Generating Bioactive Natural Product-inspired Molecules with Machine Intelligence

Petra Schneider, Karl-Heinz Altmann, and Gisbert Schneider*

Abstract: The computer-assisted design of new chemical entities has made a leap forward with the development of machine learning models for automated molecule generation. The overarching goal of this conceptual approach is to augment the creativity of medicinal chemists with a machine intelligence. In this Perspective we highlight prospective applications of “de novo” drug design and target prediction, aiming to generate natural product-inspired bioactive compounds from scratch. A virtual chemist transforms pharmacologically active natural products into new, easily synthesizable small molecules with desired properties and activity. Computational activity prediction and automated compound generation offer the possibility to systematically transfer the wealth of pharmacologically active natural products to synthetic small molecule drug discovery. We present selected prospective examples and dare a forecast into the future of natural product-inspired drug discovery.

Keywords: Artificial intelligence · Drug discovery · Machine learning · Natural product

1. Introduction

Revitalizing the concept of natural product-inspired drug discovery presents a unique opportunity for future healthcare. Evidently, natural products have traditionally inspired chemists and filled the industry pipelines.[1] Approximately half of the new drugs approved by the U.S. Food and Drug Administration (FDA) are of natural origin or natural product-derived synthetic compounds.[2] For example, marine natural products are of particular interest as anticancer and antiviral agents, with several approved drugs.[3] Similarly, certain edible algae have recently been identified as sources of potential anti-obesity agents.[4] However, despite their unrivalled appeal for pharmaceutical discovery as a source of inspiration, natural products as drugs have to some degree fallen into disgrace in the medicinal chemistry community because of supply problems (ecological sustainability), partially lengthy and costly total syntheses (ecological sustainability), and an often

*Correspondence: Prof. Dr. G. Schneider, E-mail: gisbert@ethz.ch, Dept. Chemistry and Applied Sciences, ETH Zurich, CH-8093 Zurich, Switzerland
unknown molecular mode of action (scientific sustainability).\textsuperscript{[5]} Advances in computer-assisted drug discovery, spurred in part by stellar method development in both machine learning (as a subdomain of ‘artificial intelligence’) and biotechnology,\textsuperscript{[6,7]} bear promise to overcome some of these limitations.\textsuperscript{[8,9]} In this Perspective we highlight the ability of ligand-based machine learning approaches to generate new chemical structures from scratch by molecular ‘de novo’ design, and explain how these tools can be used to obtain natural product-inspired drug-like chemical entities (Figs. 1). Selected retrospective and prospective applications illustrate the feasibility of this innovative drug design concept as an amalgamation of traditional and modern medicinal chemistry.

2. Molecular de novo Design

Molecular ‘scaffold hopping’ aims to identify pairs of molecules that have markedly different chemical structures but share a certain function of interest, e.g. binding to the same macromolecular target.\textsuperscript{[10]} De novo structure generation provides such sets of structurally diverse molecules that have certain features in common.\textsuperscript{[11]} In fact, automated de novo molecule construction enables medicinal chemists to computationally access a virtually infinite chemical space. Scaffold hopping by de novo design seems particularly suited for finding structurally novel hit and lead compounds.\textsuperscript{[12]} The approach ideally complements virtual chemical library screening\textsuperscript{[13]} but requires synthesizing the computer-generated molecules.

Molecular design with machine intelligence includes both rule-based and rule-free approaches.\textsuperscript{[14]} Rule-based methods use sets of molecular building blocks and chemical transformations (e.g. virtual reaction schemes) for molecule construction. The DOGS (Design of Genuine Structures) algorithm belongs to this class of methods.\textsuperscript{[15]} In contrast, rule-free ‘generative’ methods sample new molecules from a learned statistical distribution of the training data, which usually requires a large set of known molecules.\textsuperscript{[16]} Most of the contemporary generative approaches build on deep neural networks,\textsuperscript{[17]} Recurrent neural networks with long short-term memory as ‘chemical language models’,\textsuperscript{[18]} variational autoencoders,\textsuperscript{[19]} generative adversarial networks,\textsuperscript{[20]} graph neural networks,\textsuperscript{[21]} and various other deep learning architectures,\textsuperscript{[22]} have been proposed for this purpose. Rule-free and rule-based methods alike have been successfully employed for prospective small molecule drug design, resulting in new bioactive compounds.\textsuperscript{[23]} Importantly, both contemporary de novo methods deliver synthetically feasible molecular designs, thereby overcoming a stigma of earlier approaches.\textsuperscript{[24]} In contrast to generative models, rule-based methods do not require large sets of training data. These methods are applicable when only a single active molecule (the ‘template’ molecule for ligand-based de novo design) or a model of the binding pocket (for structure-based de novo design) is known.

