doi:10.2533/chimia.2019.1006 Chimia 73 (2019) 1006–1011 © Swiss Chemical Society

De novo Molecular Design with Generative Long Short-term Memory

Francesca Grisoni and Gisbert Schneider*

Abstract: Drug discovery benefits from computational models aiding the identification of new chemical matter with bespoke properties. The field of *de novo* drug design has been particularly revitalized by adaptation of generative machine learning models from the field of natural language processing. These deep neural network models are trained on recognizing molecular structures and generate new molecular entities without relying on pre-determined sets of molecular building blocks and chemical transformations for virtual molecule construction. Implicit representation of chemical knowledge provides an alternative to formulating the molecular design task in terms of the established, explicit chemical vocabulary. Here, we review *de novo* molecular design approaches from the field of 'artificial intelligence', focusing on instances of deep generative models, and highlight the prospective application of long short-term memory models to hit and lead finding in medicinal chemistry.

Keywords: Chemoinformatics · Deep learning · Drug discovery · LSTM · Neural network

Francesca Grisoni received her PhD in 2016 at the University of Milano-Bicocca in the research group of Prof. Roberto Todeschini, with a dissertation on Quantitative Structure–Activity Relationship (QSAR) and chemometric methods for bioaccumulation prediction. Since 2013, she has been working on QSAR development, *in silico* prediction of bioactivity and toxicity, machine learning, chemometrics and molecular descriptors. In 2019, she joined the Computer-Assisted Drug Design group of Prof. Gisbert Schneider at ETH Zurich as a Postdoctoral Researcher. Her current research focuses on generative artificial intelligence, deep learning and computer-assisted *de novo* design.

Gisbert Schneider is a full professor at ETH Zurich, holding the Chair for Computer-Assisted Drug Design. His research focuses on the integration of artificial intelligence into practical medicinal chemistry. His career has led him from the Pharmaceuticals Division at Roche, Basel, to academia, initially to the Goethe-University in Frankfurt where he held the Beilstein Endowed Chair for Chem- and Bioinformatics, and then to his current position at ETH Zurich. Schneider coined the terms 'scaffold hopping' and 'frequent hitter'. He is an elected Fellow of the University of Tokyo, and the recipient of the 2018 Herman Skolnik Award for his seminal contributions to *de novo* design of bioactive compounds.

1. Machine Intelligence in *de novo* Design

Computer-assisted molecular design has long been considered an opportunity for drug discovery. With the renewed relevance of 'artificial intelligence' (AI) research, the field is now in the midst of a surge of interest, catalyzed by advances in data processing power, the availability of software solutions for machine learning, and the development of innovative AI tools.[1–3] Part of the appeal of applying AI in drug design lies in the potential to develop data-driven model building processes to navigate datasets arising from experimental compound screening, generate new molecular structures, and prioritize the alternatives.[4] Given the complexity of multi-dimensional decision making in drug discovery, the key question is whether AI can help us identify better drug candidates faster. AI will probably play a role in answering this question.[5] These algorithms enable a computer system to interact with an environment and achieve goals in a wide range of settings,[6] from robotics to forecast.[7–10] There is a rich history of machine learning in drug discovery and design,[11] *e.g.* by guiding experimental testing, compound synthesis and library generation.[12–16] *De novo* design, *i.e.* the generation of a molecule with desired properties from scratch,[12] is one of the most challenging tasks for materials science and drug discovery.

One application of machine learning is quantitative structure–activity relationship (QSAR) modeling, which is based on the premise that the molecular structure is responsible for the molecule bioactivity and links molecular descriptors to experimentally-determined molecular properties with machine learning algorithms.[17–21] Compared to the 'direct' QSAR approach (*i.e.* inferring properties from molecular structures), 'inverse' QSAR[22,23] modeling (*i.e.* designing structures with desired properties) has an intrinsically higher complexity, *e.g.* due to the inversion of model equations, the presence of multiple solutions and difficulties in the reverse-decoding of molecular descriptors. The estimated cardinality of chemical space (10^{60} molecules)[24] renders combinatorial optimization and structure enumeration computationally demanding and limited with regard to the size of the virtual compound libraries. Consequently, the scope of *de novo* design based on virtual molecule enumeration is limited. Rule-based approaches for fragment assembly (*e.g.* evolutionary algorithms,[25–27] structure-based linking and growing,[28,29] reaction-driven design[30–32]) offer practical solutions when libraries of building blocks, molecule construction rules, and suitable scoring functions are available. Generative deep learning has emerged as a complementary approach to rule-based *de novo* design. These

*Correspondence: Prof. G. Schneider, E-mail: gisbert.schneider@pharma.ethz.ch
ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, CH-8093 Zurich, Switzerland

machine learning models represent and construct chemical entities implicitly, *i.e.* without the need of hard-coded design rules, pre-defined building blocks and scoring functions.

