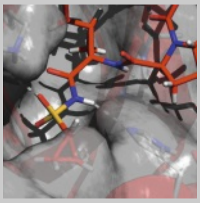
doi:10.2533/chimia.2018.648 Chimia 72 (2018) 648–649 © Swiss Chemical Society



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

Practical Aspects of Machine Learning for the Design-Synthesis-Purify-Assay Workflow

Finton Sirockin* and Nikolaus Stiefl*

*Correspondence: Dr. F. Sirockin, Dr. N. Stiefl, Novartis Institutes for Biomedical Research, Fabrikstrasse 2, Novartis Campus, CH-4056 Basel, E-mail: finton.sirockin@novartis.com; nikolaus.stiefl@novartis.com

Keywords: Automated synthesis · Auto-updating learning systems · Machine learning

As already discussed in this column,[1] Machine Learning, *via* its Deep Learning incarnation which came back into the limelight in the past few years,[2,3] or the more classical approaches, such as Multiple Linear Regression[4] (MLR), Random Forest[5] (RF), K-Nearest Neighbour[6] (KNN), *etc.*, is invading drug design and synthesis. Apart from learning how to make molecules[1] or optimizing reaction conditions,[7] what molecules to make is key in all medicinal chemistry projects. For this, both classification and regression methods have been successfully applied to predict activity (QSAR) and properties (QSPR), *cf.* references in refs [8,9].

Model performance, however, is greatly affected by the quantity and quality of available data. The typical situations encountered in real-life Medicinal Chemistry projects are (1) one or two active compounds are known as starting points, often from patents or literature, with no other information; (2) one or two mildly active starting points and a handful of inactives; (3) a few mildly active compounds and many inactives, a typical MTS or HTS situation; (4) a reasonably sized dataset with well spread activity data, a situation usually occurring in later stages of the drug design cycle. The two first cases make designing predictive models nearly impossible. The third case is more tractable, but one should probably start creating a model for inactive-compound prediction in that situation.

With recent progress in automated synthesis[10] and integrated Design – Synthesis – Purification – Assay(s) (DSPA) platforms,[11,12] auto-updating learning systems are key. For such systems to be efficient, they must combine good predictive power and fast model updating on multiple endpoints, with the capacity to automatically generate realistic and chemically tractable (in the context of an automated platform) virtual molecules to be assessed and selected by the models.

While deriving single endpoint predictions (*e.g.* activity against target, cellular activity, permeability, solubility, *etc.*) is often tractable,[13–16] the simultaneous optimization of multiple parameters,[17] frequently contradictory with each other, is a combination of various tough problems, among others: (1) uncertainty in and partial availability of experimental data; (2) different endpoints are driven by different chemical and physicochemical properties; (3) varying endpoint importance on the overall profile. Even though the Zeitgeist suggests that Artificial Intelligence (AI) can predict any task without understanding the underlying physical concepts, this is simply wrong in this case. As there is no golden bullet the theoretical medicinal chemist has to use the full armory of cheminformatics and statistical methods to address them.

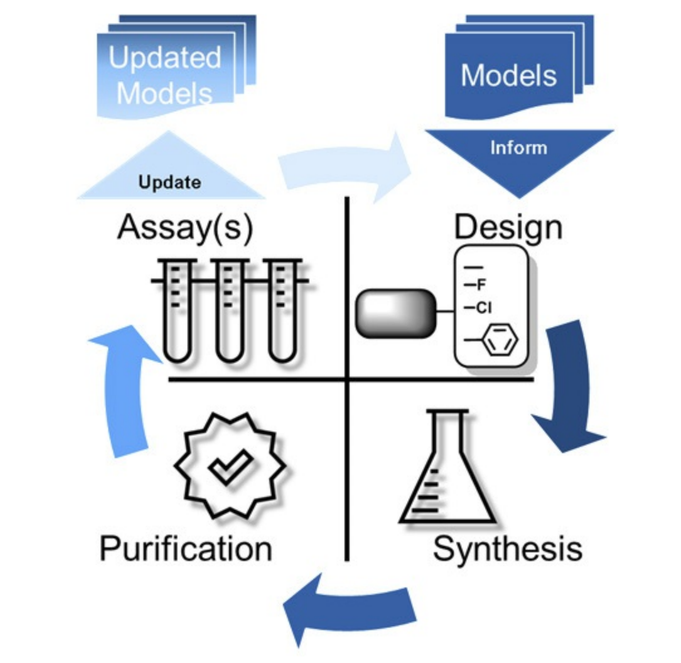


Fig. 1. Design-Synthesis-Purification-Assay iterative cycle.

