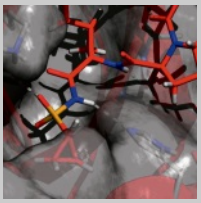
doi:10.2533/chimia.2018.70 Chimia 72 (2018) 70–71



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

Deep Learning Invades Drug Design and Synthesis

Josep Arús-Pous, Daniel Probst, and Jean-Louis Reymond*

*Correspondence: Prof. Dr. J.-L. Reymond, National Center of Competence in Research NCCR TransCure, Department of Chemistry and Biochemistry, University of Bern, jean-louis.reymond@dcb.unibe.ch

Keywords: Chemical space · Drug design · Machine learning · Retrosynthesis

Discovering a new drug goes as follows: given a medical need, identify the underlying biological mechanism, and design a molecule acting *via* this mechanism to produce the desired effects.[1] This sounds simple, but in truth it's not easy at all. Leaving aside biology, an important part of the problem is the chemistry: there are just too many molecules to choose from, perhaps as many as 10^{60} for all drug-like small molecules. Even with the help of computers one cannot enumerate more than a few billions of them, let alone predict their possible biological activity or how to synthesize them.[2] Of course, we don't really need to look at all these molecules. We can rely on trained medicinal chemists to make educated guesses on which ones to make and test first, often aided by modeling and automated searches, and this works well enough that trial-and-error cycles eventually succeed.[3]

Here comes a disruptive idea: can we automate the process and take the chemist out of the equation? Recent papers suggest that this might become possible using deep learning. Deep learning is an umbrella term for machine learning methods based on artificial neural networks,[4] by which a software first learns from training data, and then is able to perform complex tasks or predictions.[5] Parallelization of computer calculations, graphic cards and software frameworks such as TensorFlow[6] have made deep learning practical for writing music,[7] translating languages,[8] and even playing (and winning) Go.[9]

Let's first look at drug design, as published by Segler *et al.*,[10] Gupta *et al.*,[11] Ertl *et al.*,[12] and Popova *et al.*[13] All four studies rely on generative recurrent neural networks (RNN) of the type used for translating languages[8] and writing music[7] to learn to write chemical structures in the form of SMILES (Simplified molecular input line entry system), which is an unambiguous compact line notation for molecules which efficiently replaces nomenclature (Fig. 1a).[14] The RNN is first trained on SMILES from bioactive molecules taken from the database ChEMBL,[15] and surprisingly learns to write valid SMILES for ChEMBL-like molecules. In a second step, the RNN is fine-tuned with SMILES of molecules of a particular bioactivity class as input, which lead it to produce more SMILES of molecules for this specific class. Some of these generated molecules can be predicted to be as active as the molecules used for training the RNN, but have substantial structural differences to them, making them in principle novel, for example for the case of trypsin inhibitors (Fig. 1b).[11]

We now need a chemical synthesis for each of the molecules produced by the RNN to test if any of them are in fact active.[16]

Segler *et al.* also propose an automated solution for that problem based on deep learning.[17] They applied the coupling of Monte-Carlo Tree Search (MCTS) with neural networks introduced by Silver *et al.* for AlphaGo,[9] a computer program capable of beating masters in the game of Go. While the neural networks of AlphaGo were trained on Go games played by humans, Segler *et al.* trained their neural networks on reactions extracted from the Reaxys database.[18] The MCTS was then supplemented with these neural networks, speeding up the convergence and optimizing the outcome of the probabilistic Monte-Carlo algorithm. Applying this approach to the search for retrosynthesis routes has shown that a search algorithm enhanced by neural networks trained on human-produced data can surpass previous implementations such as expert systems which follow a strict set of rules organized in a decision tree both in speed and predictive quality.[19] The synthetic schemes proposed by the MCTS were judged quite convincing by synthetic chemists and sometimes identical to actual synthetic routes (Fig. 1c).[20]

As Segler *et al.* suggest,[17] automated retrosynthesis can be coupled to automated drug design and result in a fully automated drug design cycle. The machine is fed with a trained RNN and a target, automatically designs new possible bioactive compounds, checks for the ones that can be synthesized easily, attempts these syntheses,[21] and tests them in an automated bioassay. If any of the steps fails, the machine is patient and tries again and again, until some of the predicted molecules are successfully obtained and found to be active. The structure–activity data obtained from the synthesized molecules, both active and inactive compounds, can then be fed back to the RNN to guide further design cycles for optimization.