2.1 Generating Natural Product-inspired Molecules with Machine Intelligence

Computational de novo design is increasingly employed for automated small molecule construction, starting from mostly synthetic small molecules as design templates or as training data for the underlying machine learning models. The first applications of generative deep learning (chemical language model) to natural products were published in 2018.\textsuperscript{[125]} Several novel nuclear hormone receptor modulators with micromolar to nanomolar activity were obtained. Selected recent examples of prospective studies aiming to ‘scaffold hop’ from natural products to isofunctional synthetic molecules are compiled in Fig. 2. Of note, in each case, only a very small number of known bioactive natural products (one to six molecules) served as design template(s) for rule-based compound construction (Fig. 2a–c), or deep generative models (Fig. 2d), respectively.

Two independent design runs with DOGS software, taking either valerenic acid (1) or dehydroabiatic acid (2) as design template, resulted in the identical computer-generated molecule 3.\textsuperscript{[36]} This observation highlights the scaffold hopping capability of rule-based molecular design. Target prediction with SPIDER software\textsuperscript{[22]} predicted retinoid X receptor (RXR) agonism for the two natural products and the de novo molecule. Tetrahydroindole derivative 3 was obtained in a microwave-assisted two-step synthesis. Full dose-response analysis revealed low micromolar potency of the natural product templates and de novo generated compound 3 on all three RXR subtypes.

In an attempt to morph structurally more intricate natural products to small synthetic mimetics DOGS software was applied to the marine natural product marinopyrrole A (4), which led to computer generated molecule 5.\textsuperscript{[28]} Compound 5 was obtained in a two-step synthetic route, as suggested by the design software. Target prediction suggested cyclooxygenase-1 (COX-1) inhibition as a hitherto unknown activity of the anticancer agent marinopyrrole A and the de novo molecule. In vitro testing confirmed compounds 4 (IC\textsubscript{50} = 7.7±3.9 μM) and 5 (IC\textsubscript{50} = 0.10±0.05 μM) as direct COX-1 inhibitors. Of note, molecule 5 inhibited COX-1 with low-nanomolar potency in platelets (IC\textsubscript{50} = 0.009±0.001 μM). This de novo designed natural product mimetic preferentially inhibits the biosynthesis of COX-1–derived products in human platelets and monocytes. Compound 5 behaved similar to indomethacin with regard to COX-1 inhibition in platelets and showed greater COX-1 selectivity. The unique binding mode of this de novo generated compound was confirmed by X-ray structure determination of the ligand-enzyme complex. This example demonstrates the applicability of ligand-based molecular design to practical natural product-inspired medicinal chemistry.

In another study, the structurally complex natural product (–)-englerin A (6), a known inhibitor of transient receptor potential (TRP) cation channels, served as design template for DOGS.\textsuperscript{[29]} Using two different computational scoring methods (pharmacophore-based and shape-based), two different de novo generated molecules were prioritized for synthesis. Compounds 7 and 8 were afforded in 3-step and 2-step synthetic protocols, respectively, as suggested by the software. Activity determination confirmed the natural product and the computer-generated molecules as potent inhibitors of TRPML8 (K\textsubscript{i} = 0.2–0.3 μM). Importantly, the chemical constitution of the natural product templates used for rule-based de novo design served as the only reference information for automated ligand-based molecule construction. This computational approach might, therefore, prove particularly useful in “low-data”
situations that are restrictive for de novo drug design with data-hungry generative deep learning.