2. Generative Deep Learning

Fueled by the increase in computing performance, statistical advances and data availability,[33–35] *de novo* molecular design with generative approaches employs deep learning methods, in particular artificial neural network models (Fig. 1).[36–39] 'Deep' neural networks contain more than one layer of non-linear signal processing (Fig. 1).[40] Compared to 'shallow' machine learning architectures (*e.g.* neural networks with a single hidden layer, support vector machines, random forests), certain deep network models are able to grasp more complex data structures ('features of features') and need smaller amounts of labeled data for model development.[2,41,42]

Generative deep learning aims to model the underlying distribution of a given set of samples and, by sampling from the modelled distribution, generate new data points without the need of hard-coded design rules (Fig. 2).[49] Methods such as long short-term memory (LSTM) networks,[50] and variational autoencoders[45] have been proposed for this purpose (Fig. 1b).[37,51–56] These generative tools learn from known molecules which are represented in a suitable way (*e.g.* molecular graphs, simplified molecular input line entry systems (SMILES) string notations,[57] amino acid sequences) to generate representations of novel molecules – in the so-called 'end-to-end' (*e.g.* SMILES-to-SMILES) fashion. Generative deep learning methods can form internal representations of molecules, without the need of human-engineered rules for numerical encoding or sample generation (*e.g.* molecular descriptors, molecule assembly rules) (Fig. 2b).

2.1 *Long Short-term Memory Networks*

Of the various generative deep learning approaches employed for *de novo* design, LSTM networks (LSTMs)[50] have been studied in considerable detail.[56,58–60] LSTMs are recurrent neural networks borrowed from the field of natural language processing (Fig. 1b).[61] They are trained on sequential data, *e.g.* sequences of words or characters ('tokens', where a token is one discrete element of the sequence), and learn to predict one token at a time, based on the preceding portions of the sequences and a probability estimation. Once trained, the model can be used to generate novel sequences, by using a 'start' token as input and generating one token at a time, until the 'end' token is produced. LSTMs possess a memory unit (or memory cell, c_t, Fig. 3), which encodes information on inputs that have been observed before; additionally, memory gates control the information flow from past events to future predictions.[62] The input gate controls the extent to which the new input influences the cell state; the forget gate controls what to keep from the previous cell state; and the output gate controls the extent to which the updated cell state is used to compute the new hidden state value (s_t, Fig. 3). In this way, important sequence features can be carried along over long time spans, thereby capturing long-distance dependencies in the input sequences.[63]

LSTMs for *de novo* molecular design have been mostly used to generate SMILES strings[64–66] and amino acid sequences.[67,68] Fewer studies have focused on other string-based molecular representations (*e.g.* InChI,[51] DeepSMILES[66,69]). The theoretical concept of applying LSTMs to molecule design was proposed by Bjerrum and Threlfall,[60] who obtained 'chemically plausible' molecules according to retrosynthetic analysis using SMILES strings. Nagarajan *et al.*[68] trained an LSTM model to generate antimicrobial peptides (AMPs) using the amino acid sequence. Later, Segler *et al.*[56] pre-trained an LSTM model on 1.4 million

