mean values (circle), and outlier boundaries (whiskers, $1.5 \times$ interquartile range).

learned certain “semantic” ligand features that are relevant for binding to macromolecules, such as their molecular shape and partial charge patterns.

Prospective Experimental Validation

Three beam search designs were synthesized and characterized *in vitro*. We selected them based on their beam search score. From the five most likely designs (Figure 2b), we selected molecules **1** and **2**, which were ranked first and third. Compound **2** showed the highest Tanimoto similarity (Morgan fingerprints) to a known RORγ modulator (Figure 2b). The scaffolds of both compounds were also prominent among the beam search designs not included in the top 5, suggesting a structural preference. The scaffold of **1** also appeared in the

design ranked 6th. Molecules ranked 10th and 13th resembled compound **2**. Hence, we additionally chose compound **3** of this abundant chemotype from rank 13 for prospective validation. Compounds **1–3** were synthesized according to Scheme 1.

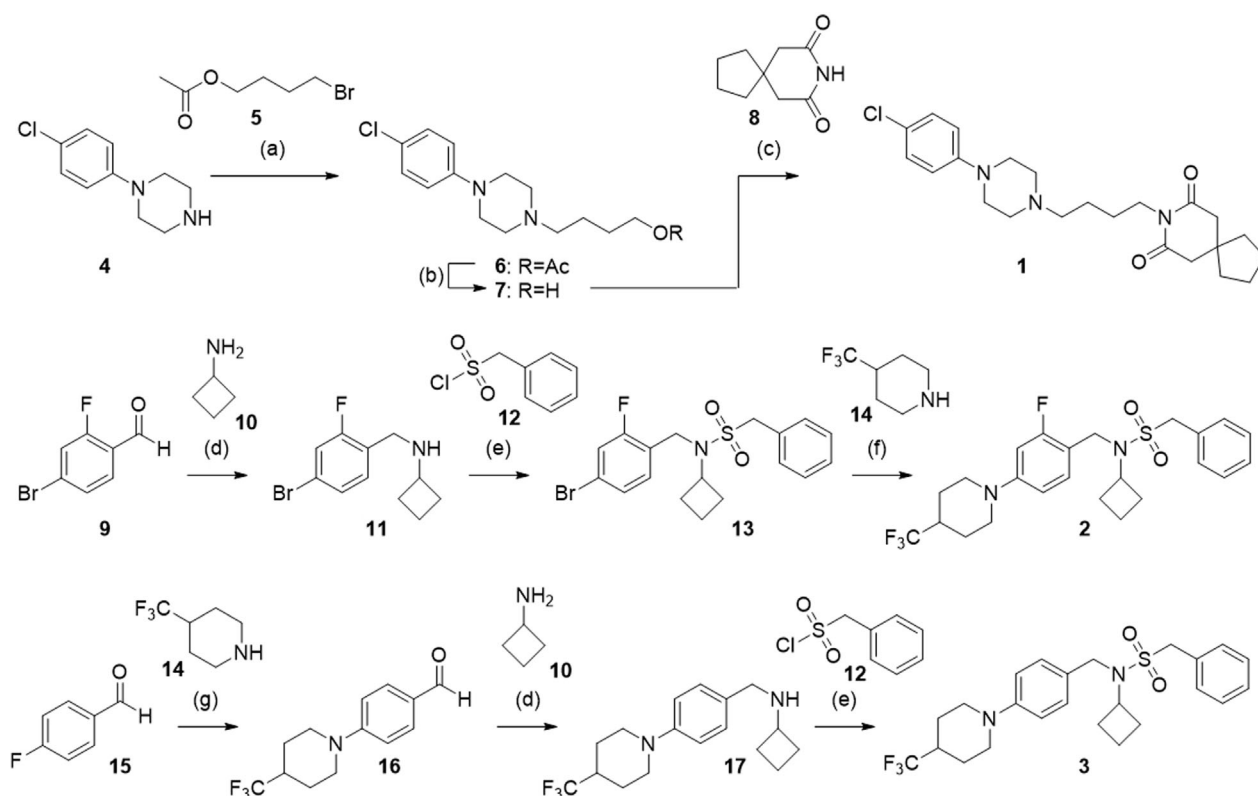
For preparation of **1**, (4-chlorophenyl)piperazine (**4**) was treated with 4-bromobutyl acetate (**5**) to obtain the ester-protected intermediate **6** which after alkaline ester hydrolysis to **7** was suitable for Mitsunobu reaction with 8-azaspiro-[4.5]decane-7,9-dione (**8**) to obtain the top-ranked computational design **1**. Preparation of **2** started from 4-bromo-2-fluorobenzaldehyde (**9**) which was reacted with amine **10** to obtain **11** by reductive amination followed by sulfonamide coupling with **12** to yield **13**. Eventually, the 4-trifluoromethylpiperidine substituent was introduced to **13** under Buchwald–Hartwig conditions with **14** yielding compound **2**. The structurally related design **3** was prepared via a different route starting from a nucleophilic aromatic substitution of 4-fluorobenzaldehyde (**15**) with 4-trifluoromethylpiperidine (**14**) to **16**. The nucleophilic aromatic substitution provided substantially higher yield (see Scheme 1) than the Buchwald–Hartwig reaction but could not be employed in the synthesis of **2** because of the potential formation of regioisomers. Reductive amination of **16** with cyclobutaneamine (**10**) to **17**, followed by sulfonamide formation with phenylmethanesulfonyl chloride (**12**), afforded the computationally designed compound **3**.

