doi:10.2533/chimia.2019.1001 Chimia 73 (2019) 1001–1005

Reducing the Concepts of Data Science and Machine Learning to Tools for the Bench Chemist

Richard A. Lewis*, Peter Ertl, Nadine Schneider and Nikolaus Stiefl

Abstract: Machine Learning and Data Science have enjoyed a renaissance due to the availability of increased computational power and larger data sets. Many questions can be now asked and answered, that previously were beyond our scope. This does not translate instantly into new tools that can be used by those not skilled in the field, as many of the issues and traps still exist. In this paper, we look at some of the new tools that we have created, and some of the difficulties that still need to be taken care of during the transition from a project run by an expert, to a tool for the bench chemist.

Keywords: Data science · Machine learning

Richard Lewis obtained his first degree in chemistry at Cambridge, followed by a PhD in drug design with Philip Dean. After a Fulbright fellowship with Tack Kuntz and a Royal Commission fellowship with M. Sternberg, he joined Jon Mason's group at RPR, then moving to Lilly to lead the European CADD team, before coming to Novartis in 2004. His major research interests include drug design, cheminformatics and development of novel methods for CADD. Richard is author of more than 60 publications covering all areas of CADD and cheminformatics.

Peter Ertl is Director and Leading Scientist at the Novartis Institutes for Biomedical Research in Basel. His major research interests include analysis and visualization of chemical space, bioisosteric design and interactive web tools supporting work of medicinal chemists. Peter is author of more than 100 publications covering all areas of cheminformatics and computational chemistry and several globally used cheminformatics algorithms and methods. In the cheminformatics community, he is best known as the author of the JSME JavaScript editor enabling molecule input on the web and a fast fragment based method to calculate molecular polar surface area.

Nadine Schneider obtained a BSc and MSc in Bioinformatics from the Saarland University in Germany. She did her PhD in Molecular Modeling in the group of Prof. Dr. Matthias Rarey at the University of Hamburg, Germany. In her PhD, she worked on a novel protein–ligand scoring function which was integrated in the commercial modeling software SeeSAR (BioSolveIT GmbH). In 2014, she joined the Novartis Institutes for BioMedical Research (NIBR) in Basel for a postdoc focusing on Cheminformatics and Data Science under supervision of Dr. Gregory Landrum and Dr. Nikolaus Stiefl. Since 2017 she is an investigator in the Computer-Aided Drug Design team in Global Discovery Chemistry in NIBR, Basel.

Nikolas Stiefl gained a degree in pharmacy at the University of Wuerzburg, after which he joined the computational chemistry group of Mike Hann at GlaxoWellcome. This was followed by a PhD in cheminformatics in the group of Prof. Knut Baumann and a postdoctoral fellowship at Eli Lilly. He then joined the CADD group of Richard Lewis at Novartis, working on a wide variety of topics such as drug discovery project support, software development, new scientific algorithms, and infrastructure tasks. He has published over 30 papers in these areas.

1. Introduction

Drug discovery has been a difficult endeavour with only low rates of success.[1] Many initiatives have been tried to improve these low rates, as even a minute improvement will lead to great rewards. During the history of drug discovery, a lot of data and information has been gathered but not organised and exploited in a meaningful way. If the bench chemist could harness the power of this knowledge, decision making could be made on a more rational basis, and hopefully the discovery goals could be reached with fewer compounds made and assayed. This requires the investment in data science to prepare the tools needed to assist the bench chemist. There is still so much that we do not know about human biology, that the combination of experience and domain knowledge with modelling will still be better than the two separated.

1.1 *What is Data Science?*

Data science can be organised into six areas, according to Donoho.[2]

1. Data exploration and preparation
2. Data representation and transformation
3. Computing with data
4. Data modelling
5. Data visualization and presentation
6. Science and data science

*Correspondence: Dr. R. A. Lewis, E-mail: richard.lewis@novartis.com
Computer-Aided Drug Design, Global Discovery Chemistry,
Novartis Institutes of BioMedical Research, CH-4056 Basel

We will use this scheme to organise the paper into similar sections, looking towards a final goal of providing tools to the chemist that can predict reliably, with information about confidence and applicability of the model, in a way that the results are simple to interpret by scientists without high levels of data science skills.