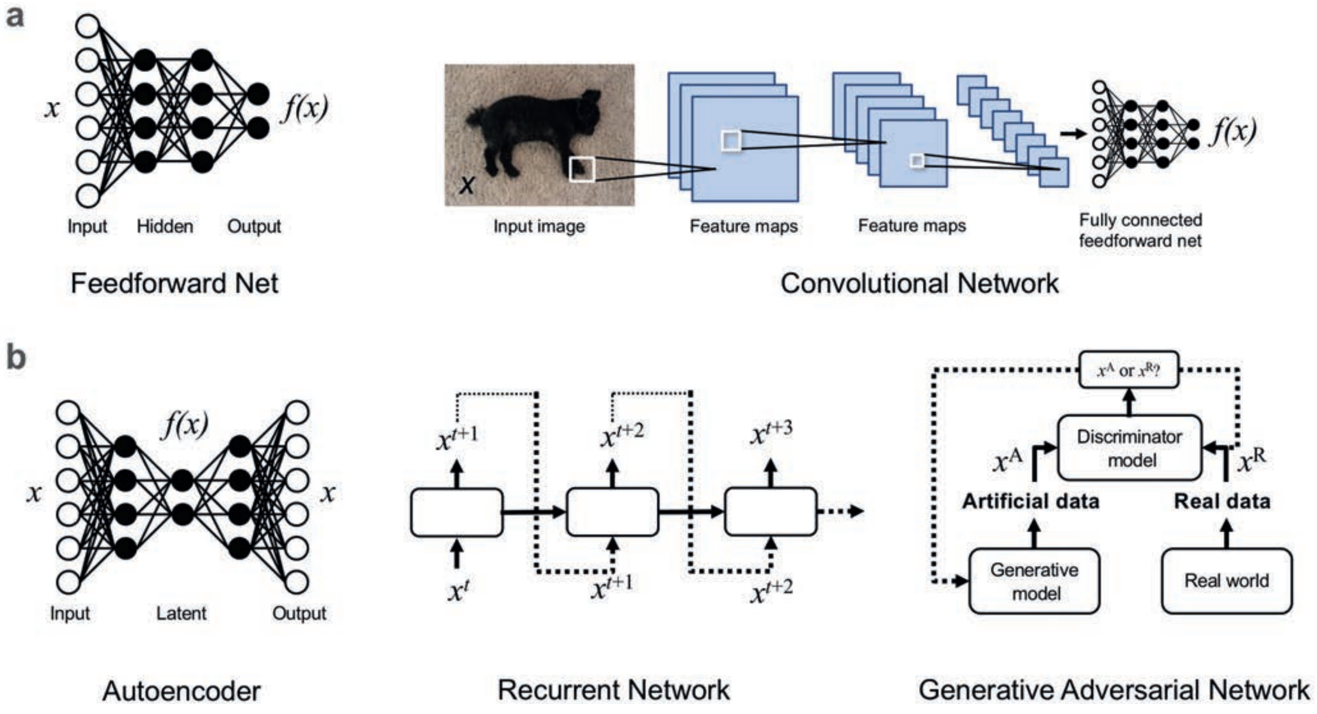

Fig. 1. Schematics of popular deep learning methods for hypothesis generation for virtual compound screening and *de novo* drug design. (a) Predictive approaches produce a qualitative or quantitative output as a function of the input *x*. A *feedforward net* (left) is a universal function approximator, where each layer of signal processing units (filled circles, 'neurons') captures increasingly complex features of *x* for computing *f(x)*. *Convolutional neural networks*[43,44] (right) are based on the successive application of filters to the pixels of the input image, and on convolution operations to capture the input-output relationships. (b) Generative models can be used to sample novel instances of *x* (*e.g.* molecules). *Variational Autoencoders*[45–47] (left) learn continuous data representations (encodings, *f(x)*), which are used to sample novel data points (*x*). *Recurrent networks* (middle) contain a recurrent layer (or cell) that is able to handle sequential data, by producing characters ('tokens') at time step *t* (x^t), based on the previous tokens of a sequence ($\{x^1,\ldots x^{t-1}\}$). *Generative adversarial networks*[48] (GANs; right) are composite deep models, where one network generates candidate molecules ('generator') and the other evaluates them ('discriminator'). Competition between the two networks leads to an improvement of both the generator and the discriminator model.

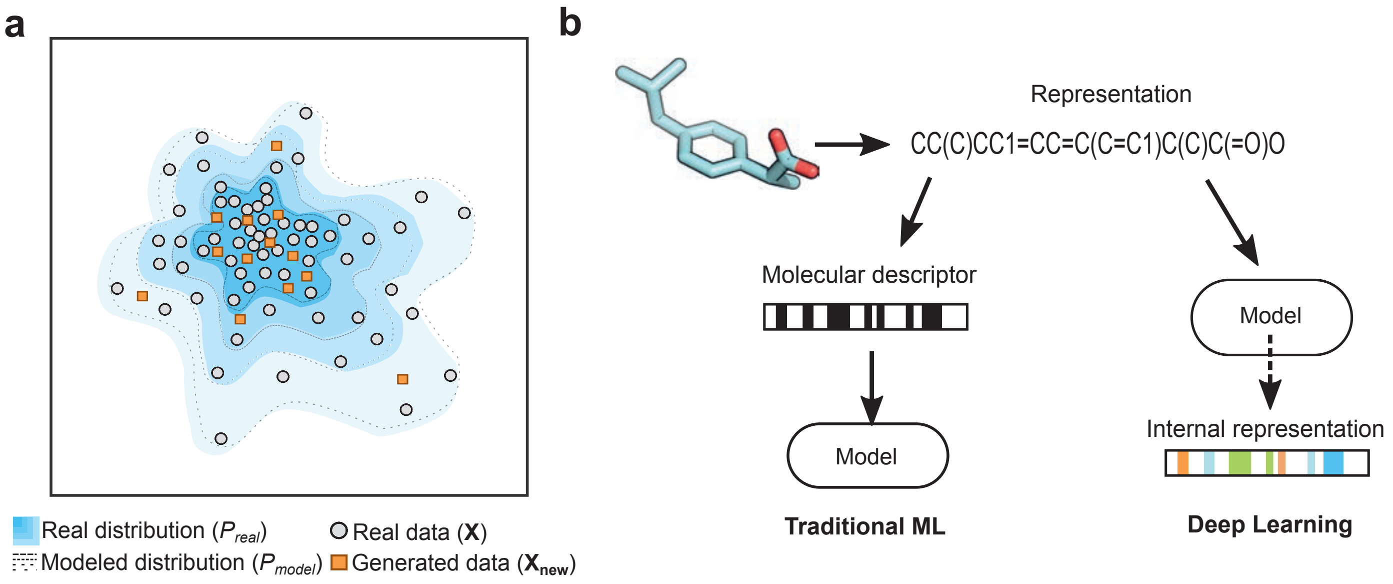
a
b
Representation
CC(C)CC1=CC=C(C=C1)C(C)C(=O)O
Molecular descriptor
Model
Traditional ML
Model
Internal representation
Deep Learning
Real distribution (P_{real})
Real data (**X**)
Modeled distribution (P_{model})
Generated data ($\mathbf{X}_{new}$)

Fig. 2. Deep learning for *de novo* molecular design. (a) Schematic representation of a generative model. Starting from real data (**X**) drawn from an (unknown) distribution (P_{data}), generative models form a probabilistic model of **X**. The modeled distribution (P_{model}) can be than used to generate new data instances ($\mathbf{X}_{new}$) that appear to be drawn from P_{data}, (b) Descriptor-based *vs* deep machine learning (ML). While descriptor-based ML relies on molecular representations that are chosen prior to model training (*e.g.* binary fingerprints, physicochemical properties), deep ML learns from more basic molecular representations (*e.g.* SMILES string, molecular graph).