In vitro characterization of compounds **1**, **2**, and **3** in Gal4-ROR hybrid reporter gene assays confirmed inverse RORγ agonism with micromolar to sub-micromolar IC_{50} values (Table 1). The top-ranked compound **1** counteracted RORγ activity with an IC_{50} value of $4.6 \mu\text{M}$. It was additionally active on RORα and RORβ, but precise IC_{50} values could not be determined due to cytotoxicity. Compounds **2** and **3** blocked RORγ activity with IC_{50} values of $0.37 \mu\text{M}$ (**2**) and $0.68 \mu\text{M}$ (**3**), respectively. In addition to being inverse RORγ agonists, all three synthesized designs revealed pronounced preference for the RORγ subtype, with compounds **2** and **3** possessing more than tenfold higher potency on RORγ compared to the related RORα and RORβ isoforms. These results show that the CLM with beam search sampling conserved the bioactivity of the training molecules in the computational designs.

Conclusion

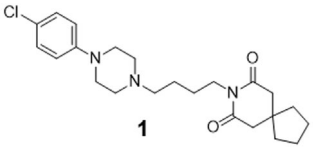
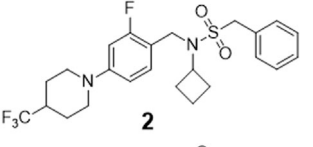
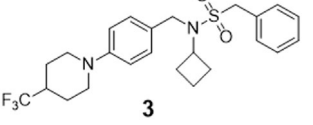
Herein, Beam search sampling from CLMs was applied to generating new molecules with desired bioactivity on the ligand-activated transcription factor RORγ. The algorithm automatically generated and scored the designs, without the need of additional prioritization rules. Prospective experimental validation yielded three novel, potent inverse agonists of the nuclear receptor with various degrees of similarity to known RORγ modulators (ranging from 0.28 to 0.71, as captured by Tanimoto similarity on Morgan fingerprints). Apparently, the beam search approach coupled with a CLM conserves structural features necessary for the desired bioactivity but still generates structurally diverse compounds in terms of fragments. This observation corroborates beam





Scheme 1. Synthesis of the CLM designs **1**, **2**, and **3**. Reagents and conditions: a) DMF, 4-DMAP, 60°C, 16 h, 48%; b) KOH, H₂O/THF/MeOH, microwave irradiation, 100°C, 30 min, 98%; c) DIAD, PPh₃, THF, 0°C→r.t., 16 h, 42%; d) NaB(OAc)₃H, HOAc, DCE, r.t., 50 h, 73%; e) 4-DMAP, pyridine, CH₂Cl₂, reflux, 16 h, 37%; f) Pd₂(dba)₃, xantphos, Cs₂CO₃, 1,4-dioxane, reflux, 16 h, 18%; g) K₂CO₃, DMSO, reflux, 48 h, 82%.