Using a generative approach (chemical language model) for natural-product inspired molecular design, computer-generated compound 9 was obtained by 2-step synthesis, and experimentally confirmed as a weak pan-RXR partial agonist (EC\(_{50}\) = 20–30 μM, 6–10 fold receptor activation).\(^\text{[31]}\) The importance of this result lies in the fact that only six natural products (known RXR agonists) were used to bias the generative chemical language model, and no explicit target prediction was performed. The machine intelligence implicitly captured the structural requirements of RXR agonists and ranked the de novo generated molecules.

### 2.2 Case Study: Designing Nucleoside Analogs with a Chemical Language Model

In light of the urgent need for novel antiviral therapeutics, de novo design may play a decisive role for future rapid hit and lead identification. To provide another example of generative de novo design, we targeted RNA-dependent RNA polymerase (RdRp) of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).\(^\text{[32]}\) aiming to obtain ideas for new SARS-CoV-2 RdRp inhibitors. For this purpose, we trained a generative chemical language model.\(^\text{[33]}\) In the first step, a generic model was obtained by learning the chemical syntax of a large set of known, bioactive druglike compounds. In the second step, this model was biased toward nucleoside analogs acting as SARS-CoV-2 RdRp inhibitors.\(^\text{[34]}\) For the medium-sized anticancer depsipeptide doliculide\(^\text{[35]}\) the software revealed prostanoid E receptor EP3 antagonism (IC\(_{50}\) = 16 ± 7 nM, \(K_i\) = 6 nM).\(^\text{[36]}\) For the polyketide archaeozolide A, a known inhibitor of V-ATPase, farnesoid X receptor (FXR) agonist, EC\(_{50}\) = 0.2 μM) and other hitherto unknown targets could be identified.\(^\text{[37]}\)

Aside from providing a straightforward access to target and activity prediction for large natural product structures, e.g. macrocycles or peptides. Aiming at partially alleviating this issue, one can virtually dissect the large natural product into smaller portions and perform target predictions for the resulting “drug-sized” fragments.\(^\text{[38]}\) Fig. 4 shows three such examples of new target identification with TIGER. Resveratrol\(^\text{[36]}\) (9) is a small natural product, for which estrogen receptor beta antagonism was predicted and experimentally confirmed (\(K_i\) = 0.4 μM).\(^\text{[39]}\) For the medium-sized anticaner depsipeptide doliculide\(^\text{[40]}\) the software revealed prostanoid E receptor EP3 antagonism (IC\(_{50}\) = 16 ± 7 nM, \(K_i\) = 6 nM).\(^\text{[36]}\) For the polyketide archaeozolide A, a known inhibitor of V-ATPase, farnesoid X receptor (FXR) agonist, EC\(_{50}\) = 0.2 μM) and other hitherto unknown targets could be identified.\(^\text{[37]}\)

### 3. De-orphaning the Targets of Bioactive Natural Products with Machine Intelligence

Targeted natural product-inspired drug discovery requires the knowledge of the pharmacologically relevant macromolecular binding partners of the natural products. Numerous software tools are available for target and activity prediction, ranging from structure-based (e.g. docking) to ligand-based (e.g. substructure-, pharmacophore-, shape-based) methods.\(^\text{[41]}\) No method is perfect but all have their individual sweet spot. One of the most successful and widely applied tools is TIGER (Target Inference GENeratorR), which has proven applicable to natural products.\(^\text{[33]}\) The TIGER algorithm works on the two-dimensional chemical structure (chemical constitution) of the ligand and does not take the target structure into account. It is thus applicable to a wide range of targets and ligands. Most target prediction tools, including TIGER, were developed using small molecule reference data. Their prediction accuracy typically suffers when applied to larger natural product structures, e.g. macrocycles or peptides. Aiming at partially alleviating this issue, one can virtually dissect the large natural product into smaller portions and perform target predictions for the resulting “drug-sized” fragments.\(^\text{[38]}\)