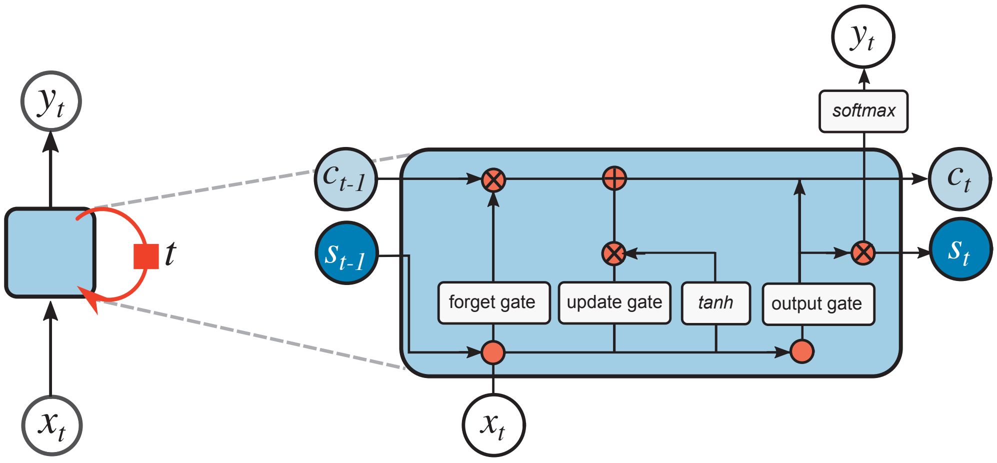

Fig. 3. Simplified representation of an LSTM with one neuron (left), which is a special type of recurrent neural network. LSTMs model a dynamic system, in which the network state at any *t*-th time point depends both on the current observation (x_t) and on the previous state (at *t*-1), and is used to predict the output (y_t). At each *t*-th time step, the network uses the input x_t to generate a predicted character, y_t, starting from the input x_t; this procedure is repeated for each *t*-th time step. In the generative setting, the LSTM model is trained to predict the next character, so that $y_t=x_{t+1}$. The zoom-in (right) shows an LSTM neuron, where c_t represents the memory cell which encodes information on the input that has been observed up to the *t*-th time step, and s_t represents the network state ('+' and '*x*' symboles indicate the sum and Hadamard product, respectively).

molecules, and then 'fine-tuned' it by using sets of molecules with desirable biological properties (transfer learning). Transfer learning – the task of transferring the knowledge of a previously trained model to a related, more specific, task on which less training samples are available[65] – proved useful to bias the model towards focused regions of chemical space, after learning the SMILES grammar in the pre-training step.[56] Since these first theoretical studies, LSTMs have been increasingly used for *de novo* design (*e.g.* refs [38,70–74]). In what follows, we focus on selected prospective applications of LSTMs in medicinal chemistry from our laboratory.

3. Prospective Application of Generative LSTM Models

Only few studies have focused on the prospective experimental testing of generative models for molecular design.[37,52,68,75,76] Recently, we have applied an LSTM model to fragment-based molecule design,[64] and to generating new bioactive nuclear receptor modulators from scratch.[75] This LSTM model was pretrained on approximately 500,000 bioactive molecules from ChEMBL22[77] (K_D, K_i, IC/EC_{50}<1 µM) and then fine-tuned on 25 fatty acid mimetics with known agonistic activity on retinoid X receptors (RXR) and/or peroxisome proliferator-activated receptors (PPAR).[64] From the fine-tuned model, 1000 SMILES were generated, starting from the fragment '–COOH' and ranked according to (i) their pharmacophore similarity to known bioactives,[78,79] and (ii) the computationally predicted biological target.[80] Five top-ranked compounds were selected and tested *in vitro* for their activity on PPAR and RXR receptors. Four out of five selected designs activated PPAR and/or RXR subtypes, and showed EC_{50} values ranging from 14±2 µM to 0.13±0.01 µM with different selectivity profiles[75] (Fig. 4).

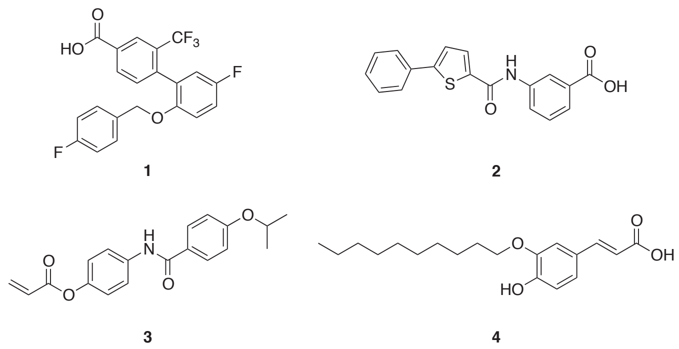

Fig. 4. Nuclear receptor modulators designed by a generative LSTM model.[75,76] **1**, **2**: Dual modulators of retinoid X receptors (RXR) and peroxisome proliferator-activated receptors (PPAR)[75]; **3**, **4**: Natural-product inspired modulators of RXR.[76]

A similar strategy was adopted to obtain anticancer peptides (ACPs) *de novo*.[52] An LSTM model based on amino-acid sequences of peptides[67] was pre-trained on a computer-generated set of 10,000 presumably alpha-helical and amphipathic peptides.[81] The model was fine-tuned on 26 anticancer peptides with low-micromolar activity against MCF7 cancer cells and diverse selectivity on human erythrocytes. After this transfer learning step, 1000 peptides were sampled, and the 12 top-ranking sequences based on similarity analysis and property prediction[81–83] were synthesized and tested *in vitro*. Ten out of the 12 peptides were bioactive (EC_{50} values between 16.1±0.3 µM and 101±4 µM on MCF7 cancer cells) and inherited the selectivity profile of the fine-tuning peptides with a therapeutic index ranging from 1 to >11 (Table 1).[81] Both of these prospective studies confirm the potential of generative LSTMs for focused compound libraries design, without the need of explicit structure–activity rules or QSAR models.