Can this approach work? Would it be faster and more economical than a project team? Would this be creative and solve problems? The authors of the publications discussed above are very much aware of the limitations of deep learning, which is extremely good at performing automated tasks within a defined range, but not necessarily more. Molecules created using deep learning might only look like real drugs, as much as music composed by deep learning sounds like the real thing but is not.[22] Since we don't know the answer, the confrontation of man against machine will be fascinating to follow, and the verdict will be severe: either the machine succeeds, perhaps because drug design does not need human genius like music does, and chemists will have to rethink how to work. Or the machine fails, sending computer scientists back to the drawing board.

Received: December 5, 2017

1] S. Sinha, D. Vohora, in 'Pharmaceutical Medicine and Translational Clinical Research', Academic Press, Boston, **2018**, pp. 19-32.
[2] M. Awale, R. Visini, D. Probst, J. Arus-Pous, J. L. Reymond, *Chimia* **2017**, *71*, 661.
[3] a) K. H. Bleicher, H. J. Bohm, K. Muller, A. I. Alanine, *Nat. Rev. Drug Discov.* **2003**, *2*, 369; b) K. Smietana, M. Siatkowski, M. Møller, *Nat. Rev. Drug Discov.* **2016**, *15*, 379.
[4] P. J. Werbos, 'The roots of backpropagation: from ordered derivatives to neural networks and political forecasting', Wiley-Interscience, **1994**.
[5] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85.
[6] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V.