2. Data Exploration and Preparation

It is said[3] that 80% of the work in preparing models is to understand and clean the primary data. This curation effort is particularly important when dealing with public data sources derived from primary data.[4] Our own studies indicate that most biological assays have a log error of 0.2–0.3 units, depending on the complexity of the assay. We will return to this observation in the section on data presentation. There are also artefacts and outliers in the data, some of which might be genuine, others caused by more mundane effects such as edge effects on a screening plate. The use of geometric averages (arithmetic averages in log units) is encouraged as the data is explored. Internal tools have been developed to build rudimentary models from any data set, to explore how much initial signal there is in the data, before starting on the expensive data cleaning process. The data should also be examined in terms of the guidelines laid down for QSAR data sets.[5] It is often found that the data set is not suitable for modelling; a common cause is an insufficient number of actives in the data set. In Novartis, this is addressed by the design and synthesis of project-focussed libraries around hits, to amplify the signal.

2.1 *Data Representation and Transformation*

In most pharma companies, the primary observations (assay results) can be contained in several different databases and formats. Considerable efforts are being made to unify the means of access to the data, so that the user does not have to work hard to merge data from screening and *in vivo* models. Although most chemists deal in the concept of IC50, this is secondary data derived from fitting of a sigmoid function to a dose-response curve. Visualization of the curves is important to the evaluation of the quality of the data before transformation; we have developed tools to facilitate this manual checking of the primary data. As chemists are interested in structure–activity relationships, a suitable set of descriptors for structures needs to be found. In Novartis, descriptors available in rdkit[6] are used first although many other descriptors can be computed; we have developed new descriptors for chirality to address gaps seen in certain projects.[7] Descriptor selection is based on the observation to be modelled, and the size of the data set, so that results can be obtained in a reasonable time given the compute and disk resources available; the results should be interpretable by the chemist.

2.2 *Computing with Data*

In the past, model building was performed on any platform of experience, from Fortran, to within commercial packages. This caused many issues with maintenance (rebuilding, licensing) of models. We also have a goal of making models open source when publishing them. This has moved us towards the adoption of open source paradigms, such as python, together with rdkit and scikit learn as our primary tools to work with data. We can capture the data used to train and test the models, the versions of the software and the parameters and descriptors. For models that span across multiple projects (for example solubility models or similar), we set up a model review team of experts that carefully review models prior to their release. Checks are also put in place to see when the performance of a model is becoming unacceptably degraded for new data, so that it can be rebuilt when necessary. The use of standardised model building protocols makes this process much easier and with lower maintenance costs. There is still a philosophical issue for the users when models (and hence predictions) change. Versioning of models needs to be carefully documented.

2.3 *Data Modelling*

Models can either be explanatory or predictive. An example of a well-known explanatory model would be the rule of 5.[8] Explanatory models describe the data set they were built for very well, but do less well when the prediction being made extrapolates from the training data sets. This is a particular danger for neural net models, and careful control experiments have to be run. However, when the data set is very large, these techniques can be very valuable. In the field of generative chemistry, one can train a network to learn the grammar of chemical structures, expressed as SMILES strings, and generate many new valid structures;[9] examples are given in Fig. 1. The next level is to bias the structure generation towards the much smaller space of a discovery project, where there is insufficient data to build a standalone model.[10] Generative chemistry, like its predecessor *de novo* chemistry, can produce challenging structures; our SA score tool[11] has proved to be very useful to help prioritise suggested structures.

Predictive models are much more useful in the discovery context, as the goal is to move beyond what is known for provided structures that have been optimised and tailored to the goals of the discovery project. This requires careful selection of the data used to train the model, to test the model, and to validate the model as it is being used. We use time-splits,[12] as this is the most realistic approximation to the process of drug discovery, where new chemotypes are being discovered and elaborated on, rather than the continuous exploitation of existing scaffolds. The performance is driven often by the mean average error between the predictions and the experiment. This gives an indication of the confidence of the prediction. Gaussian process modelling is also being used, as the error in prediction is not constant across a data set (Fig. 2). It is smaller in regions of high numbers of structures and lower in sparsely populated regions. Predictive models may not be very explanatory, in the sense they do not explain the reasoning behind the prediction, so it is not obvious how to exploit the predictions through the design of new structures. This might be partially solvable through Inverse Design[13] or generative chemistry (covered above).