Recently, generative deep learning has been used for natural-product-inspired *de novo* design.[76] Natural products (NPs) and their molecular scaffolds have always been a source of inspiration to medicinal chemists, largely owing to their often unexplored scaffolds for the discovery of chemical probes and drugs.[84,85] In a prospective application, an LSTM model was utilized for generating new molecules inspired by NPs.[76] Again, a pre-trained LSTM model[75] was fine-tuned using only one, three or six natural RXR modulators from literature.[86–89] The number of NPs for LSTM fine-tuning was shown to affect the natural-product likeness[90] of the generated designs, thereby showing the potential of generative deep learning to bridge the chemical space of synthetic molecules and natural products in an application-tailored fashion. The model that was fine-tuned with six NPs was then used to generate two novel RXR modulators (**3** and **4**, EC_{50} values ranging from 15.7±0.8µM to 29±5 µM, Fig. 4).[76] The results also showed that the deep learning model can generate molecules lying at the interface between synthetic bioactive molecules and natural products.

4. Outlook

LSTMs and other generative models have been added to the medicinal chemist's toolkit, with visible success in pioneering experimental applications. We expect these prospective applications to be the first of many others to come. In fact, we envisage these tools to be rapidly incorporated in standard *de novo* design workflows in the near future, to accelerate the exploration of the chemical space in search of novel bioactive matter. At the same time, the rapid progress in the field of machine intelligence bears the potential to further the capacity and efficiency of deep learning methods for *de novo* design, especially in low-data regimes and with limited *a priori* chemical and biological information. AI supported *de novo* structure generation is a welcome first step. However, it is a long way from isolated proof-of-concept studies to developing new medicines. Progressing the computer-generated compounds towards a clinical drug candidate still remains to be demonstrated, as well as the anticipated significant cycle time and cost reductions in the generation of a novel clinical drug candidate. A curious but cautious approach may thus be advisable, given the required shift from the established discovery processes in medicinal chemistry to science that includes and values the contribution of AI.

Judging from other fields of application, *e.g.* in the music industry, the potential of generative AI for drug design is, however, largely untapped and might extend beyond serving as a simple generation/recommendation system. The capacity of certain deep learning algorithms to autonomously capture complex patterns in high-dimensional data might deliver new insights into synthesis optimization, ligand–receptor interaction, and underlying mechanisms of pharmacological action. An interpretable chemistry-savvy AI will undoubtedly be helpful (*e.g.* Fig. 5). Exploring this untapped potential will require a synergy between medicinal chemists, chemoinformaticians and statisticians, to map data inputs and model outputs to chemical and biological knowledge.

Table 1. Bioactive anticancer peptides generated *de novo* with an LSTM model.[52] The half-inhibitory concentration (IC_{50}) against MCF7 cancer cells, the half-hemolytic concentration (HC_{50}), and the therapeutic index (T.I. = HC_{50}/EC_{50}) are given (*mean* ± SEM, *N*=3).

Amino acid sequence	IC_{50} [*µM*]	HC_{50} [*µM*]	T.I.
KLWKKIEKLIKKLLTSIR	47±3	236±13	5.1±0.6
YIWARAERVWLWWGKFLSL	56±3	>400	>7
DLFKQLQRLFLGILYCLYKIW	47±4	132±16	2.8±0.6
AIKKFGPLAKIVAKV	95±4	>400	>4
RWNGRIIKGFYNLVKIWKDLKG	42±4	89±6	2.1±0.3
KVWKIKKNIRRLLHGIKRGWKG	34±4	>400	> 11
GFWARIGKVFAAVKNL	101±4	>400	> 4
AFLYRLTRQIRPWWRWLYKW	45.5±0.8	34±5	0.7±0.1
RIWGKHSRYIKIVKRLIQ	50±10	>400	>8
QIWHKIRKLWQIIKDGF	16.1±0.3	23±5	1.4±0.3

These efforts will be key to expand the abilities of creative machines for molecule generation and pattern recognition. Medicinal chemists will be confronted with increasingly more complex data and drug target hypotheses. At the same time, we have to concede our limited knowledge of human biology and pathophysiology. AI needs to provide answers flexibly, as drug discovery knowledge develops. If successful in the long run, the envisaged collaborative drug design engine may not only imitate but exceed human decision making as a core aspect of the molecular design process.

Acknowledgements

This research was financially supported by the Swiss National Science Foundation (SNSF, grant no. 205321_182176), the Novartis Forschungsstiftung (FreeNovation: AI in Drug Discovery), and the ETH RETHINK Initiative.

Competing interest statement

The authors declare the following potentially competing financial interest: G.S. is a co-founder of inSili.com LLC, Zurich, and a consultant to the life science industry.

Note

Several references point to non-peer-reviewed texts and preprints. These are cited to account for the actuality of the topic of this article.

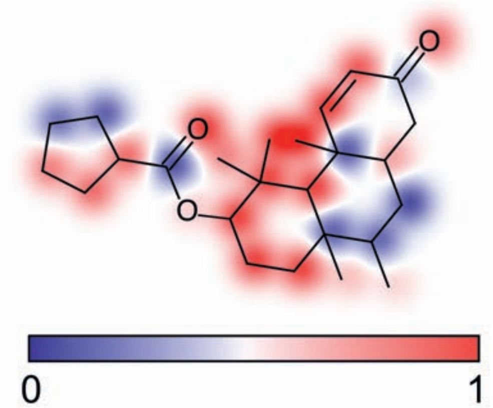

Fig. 5. Molecule designed *de novo* by a generative LSTM.[75] After pretraining,[75] the model was fine-tuned on 140 androgen receptor modulators[91] and used for molecule generation. Atoms are colored based on the LSTM probability assigned to the respective tokens, as utilized for SMILES generation (from 0 [low probability] to 1 [high probability]). High probability regions resemble known pharmacophoric features of AR binders, that is core hydrophobic regions and specific hydrogen-bond acceptor patterns, responsible for the hydrophobic interaction and the molecule positioning in the receptor pocket of the androgen receptor.[91–93] Interpretable AI might guide medicinal chemists in hit-to-lead optimization and provide additional mechanistic insights into structure-activity relationships.