### 3.1 Revealing Targets of Complex Natural Products: A Prospective Study

The marine natural product (−)-zampanolide\(^\text{[20]}\) 20 is a microtube-stabilizing antiproliferative macrolide from the Togan sponge Cacospongia mycofijensis (Fig. 5).\(^\text{[39]}\) Its total synthesis was achieved in 2012,\(^\text{[40]}\) and several structural analogues have been synthetically obtained ever since.\(^\text{[41]}\) The unmodified natural product (12) potently inhibits the growth of different human cancer cells in vitro with nanomolar IC\(_{50}\) values, whereas its synthetic de-
The derivative desTHP-(–)-zampanolide (13) is >100-fold less potent.\cite{40} In contrast to 12, 13 shows no measurable binding to stabilized microtubules up to concentrations of 25 µM (unpublished data). We took this information as a starting point to identify potential human drug targets of compound 13. Based solely on the two-dimensional structure of compound 13, TIGER software (version 16.10) suggested cholecystokinin receptor B (TIGER score = 90), N-formyl peptide receptor 2 FPR2 (TIGER score = 24.5), and prostanoid receptor EP3 (TIGER score = 22.3) as the most confidently predicted targets. The suggestion of FPR2 seemed particularly interesting because of its vital role in cell differentiation and chemotaxis.\cite{42} Experimental validation corroborated compound 13 as a partial agonist of FPR2 \textit{in vitro} (75% max. receptor activation, EC$_{50}$ = 5±1 µM, Fig. 6).\cite{43} This prospective example positively advocates the use of machine learning models for ligand-based target prediction and ligand ‘de-orphaning’.

4. Conclusions and Outlook

Computational \textit{de novo} molecular design has proven value for generating reasonable chemical structures. These computer-generated molecules provide working hypotheses for chemical synthesis and experimental testing. Concluding from several successful prospective applications to natural products, this approach seems suited for finding synthetic analogues and mimetics of bioactive natural products. Both rule-based and rule-free machine learning methods for \textit{de novo} structure generation can be used for this purpose. Remaining challenges are the structural complexity of certain natural products and the functional relevance of certain substructural elements, \textit{e.g.} carbohydrate moieties, among others.
Also, important activity-determining structural features, e.g., stereocenters, are insufficiently accounted for by the currently available ligand-based de novo design methods.

Recently, there have been attempts to integrate computational molecule construction and activity prediction into a single generative model. These computational models, if proven successful in diverse prospective settings, will be able to automatically generate new chemical entities with desired properties and biological activity, without the need for explicit activity prediction. Certain deep learning approaches have already been shown applicable in this regard. Geometric deep learning techniques allow to explicitly consider three-dimensional chemical features for model building. However, this approach has not been applied to natural products yet. A further area of development is the combination of deep learning with a rule-based molecule construction process, thereby allowing chemists to explore the virtual chemical space that is accessible with the molecular building blocks that are readily available in the laboratory. In the near future we expect to see fully automated laboratories implementing design-make-test-optimize cycles, thereby enabling the full exploration of the structural and functional diversity of natural products for de novo design. According to the World Health Organization, infectious diseases represent a leading global public health threat in the 21st century. At the same time, populations are growing and ageing owing to successes against infections. This situation paradoxically raises the risk of developing chronic diseases. The proposed drug design approach could enable a renaissance of natural product-inspired pharmaceutical research by amalgamating modern medicinal chemistry with generative artificial intelligence. This matchless setting will alleviate some of the limitations of traditional natural product-based drug discovery, and harness the healing power of naturally evolved solutions for future synthetic medicines.

Acknowledgements

This research was supported by the Swiss National Science Foundation (grant no. 205321_182176 to G.S.).

Conflict of Interest Statement

P.S. and G.S. declare a potential financial conflict of interest as in their role as founders of inSili.com LLC and consultants to the pharmaceutical industry.

Received: April 13, 2022

Fig. 5. Structures of the antiproliferative marine macrolide (−)-zampanolide and its synthetic derivative desTHP-(−)-zampanolide.

Fig. 6. Activation of N-formyl peptide receptor 2 (FP2) by desTHP-(−)-zampanolide (13). Data shown from n = 2 independent experiments.