The data sets in pharma are also very sparse, that is, only a few percent of the data matrix is filled in (where rows represent compounds and columns assays). Scientists in the Novartis Emeryville site have developed methods for imputing the missing data[14] so that we can still model the missing data and feed that into further programs. It is necessary to make clear which data was measured and which imputed if the model results are to be included in the central data repository. Novartis is also part of the MELLODDY IMI project;[15] it has been shown that combining data from different sources can improve all models. The challenge is how to share information without compromising intellectual property. We are also exploring how to use public data sources, for example, ChemBl,[16] so that models and protocols can be shared.

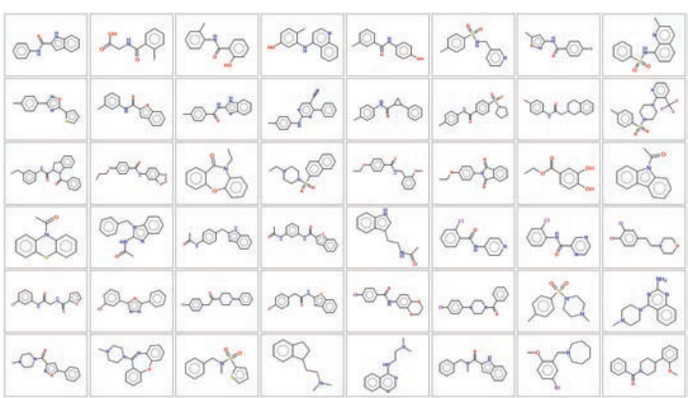

Fig. 1. A random selection of structures generated by our LSTM model.[9]

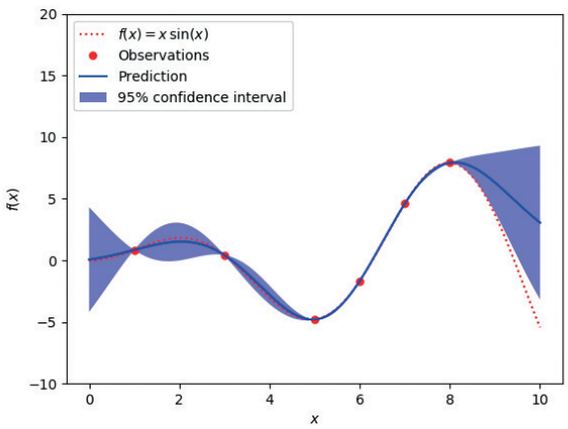

Fig. 2. An illustration of how estimation error in a prediction varies with data density for a Gaussian process model, taken from *scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html*.

2.4 *Data Visualisation and Presentation*

Many of the topics discussed above have been described from the point of view of a data scientist. However the models will have very little impact on the design- and decision-making process unless the conclusions are presented in an understandable way.[17] We have discussed the need for results to be presented as a prediction with a confidence. For categorical models (a compound is classed as soluble or insoluble), we can go a step further. The limits of what we mean by soluble or insoluble can be set in agreement with the users. The performance of the classification can then be used in the output (the prediction that this compound has a solubility < 10 microM will be right in 90% of cases). There is also the very important class, 'Not conclusive'. This means that the model does not know enough to make a confident prediction and would benefit from the data provided by making and assaying the compound. Identifying gaps in the model is key to improving the model. In addition, providing only predicted classes (*i.e.* anything that is not 'Not conclusive') with very high confidence to end users is important as it creates trust in the models for prioritising compounds prior to synthesis. When a numerical prediction is made, it is important for the prediction not to be given to excessive number of significant figures, especially when the input data is only accurate to 0.3 log units. This is a common failing of most user interfaces.

We have experimented with glowing molecules,[18] where the atoms in the molecule that contribute most to the observation are highlighted. It has been very important to also perform sensitivity analysis, as there is anecdotal evidence that the regions identified as important can vary strongly according to the nature of the training set. This should not be the case. The use of matched molecular pairs or series seems to be preferred by the chemist, as it can also tie in with the synthetic route being used, given more directed exploration of the SAR in a faster make-assay cycle.