[1] G. Schneider, *Nat. Mach. Intell.* **2019**, *1*, 128, DOI: 10.1038/s42256-019-0030-7.

[2] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, 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, G. *Nat. Rev. Drug Discov*. **2019**, accepted.

[3] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, S. *Chem. Rev*. **2019**, *119*, 10520, DOI: 10.1021/acs.chemrev.8b00728.

[4] Schneider G. *Nat. Rev. Drug Discov*. **2018**, *17*, 97, DOI: 10.1038/nrd.2017.232.

[5] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao S. *Nat. Rev. Drug Discov*. **2019**, *18*, 463, DOI: 10.1038/s41573-019-0024-5.

[6] S. Legg, M. Hutter, *Minds Mach.* **2007**, *17*, 391, DOI: 10.1007/s11023-007-9079-x.

[7] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, A. Clare, *Science* **2009**, *324*, 85, DOI: 10.1126/science.1165620.

[8] M. Q. Raza, A. Khosravi, *Renew. Sustain. Energy Rev.* **2015**, *50*, 1352, DOI: 10.1016/j.rser.2015.04.065.

[9] B. Li, B. Hou, W. Yu, X. Lu, C. Yang, *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 86, DOI: 10.1631/FITEE.1601885.

[10] Y. Seo, S. Kim, O. Kisi, V. P. Singh, *J. Hydrol.* **2015**, *520*, 224, DOI: 10.1016/j.jhydrol.2014.11.050.

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547, DOI: 10.1038/s41586-018-0337-2.

[12] P. Schneider, G. Schneider, *J. Med. Chem.* **2016**, *59*, 4077, DOI: 10.1021/acs.jmedchem.5b01849.

[13] F. Grisoni, C. S. Neuhaus, M. Hishinuma, G. Gabernet, J. A. Hiss, M. Kotera, G. Schneider, *J. Mol. Model.* **2019**, *25*, 112, DOI: 10.1007/s00894-019-4007-6.

[14] J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468, DOI: 10.1002/wcms.1183.

[15] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281, DOI: 10.1021/acs.accounts.8b00087.

[16] M. H. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604. DOI: 10.1038/nature25978.

[17] S. J. Y. Macalino, V. Gosu, S. Hong, S. Choi, *Arch. Pharm. Res.* **2015**, *38*, 1686, DOI: 10.1007/s12272-015-0640-5.

[18] D. Merk, F. Grisoni, K. Schaller, L. Friedrich, G. Schneider, *ChemistryOpen* **2019**, *8*, 7, DOI: 10.1002/open.201800156.

[19] S. Vilar, T. Lorberbaum, G. Hripcsak, N. P. Tatonetti, *PLoS ONE* **2015**, *10*, e0129974, DOI: 10.1371/journal.pone.0129974.

[20] R. M. Marrero, Y. Marrero-Ponce, S. J. Barigye, Y. E. Díaz, R. Acevedo-Barrios, G. M. Casañola-Martín, M. G. Bernal, F. Torrens, F. Pérez-Giménez, *SAR QSAR Environ. Res.* **2015**, *26*, 943, DOI: 10.1080/1062936X.2015.1104517.

[21] P. V. Oliferenko, A. A. Oliferenko, G. I. Poda, D. I. Osolodkin, G. G. Pillai, U. R. Bernier, M. Tsikolia, N. M. Agramonte, G. G. Clark, K. J. Linthicum, A. R. Katritzky, *PLoS ONE* **2013**, *8*, e64547, DOI: 10.1371/journal.pone.0064547.

[22] T. Miyao, H. Kaneko, K. Funatsu, *J. Chem. Inf. Model.* **2016**, *56*, 286. DOI: 10.1021/acs.jcim.5b00628.

[23] T. Miyao, M. Arakawa, K. Funatsu, *Mol. Inform.* **2010**, *29*, 111. DOI: 10.1002/minf.200900038.

[24] C. M. Dobson, *Nature* **2004**, *432*, 824, DOI: 10.1038/nature03192.

[25] R. V. Devi, S. S. Sathya, M. S. Coumar, *Appl. Soft Comput.* **2015**, *27*, 543. DOI: 10.1016/j.asoc.2014.09.042

[26] D. Douguet, E. Thoreau, G. Grassy, *J. Comput. Aided Mol. Des.* **2000**, *14*, 449. DOI: 10.1023/A:1008108423895.

[27] S. Kamphausen, N. Höltge, F. Wirsching, C. Morys-Wortmann, D. Riester, R. Goetz, M. Thürk, A. Schwienhorst, *J. Comput. Aided Mol. Des.* **2002**, *16*, 551. DOI: 10.1023/A:1021928016359.

[28] A. C. Anderson, *Chem. Biol.* **2003**, *10*, 787. DOI: 10.1016/j.chembiol.2003.09.002.