Table 1: Activity of de novo designs **1**, **2**, and **3** on RORs in Gal4 hybrid reporter gene assays. Data are reported as mean ± S.E.M., *n* ≥ 4.

Structure and ID	ROR α	IC ₅₀ [μM] ROR β	ROR γ
 1	> 10	> 10	4.6 ± 0.5
 2	23 ± 3	22 ± 1	0.37 ± 0.05
 3	10 ± 1	7.6 ± 0.5	0.68 ± 0.07

search sampling as a technique for the de novo design of bioactive molecules by a CLM. The computational and experimental results suggest two attractive properties of the beam search algorithm. Firstly, by searching for the most likely molecules a CLM can generate, the beam search algorithm probes the suitability of a CLM for the given task. Evaluation of the resulting designs allows to check the

compliance of the CLM designs with the design objectives and to assess the success of fine-tuning. This is in contrast to standard temperature sampling, which might lead chemists to consider designs that are not likely according to the model. Secondly, beam search sampling could potentially overcome the need for external compound prioritization. It should be noted, however, that the number of designs that can be sampled by beam search is limited compared to temperature sampling, which can virtually generate an infinite number of chemical structures. The two techniques complement each other, and both offer characteristic advantages. The desired application should guide the choice of either strategy. If corroborated in future prospective studies, beam search sampling may help to further the applicability of CLMs for de novo molecular design in medicinal chemistry.

Acknowledgements

This research was supported by the Swiss National Science Foundation (grant no. 205321_182176 to G.S.), the RE-THINK initiative at ETH Zurich, and the Novartis Forschungsstiftung (FreeNovation grant “AI in Drug Discovery” to G.S.). Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

G.S. declares a potential financial conflict of interest as a founder of inSili.com GmbH, Zurich, and in his role as consultant to the pharmaceutical industry.

Keywords: de novo design · deep learning · drug discovery · neural network · nuclear receptor