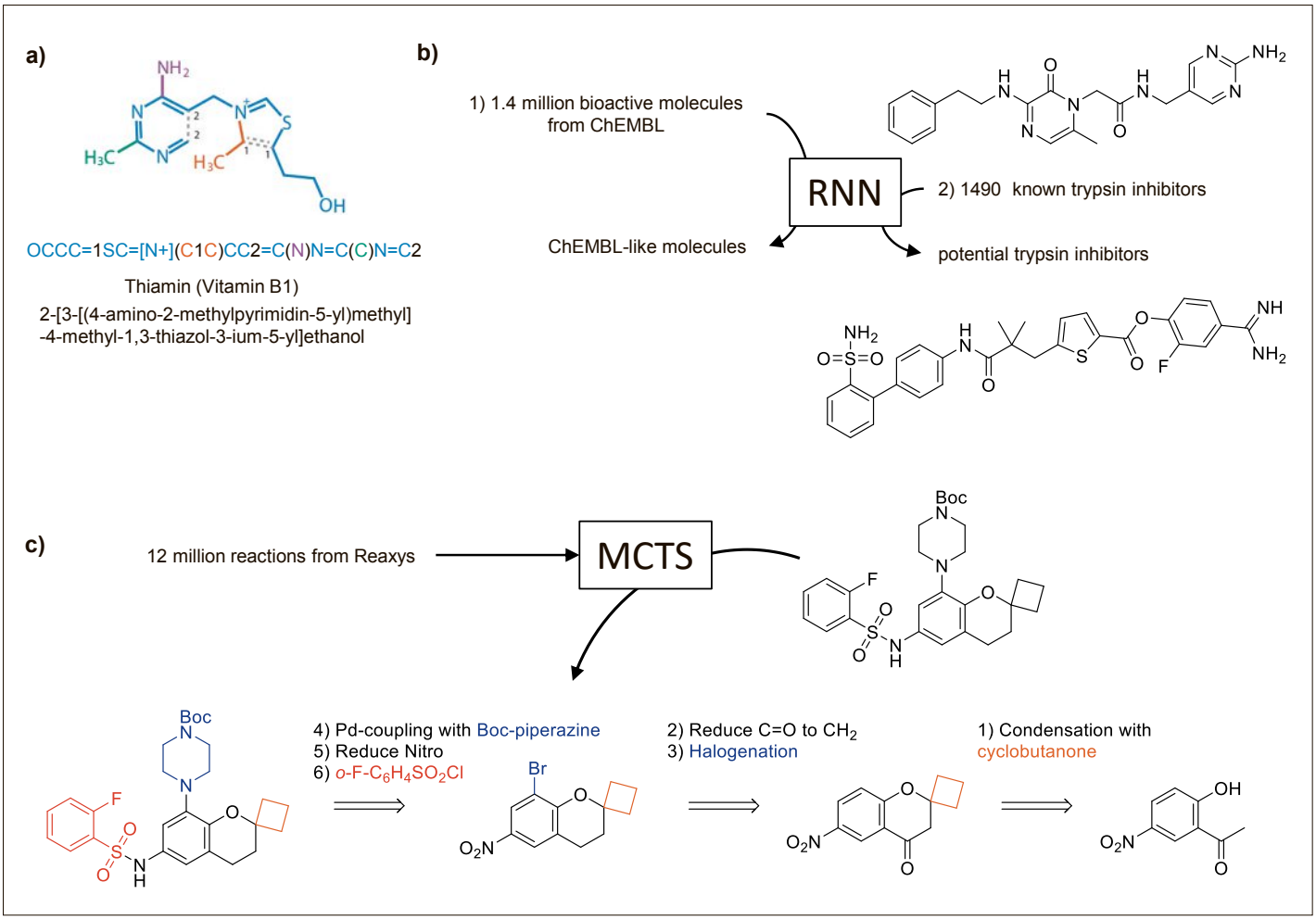


Fig. 1. a) Color-coded structural formula of thiamin, the corresponding SMILES notation (numbers in SMILES mark ring closures), and the systematic name. b) A generative recurrent neural network (RNN) is first trained with SMILES from the database ChEMBL, thus learning to write valid SMILES of ChEMBL-like molecules. The RNN is fine-tuned in a second step with bioactive compounds for a specific class (here trypsin inhibitors), and then produces valid SMILES representing potential new bioactive compounds of this class. Example molecules taken from ref. [11]. c) A Monte-Carlo Tree Search (MCTS) algorithm is enhanced by neural networks trained on reaction data from Reaxys. The algorithm can then propose retrosynthetic schemes for any molecule. The example retrosynthesis of a drug intermediate produced by the MCTS is identical to a documented synthetic route,[20] as discussed in ref. [17].

Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, G. Research, *arXiv:1603.04467* **2016**.

[7] D. Eck, J. Schmidhuber, Technical Report No. IDSIA-07-02, **2002**.

[8] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, J. Dean, *Trans. Assoc. Computational Linguistics* **2017**, *5*, 339.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **2016**, *529*, 484.

[10] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *arXiv:1701.01329* **2017**.

[11] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inform.* **2017**, *36*, 1700111.

[12] P. Ertl, R. Lewis, E. Martin, V. Polyakov, *arXiv:1712.07449v2*, **2017**.

[13] M. Popova, O. Isayev, A. Tropsha, *arXiv:1711.10907*, **2017**.

[14] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

[15] 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.

[16] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inform.* **2018**, *37*, 1700153.

[17] M. H. S. Segler, M. Preuss, M. P. Waller, *arXiv:1708.04202* **2017**.

[18] A. J. Lawson, J. Swienty-Busch, T. Géoui, D. Evans, in 'The Future of the History of Chemical Information', Vol. 1164, American Chemical Society, **2014**, pp. 127-148.

[19] a) E. J. Corey, X.-M. Cheng, 'The Logic of Chemical Synthesis', John Wiley & Sons, Inc., New York, **1995**; b) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928; c) S. Szymkuc, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904; d) M. H. S. Segler, M. Preuß, M. P. Waller, *arXiv:1702.00020* **2017**.

[20] R. V. S. Nigori, R. Badange, V. Reballi, M. Khagga, *Asian J. Chem.* **2015**, *27*, 2117.

[21] a) J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, *Science* **2015**, *347*, 1221; b) S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Myers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449.

[22] a) F. Marchesani, YouTube *https://www.youtube.com/watch?v=j60J1cGINX4* (accessed Dec 1, **2017**); b) F. Pachet, YouTube *https://www.youtube.com/watch?v=LSHZ_b05W7o* (accessed Dec 1, **2017**).

Can you show us your Medicinal Chemistry and Chemical Biology Highlight?
Please contact: Dr. Cornelia Zumbrunn, Idorsia Pharmaceuticals Ltd., Hegenheimermattweg 91, CH-4123 Allschwil,
E-mail: cornelia.zumbrunn@idorsia.com