[29] L. G. Ferreira, R. N. Dos Santos, G. Oliva, A. D. Andricopulo, *Molecules* **2015**, *20*, 13384. DOI: https://doi.org/10.3390/molecules200713384.

[30] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, G. Schneider, *PLoS Comput. Biol.* **2012**, *8*, e1002380. DOI: 10.1371/journal.pcbi.1002380.

[31] B. Spänkuch, S. Keppner, L. Lange, T. Rodrigues, H. Zettl, C. P. Koch, M. Reutlinger, M. Hartenfeller, P. Schneider, G. Schneider, *Angew. Chem. Int. Ed.* **2013**, *52*, 4676. DOI: 10.1002/anie.201206897.

[32] A. Button, D. Merk, J. A. Hiss, G. Schneider, *Nat. Mach. Intell.* **2019**, *1*, 307. DOI: 10.1038/s42256-019-0067-7.

[33] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436, DOI: 10.1038/nature14539.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *J. Mach. Learn. Res.* **2014**, *15*, 1929. http://jmlr.org/papers/v15/srivastava14a.html.

[35] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, J. C. Phillips, *Proc. IEEE* **2008**, *96*, 879, DOI: 10.1109/JPROC.2008.917757.

[36] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241, DOI: 10.1016/j.drudis.2018.01.039.

[37] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, *Nat. Biotechnol.* **2019**, *37*, 1038, DOI: 10.1038/s41587-019-0224-x.

[38] K. Kim, S. Kang, J. Yoo, Y. Kwon, Y. Nam, D. Lee, I. Kim, Y.-S. Choi, Y. Jung, S. Kim, W.-J. Son, J. Son, H. S. Lee, S. Kim, J. Shin, S. Hwang, *Npj Comput. Mater.* **2018**, *4*, 67, DOI: 10.1038/s41524-018-0128-1.

[39] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360. DOI: 10.1126/science.aat2663.

[40] C. D. Elton, Z. Boukouvalas, M. D. Fuge, P. W. Chung, *Mol. Syst. Des. Eng.* **2019**, *4*, 828, DOI: 10.1039/C9ME00039A.

[41] Y. Bengio, *Found. Trends Mach. Learn.* **2009**, *2*, 1, DOI: 10.1561/2200000006.

[42] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, V. Svetnik, *J. Chem. Inf. Model.* **2015**, *55*, 263, DOI: 10.1021/ci500747n.