- [1] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85–117.
- [2] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.
- [3] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241–1250.
- [4] W. P. Walters, R. Barzilay, *Acc. Chem. Res.* **2021**, *54*, 263–270.
- [5] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.
- [6] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268–276.
- [7] N. de Cao, T. Kipf, *arXiv* **2018**, <https://arxiv.org/abs/1805.11973>.
- [8] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700111.
- [9] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.
- [10] J. Bradshaw, B. Paige, M. J. Kusner, M. H. S. Segler, J. M. Hernández-Lobato, *arXiv* **2020**, <https://arxiv.org/abs/2012.11522>.
- [11] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- [12] M. A. Skinnider, R. G. Stacey, D. S. Wishart, L. J. Foster, *ChemRxiv* **2021**, <https://doi.org/10.26434/CHEMRXIV.13638347.V1>.
- [13] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2018**, *4*, 120–131.
- [14] W. Yuan, D. Jiang, D. K. Nambiar, L. P. Liew, M. P. Hay, J. Bloomstein, P. Lu, B. Turner, Q. T. Le, R. Tibshirani, P. Khatri, M. G. Moloney, A. C. Koong, *J. Chem. Inf. Model.* **2017**, *57*, 875–882.
- [15] D. Merk, F. Grisoni, L. Friedrich, G. Schneider, *Commun. Chem.* **2018**, *1*, 68.
- [16] Y. Yang, R. Zhang, Z. Li, L. Mei, S. Wan, H. Ding, Z. Chen, J. Xing, H. Feng, J. Han, H. Jiang, M. Zheng, C. Luo, B. Zhou, *J. Med. Chem.* **2020**, *63*, 1337–1360.
- [17] M. Moret, L. Friedrich, F. Grisoni, D. Merk, G. Schneider, *Nat. Mach. Intell.* **2020**, *2*, 171–180.
- [18] M. Awale, F. Sirockin, N. Stiefl, J. L. Reymond, *J. Chem. Inf. Model.* **2019**, *59*, 1347–1356.
- [19] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3320–3328.
- [20] M. Peters, S. Ruder, N. A. Smith, *arXiv* **2019**, <https://arxiv.org/abs/1903.05987>.
- [21] H. Altae-Tran, B. Ramsundar, A. S. Pappu, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 283–293.
- [22] Y. Wang, Q. Yao, J. Kwok, L. M. Ni, *arXiv* **2019**, <https://arxiv.org/abs/1904.05046>.
- [23] X. Yang, J. Zhang, K. Yoshizoe, K. Terayama, K. Tsuda, *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.
- [24] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572–1583.
- [25] D. Grechishnikova, *Sci. Rep.* **2021**, *11*, 321.
- [26] G. Benoit, A. Cooney, V. Giguere, H. Ingraham, M. Lazar, G. Muscat, T. Perlmann, J. P. Renaud, J. Schwabe, F. Sladek, M. J. Tsai, V. Laudet, *Pharmacol. Rev.* **2006**, *58*, 798–836.
- [27] D. P. Marciano, M. R. Chang, C. A. Corzo, D. Goswami, V. Q. Lam, B. D. Pascal, P. R. Griffin, *Cell Metab.* **2014**, *19*, 193–208.
- [28] Y. Hoon Kim, M. A. Lazar, *Endocr. Rev.* **2020**, *41*, 707–732.
- [29] V. B. Pandya, S. Kumar, Sachchidanand, R. Sharma, R. C. Desai, *J. Med. Chem.* **2018**, *61*, 10976–10995.
- [30] L. A. Solt, T. P. Burris, *Trends Endocrinol. Metab.* **2012**, *23*, 619–627.
- [31] S. Asimus, R. Palmér, M. Albayaty, H. Forsman, C. Lundin, M. Olsson, R. Pehrson, J. Mo, M. Russell, S. Carlert, D. Close, D. Keeling, *Br. J. Clin. Pharmacol.* **2020**, *86*, 1398–1405.
- [32] D. J. Kojetin, T. P. Burris, *Nat. Rev. Drug Discovery* **2014**, *13*, 197–216.
- [33] L. T. Lowerre, PhD Thesis Carnegie Mellon Univ. Pittsburgh, **1976**.
- [34] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735–1780.
- [35] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, J. P. Overington, *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.
- [36] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2020**, *83*, 770–803.
- [37] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, *Nat. Chem.* **2016**, *8*, 531–541.
- [38] P. Ertl, A. Schuffenhauer, in *Prog. Drug Res.*, Birkhäuser, Basel, **2008**, pp. 217–235.
- [39] P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68–74.
- [40] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci.* **2020**, *11*, 3316–3325.
- [41] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, C. Zardecki, *Acta Crystallogr. Sect. D* **2002**, *58*, 899–907.
- [42] H. L. Morgan, *J. Chem. Doc.* **1965**, *5*, 107–113.
- [43] F. Grisoni, D. Merk, V. Consonni, J. A. Hiss, S. G. Tagliabue, R. Todeschini, G. Schneider, *Commun. Chem.* **2018**, *1*, 44.
- [44] F. Grisoni, G. Schneider, *Methods Mol. Biol.* **2021**, 2266, 11–35.
- [45] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Manuscript received: March 30, 2021

Revised manuscript received: June 2, 2021

Accepted manuscript online: June 24, 2021

Version of record online: ■ ■ ■ ■ ■ ■ ■ ■ ■ ■



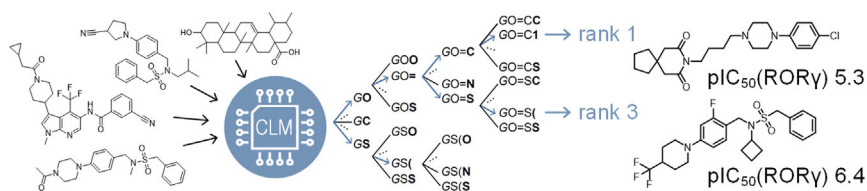
Research Articles



De novo Design

M. Moret, M. Helmstädter, F. Grisoni,
G. Schneider,* D. Merk* — ■■■■-■■■■

Beam Search for Automated Design and
Scoring of Novel ROR Ligands with
Machine Intelligence



The beam search algorithm was employed for automated molecule design and scoring from a chemical language model (CLM). Prospective application of this model-intrinsic technique with a CLM trained on inverse ROR γ agonists yielded

novel ligands of this nuclear receptor with the intended bioactivity. Beam search sampling overcomes the need for external scoring methods and extends the applicability of machine learning-driven molecular design.