The question of the applicability domain of a single model is key and warrants careful consideration of datasets and statistical methods.[18,19] If not enough data is available directly, surrogate methods can be used to fill data gaps (*e.g.* pQSAR[20]). If available data diversity is low – often driven by exclusive selection of compounds predicted to be the most active[21] – the data scientist's responsibility is to promote careful selection of the most informative compounds in order to support the model predictive power. This would remediate the problem of inconclusive model predictions, however generating and capturing negative data is too often seen as a luxury, which is a mistake.

The selection of descriptors relevant for the observable to be modelled is not trivial. Classical ADMET observables are often well modelled using physicochemical properties, while target activity, or other events involving molecular recognition, necessitates a combination of specific chemical features along with defined bulk properties, thus leading to models involving fragment-based descriptors, molecular fingerprints along with physicochemical descriptors.[22] However, if not done carefully, adding too many descriptors can lead to overtrained models[23] or the infamous spurious correlations.[24]

A more subjective task is the adaption of weights between the different endpoints in the course of an iterative optimization process. While in a regular Medicinal Chemistry project, the parameters are often optimized sequentially, a key advantage of auto-updating learning systems is the necessity to consider and implement those differing weights from the beginning. Here, the major responsibility of the data scientist is to remind the team to review and adapt the endpoints' varying importance during the course of the optimization. This in turn will cultivate a data science culture that will be necessary for the medicinal chemistry of the future.

With models in hand, molecules need to be virtually generated to be assessed by the model(s) and selected for synthesis. In the context of DSPA platforms, synthesis routes commonly involve few steps of common reactions limited by cost and building blocks availability.[25] In contrast to ref. [1], here the focus is on automatic compound generation under the aforementioned constraints. This leads to two avenues: either the building blocks are available internally or commercially or the final compounds can be synthesized from advanced intermediates.

For the former, recently, multiple vendors started to provide large sets of building blocks along with reactivity information (*e.g.* REAL Database provided by Enamine[26]). It has the advantage that the probability of success for synthesis in an automated synthesis platform context is reasonable. Beyond vendor offerings, companies leverage on their internal reactivity data pool.[27] This type of information can also be used for *de novo* design approaches[28] that fit the real-life scenario of focused library synthesis.

From the synthesis point of view, the latter avenue is more demanding and probably more prone to failure. Conversely, it allows the inclusion of project specific knowledge into the design process along with a higher potential for novelty. Even though virtual molecules can be generated with methods that do not rely on existing sets of building blocks such as Genetic Algorithms[29] (GA), Recurrent Neural Network (RNN) Deep Learning methods such as Long Short-Term Memory[30] (LSTM), Gated Recurrent Unit[31] (GRU), Stack-RNN,[32] Generative Adversarial Networks[33] (GAN) and combinations thereof,[34] none of these are specifically optimized to work with constraints of advanced intermediates, chemical reactivity and limited number of chemical steps.

For either method, in order not to generate molecules outside the applicability domain of the ML models it is necessary to either filter the generated molecules or focus the generative process itself towards the relevant portion of chemical space. Filtering can be done for both 2D[35] or 3D[36,37] approaches in a straightforward manner, typically using distances to the training set. However, one drawback of 3D methods is that they can be too computationally demanding. To focus the molecular generation itself, a generic model built on millions of drug-like molecules (*e.g.* ChEMBL[38]) is typically re-trained with molecules from the training set using so-called reinforcement techniques.[39]

Overall, even though the recent strong interest around AI would suggest that the new methods will solve all our problems, it is important to note that only by fully understanding the gist of the matter can one take full advantage of the potential of platforms such as DSPA. It will require substantial efforts from individuals to understand the science and techniques behind these methods, especially their scope and limitations. This, in combination with a radical change in company and laboratory culture, will provide the potential to fully leverage a brave new world of data science driven drug discovery and automation platforms.

Received: July 25, 2018

[1] J. Arús-Pous, D. Probst, J.-L. Reymond, *Chimia* **2018**, *72*, 70, DOI: 10.2533/chimia.2018.70.

[2] K. Baumann, G. Schneider, *Mol. Inform.* **2017**, *36*, DOI: 10.1002/minf.201780132.

[3] Y.-C. Lo, S. E. Rensi, W. Torng, R. B. Altman, *Drug Discov. Today* **2018**, DOI: 10.1016/j.drudis.2018.05.010.

[4] D. R. Cox, *J. R. Stat. Soc. Ser. B Methodol.* **1958**, *20*, 215.

[5] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947, DOI: 10.1021/ci034160g.