[43] K. Fukushima, *IEICE Tech. Rep. A* **1979**, *62*, 658.
[44] K. Fukushima, *Biol. Cybern.* **1980**, *36*, 193, DOI: 10.1007/BF00344251.
[45] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, *ArXiv Prepr. ArXiv151105644* **2015**.
[46] M. J. Kusner, B. Paige, J. M. Hernández-Lobato, in 'Proceedings of the 34th International Conference on Machine Learning-Volume 70', JMLR Org, **2017**, pp. 1945.
[47] H. Dai, Y. Tian, B. Dai, S. Skiena, L. Song, *ArXiv Prepr. ArXiv180208786* **2018**.
[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, in 'Advances in Neural Information Processing Systems', **2014**, pp. 2672.
[49] D. Foster, 'Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play', O'Reilly Media, Inc., **2019**.
[50] M. Sundermeyer, R. Schlüter, H. Ney, in 'Proceedings of the Interspeech', pp 194–197, Portland, OR, USA, **2012**.
[51] 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, DOI: 10.1021/acscentsci.7b00572.
[52] F. Grisoni, C. S. Neuhaus, G. Gabernet, A. T. Müller, J. A. Hiss, G. Schneider, *ChemMedChem* **2018**, *13*, 1300, DOI: 10.1002/cmdc.201800204.
[53] E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, *J. Chem. Inf. Model.* **2018**, *58*, 1194, DOI: 10.1021/acs.jcim.7b00690.
[54] D. Polykovskiy, A. Zhebrak, D. Vetrov, Y. Ivanenkov, V. Aladinskiy, P. Mamoshina, M. Bozdaganyan, A. Aliper, A. Zhavoronkov, A. Kadurin, *Mol. Pharm.* **2018**, *15*, 4398, DOI: 10.1021/acs.molpharmaceut.8b00839.
[55] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, A. Aspuru-Guzik, *ArXiv170510843 Cs Stat* **2017**.
[56] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2018**, *4*, 120, DOI: 10.1021/acscentsci.7b00512.
[57] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31, DOI: 10.1021/ci00057a005.
[58] M. Awale, F. Sirockin, N. Stiefl, J.-L. Reymond, *J. Chem. Inf. Model.* **2019**, *59*, 1347, DOI: 10.1021/acs.jcim.8b00902.
[59] D. Neil, M. Segler, L. Guasch, M. Ahmed, D. Plumbley, M. Sellwood, N. Brown, *ICLR 2018 Conf. Pap.* **2018**.
[60] E. J. Bjerrum, R. Threlfall, *ArXiv170504612 Cs Q-Bio* **2017**.
[61] L. C. Jain, L. R. Medsker, 'Recurrent Neural Networks: Design and Applications', CRC Press, Inc., Boca Raton, FL, USA, **1999**.
[62] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735, DOI: 10.1162/neco.1997.9.8.1735.
[63] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, *ArXiv14123555 Cs* **2014**.
[64] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700111, DOI: 10.1002/minf.201700111.
[65] 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, DOI: 10.1021/acs.jcim.6b00754.
[66] J. Arús-Pous, S. Johansson, O. Ptykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, O. Engkvist, *ChemRxiv 8639942* **2019**, DOI: 10.26434/chemrxiv.8639942.v1.
[67] A. T. Müller, J. A. Hiss, G. Schneider, *J. Chem. Inf. Model.* **2018**, *58*, 472, DOI: 10.1021/acs.jcim.7b00414.
[68] D. Nagarajan, T. Nagarajan, N. Roy, O. Kulkarni, S. Ravichandran, M. Mishra, D. Chakravortty, N. Chandra, *J. Biol. Chem.* **2018**, *293*, 3492, DOI: 10.1074/jbc.M117.805499.
[69] N. O'Boyle, A. Dalke, *ChemRxiv 7097960v1* **2018**, DOI: 10.26434/chemrxiv.7097960.v1.
[70] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminform.* **2017**, *9*, 48, DOI: 10.1186/s13321-017-0235-x.
[71] K. Zhou, S. Zhang, X. Meng, Q. Luo, Y. Wang, K. Ding, Y. Feng, M. Chen, K. Cohen, J. Xia, in 'Proceedings of the BioNLP 2018 Workshop', Association for Computational Linguistics, Melbourne, Australia, **2018**, pp. 166.
[72] M. Skalic, J. Jiménez, D. Sabbadin, G. De Fabritiis, *J. Chem. Inf. Model.* **2019**, *59*, 1205, DOI: 10.1021/acs.jcim.8b00706.
[73] B. Sattarov, I. I. Baskin, D. Horvath, G. Marcou, E. J. Bjerrum, A. Varnek, *J. Chem. Inf. Model.* **2019**, *59*, 1182, DOI: 10.1021/acs.jcim.8b00751.
[74] S. Harel, K. Radinsky, *Mol. Pharm.* **2018**, *15*, 4406, DOI: 10.1021/acs.molpharmaceut.8b00474.
[75] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, DOI: 10.1002/minf.201700153.
[76] D. Merk, F. Grisoni, L. Friedrich, G. Schneider, *Commun. Chem.* **2018**, *1*, 68, DOI: 10.1038/s42004-018-0068-1.
[77] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100, DOI: 10.1093/nar/gkr777.
[78] F. Grisoni, D. Merk, V. Consonni, J. A. Hiss, S. G. Tagliabue, R. Todeschini, G. Schneider, *Commun. Chem.* **2018**, *1*, 44, DOI: 10.1038/s42004-018-0043-x.
[79] F. Grisoni, D. Merk, R. Byrne, G. Schneider, *Sci. Rep.* **2018**, *8*, 16469, DOI: 10.1038/s41598-018-34677-0.
[80] D. Reker, T. Rodrigues, P. Schneider, G. Schneider, *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4067, DOI: 10.1073/pnas.1320001111.
[81] A. T. Müller, G. Gabernet, J. A. Hiss, G. Schneider, *Bioinformatics* **2017**, *33*, 2753, DOI: 10.1093/bioinformatics/btx285.
[82] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, G. P. S. Raghava, *Sci. Rep.* **2013**, *3*, 2984, DOI: 10.1038/srep02984.
[83] G. Gabernet, D. Gautschi, A. T. Müller, C. S. Neuhaus, L. Armbrecht, P. S. Dittrich, J. A. Hiss, G. Schneider, *Sci. Rep.* **2019**, *9*, DOI: 10.1038/s41598-019-47568-9.
[84] T. Rodrigues, D. Reker, P. Schneider, G. Schneider, *Nat. Chem.* **2016**, *8*, 531, DOI: 10.1038/nchem.2479.
[85] B. Shen, *Cell* **2015**, *163*, 1297, DOI: 10.1016/j.cell.2015.11.031.
[86] K. Nakashima, T. Murakami, H. Tanabe, M. Inoue, *Biochim. Biophys. Acta* **2014**, *1840*, 3034, DOI: 10.1016/j.bbagen.2014.06.011.
[87] H. Zhang, L. Li, L. Chen, L. Hu, H. Jiang, X. Shen, *J. Mol. Biol.* **2011**, *407*, 13, DOI: 10.1016/j.jmb.2011.01.032.
[88] H. Kotani, H. Tanabe, H. Mizukami, M. Makishima, M. Inoue, *J. Nat. Prod.* **2010**, *73*, 1332, DOI: 10.1021/np100120c.
[89] D. Merk, F. Grisoni, L. Friedrich, E. Gelzinyte, G. Schneider, *J. Med. Chem.* **2018**, *61*, 5442, DOI: 10.1021/acs.jmedchem.8b00494.
[90] P. Ertl, S. Roggo, A. Schuffenhauer, *J. Chem. Inf. Model.* **2008**, *48*, 68, DOI: 10.1021/ci700286x.
[91] F. Grisoni, V. Consonni, D. Ballabio, *J. Chem. Inf. Model.* **2019**, *59*, 1839, DOI: 10.1021/acs.jcim.8b00794.
[92] K. Pereira de Jésus-Tran, P.-L. Côté, L. Cantin, J. Blanchet, F. Labrie, R. Breton, *Protein Sci.* **2006**, *15*, 987, DOI: 10.1110/ps.051905906.
[93] H. Tamura, Y. Ishimoto, T. Fujikawa, H. Aoyama, H. Yoshikawa, M. Akamatsu, *Bioorg. Med. Chem.* **2006**, *14*, 7160, DOI: 10.1016/j.bmc.2006.06.064.