[1] A. L. Harvey, Drug Discov. Today 2008, 13, 894, https://doi.org/10.1016/j.drudis.2008.07.004
[2] E. Patridge, P. Gareiss, M. S. Kinch, D. Hoyer, Drug Discov. Today 2015, 21, 204, https://doi.org/10.1016/j.drudis.2015.01.009.
[3] a) U. Lindequist, Biomol. Ther. 2016, 24, 561, https://doi.org/10.4062/biomolther.2016.181; b) P. Shinde, P. Banerjee, A. Mandhare, Expert Opin. Ther. Pat. 2019, 29, 283, https://doi.org/10.1080/13543776.2019.1598972.
[4] C. Wan-Loy, P. Siew-Moi, Nat. Drugs 2016, 14, pii: E222, https://doi.org/10.3390/md141120222.
[5] a) G. Molinari, Adv. Exp. Med. Biol. 2009, 655, 13; b) A. G. Atanasov, S. B. Zotchek, V. M. Dirsch, International Natural Product Sciences Taskforce, C. T. Supuran, Nat. Rev. Drug Discov. 2021, 20, 200, https://doi.org/10.1038/s41573-020-00114-z.
[6] a) E. N. Muratov, R. Amaro, C. H. Andrade, N. Brown, S. Ekins, D. Fourches, O. Isayev, D. Kozakov, J. L. Medina-Franco, K. M. Merz, T. I. Opprea, V. Porokov, G. Schneider, M. H. Todd, A. Varnek, D. A. Winkler, A. V. Zakharov, A. Cherkasov, A. Troshpa, Chem. Soc. Rev. 2021, 50, 9121, https://doi.org/10.1039/D0CS01065K; b) X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, Chem. Soc. Rev. 2019, 119, 10520, https://doi.org/10.1021/acs.chemrev.8b00728.
[7] A. L. Harvey, R. Edrada-Ebel, R. J. Quinn, Nat. Rev. Drug Discov. 2015, 14, 111, https://doi.org/10.1038/nrd4510.
[8] N. E. Thomford, D. A. Senthebane, A. Rowe, D. Munro, P. Seele, A. Maroyi, K. Dzobo, Int. J. Mol. Sci. 2018, 19, 1578, https://doi.org/10.3390/ijms19061578.
[9] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, Nat. Chem. 2016, 8, 531, https://doi.org/10.1038/nchem.2479.
[10] a) G. Schneider, W. G. Neidhart, T. Giller, G. Schmid, Angew. Chem. Int. Ed. 1999, 38, 2894, https://doi.org/10.1002/(SICI)1521-3773(199904)38:19<2894::AID-ANIE2894>3.0.CO;2-F; b) Y. Hu, D. Stumpe, J. Bajorath, J. Med. Chem. 2016, 59, 4062, https://doi.org/10.1021/acs.jmedchem.5b01746.
[11] G. Schneider, U. Fechner, Nat. Rev. Drug Discov. 2005, 4, 649, https://doi.org/10.1038/nrd1799.
[12] Book: G. Schneider (ed.), De Novo Molecular Design, Wiley-VCH, Weinheim, New York, 2013, https://doi.org/10.1002/9783527677016.
[13] a) J. Lyu, S. Wang, T. E. Balus, I. Singh, A. Levit, Y. S. Moroz, M. J. O’Meara, T. Che, E. Alagaa, K. Tolmacheva, A. A. Tolmachev, B. K. Shoichet, B. L. Roth, J. J. Irwin, Nature 2019, 566, 224, https://doi.org/10.1038/s41586-019-0197-9; b) F. Gentile, J. C. Yaacoub, J. Gleave, M. Fernandez, A. T. Ton, F. Ban, A. Stern, A. Cherkasov, Nat. Protoc. 2022, 17, 672, https://doi.org/10.1038/s41596-021-00659-2.
[14] a) P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroaka, J. Listgarten, R. A. Goodnow Jr, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, G. Schneider, Nat. Rev. Drug Discov. 2020, 19, 353, https://doi.org/10.1038/s41573-019-0050-3; b) G. Schneider, Nat. Mach. Intell. 2019, 1, 128, https://doi.org/10.1038/s42256-019-0030-7.
[15] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, G. Schneider, PLoS Comput. Biol. 2012, 8, e1002380, https://doi.org/10.1371/journal.pcbi.1002380.
[16] a) Q. Vanhaelen, Y. C. Lin, A. Zhavoronkov, ACS Med. Chem. Lett. 2020, 11, 1496, https://doi.org/10.1021/acsmedchemlett.0c00088; b) Book: T. Jebara, Machine Learning: Discriminative and Generative, Kluwer Academic,