[6] W. Zheng, A. Tropsha, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185, DOI: 10.1021/ci980033m.

[7] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377, DOI: 10.1038/s41586-018-0307-8.

[8] H. Kubinyi, *Drug Discov. Today* **1997**, *2*, 538, DOI: 10.1016/S1359-6446(97)01084-2.

[9] K. Roy, R. N. Das, *Curr. Drug Metab.* **2014**, *15*, 346.

[10] K. Sanderson, *Nat. Rev. Drug Discov.* **2015**, *14*, 299, DOI: 10.1038/nrd4613.

[11] B. Desai, K. Dixon, E. Farrant, Q. Feng, K. R. Gibson, W. P. van Hoorn, J. Mills, T. Morgan, D. M. Parry, M. K. Ramjee, C. N. Selway, G. J. Tarver, G. Whitlock, A. G. Wright, *J. Med. Chem.* **2013**, *56*, 3033, DOI: 10.1021/jm400099d.

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

[13] C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178, DOI: 10.1038/194178b0.

[14] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616, DOI: 10.1021/ja01062a035.

[15] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395, DOI: 10.1021/jm00334a001.

[16] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959, DOI: 10.1021/ja00226a005.

[17] C. A. Nicolaou, N. Brown, *Drug Discov. Today Technol.* **2013**, *10*, e427, DOI: 10.1016/j.ddtec.2013.02.001.

[18] A. Tropsha, *Mol. Inform.* **2010**, *29*, 476, DOI: 10.1002/minf.201000061.

[19] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2014**, *57*, 4977, DOI: 10.1021/jm4004285.

[20] E. J. Martin, V. R. Polyakov, L. Tian, R. C. Perez, *J. Chem. Inf. Model.* **2017**, *57*, 2077, DOI: 10.1021/acs.jcim.7b00166.

[21] R. Kurczab, S. Smusz, A. J. Bojarski, *J. Cheminformatics* **2014**, *6*, 32, DOI: 10.1186/1758-2946-6-32.

[22] J. G. Cumming, A. M. Davis, S. Muresan, M. Haeberlein, H. Chen, *Nat. Rev. Drug Discov.* **2013**, *12*, 948, DOI: 10.1038/nrd4128.

[23] K. Baumann, N. Stiefl, *J. Comput. Aided Mol. Des.* **2004**, *18*, 549, DOI: 10.1007/s10822-004-4071-5.

[24] '15 Insane Things That Correlate With Each Other', *http://tylervigen.com/spurious-correlations*, accessed July 26, **2018**.

[25] A. G. Godfrey, T. Masquelin, H. Hemmerle, *Drug Discov. Today* **2013**, *18*, 795, DOI: 10.1016/j.drudis.2013.03.001.

[26] REAL Compound Libraries: New Chemical Space for Discovery - Enamine, *https://enamine.net/index.php?option=com_content&task=view&id=254%3F*, accessed July 24, **2018**.

[27] C. A. Nicolaou, I. A. Watson, H. Hu, J. Wang, *J. Chem. Inf. Model.* **2016**, *56*, 1253, DOI: 10.1021/acs.jcim.6b00173.

[28] F. Chevillard, P. Kolb, *J. Chem. Inf. Model.* **2015**, *55*, 1824, DOI: 10.1021/acs.jcim.5b00203.

[29] L. Weber, *Curr. Opin. Chem. Biol.* **1998**, *2*, 381, DOI: 10.1016/S1367-5931(98)80013-6.

[30] P. Ertl, R. Lewis, E. Martin, V. Polyakov, *ArXiv171207449 Cs Q-Bio* **2017**.

[31] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, *ArXiv14123555 Cs* **2014**.

[32] A. Joulin, T. Mikolov, *ArXiv150301007 Cs* **2015**.

[33] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharm.* **2017**, *14*, 3098, DOI: 10.1021/acs.molpharmaceut.7b00346.

[34] M. Popova, O. Isayev, A. Tropsha, *ArXiv171110907 Cs Stat* **2017**.

[35] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742, DOI: 10.1021/ci100050t.

[36] P. C. D. Hawkins, A. G. Skillman, A. Nicholls, *J. Med. Chem.* **2007**, *50*, 74, DOI: 10.1021/jm0603365.

[37] T. Cheeseright, M. Mackey, S. Rose, A. Vinter, *J. Chem. Inf. Model.* **2006**, *46*, 665, DOI: 10.1021/ci050357s.

[38] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res.* **2017**, *45*, D945, DOI: 10.1093/nar/gkw1074.

[39] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2018**, *4*, 120, DOI: 10.1021/acscentsci.7b00512.