2.4.1 *Organisation of Large Chemical Data Sets into Topics*

Handling large sets of molecules can be very complex and requires compromises that often come at the expense of interpretability. For this reason we have developed an alternative, novel approach called 'chemical topic modeling' which has been adopted from the text-mining field (the workflow is shown in Fig. 3).[19] This probabilistic framework offers an intuitive and meaningful way to organize and explore large chemical data sets. For example, on the ChEMBL database, a very heterogonous set of more than 1.6 million molecules, the method has proven its efficacy and robustness: a 100-topic model provided interesting topics like 'proteins', 'DNA' or 'steroids'. These rather general, yet nonetheless sensible and humanly understandable topics can provide the basis for further investigation. Using smaller data sets also more fine-grained information can be extracted: topics related to, for example, beta-secretase or sphingosine inhibition are found to be more commonly associated with certain chemical fragments than others (Fig. 4). A picture can be built up of the key pieces responsible for activity, suggesting further structural ideas for investigation.

3. Interfacing to Tools used by the Bench Chemist

Novartis was one of the early proponents of putting cheminformatics and modelling tools in the hands of the chemists.[20–22]

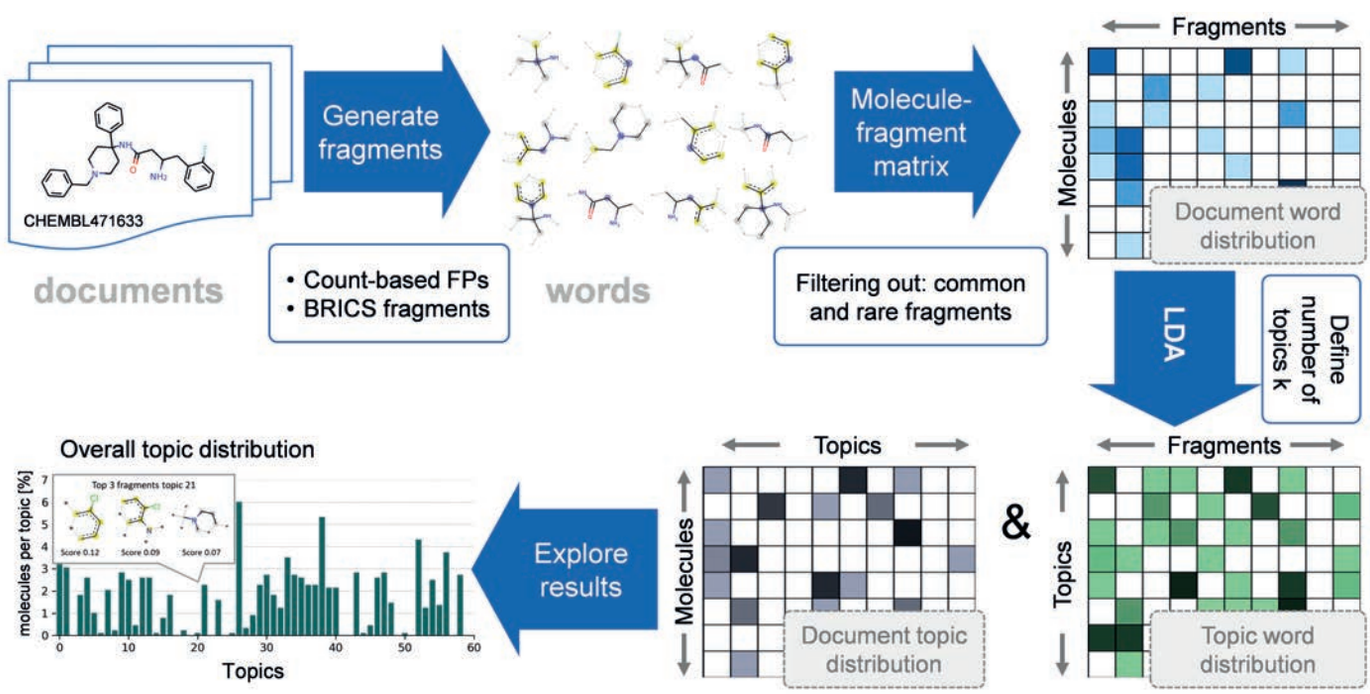

Fig. 3. Chemical topic modeling workflow. To make the connection to chemical topic modeling, the terms used in the context of topic modeling of text documents are shown as grey text. Picture adapted from Schneider *et al. JCIM* **2017**.[19] Reprinted with permission. Copyright 2017 American Chemical Society.

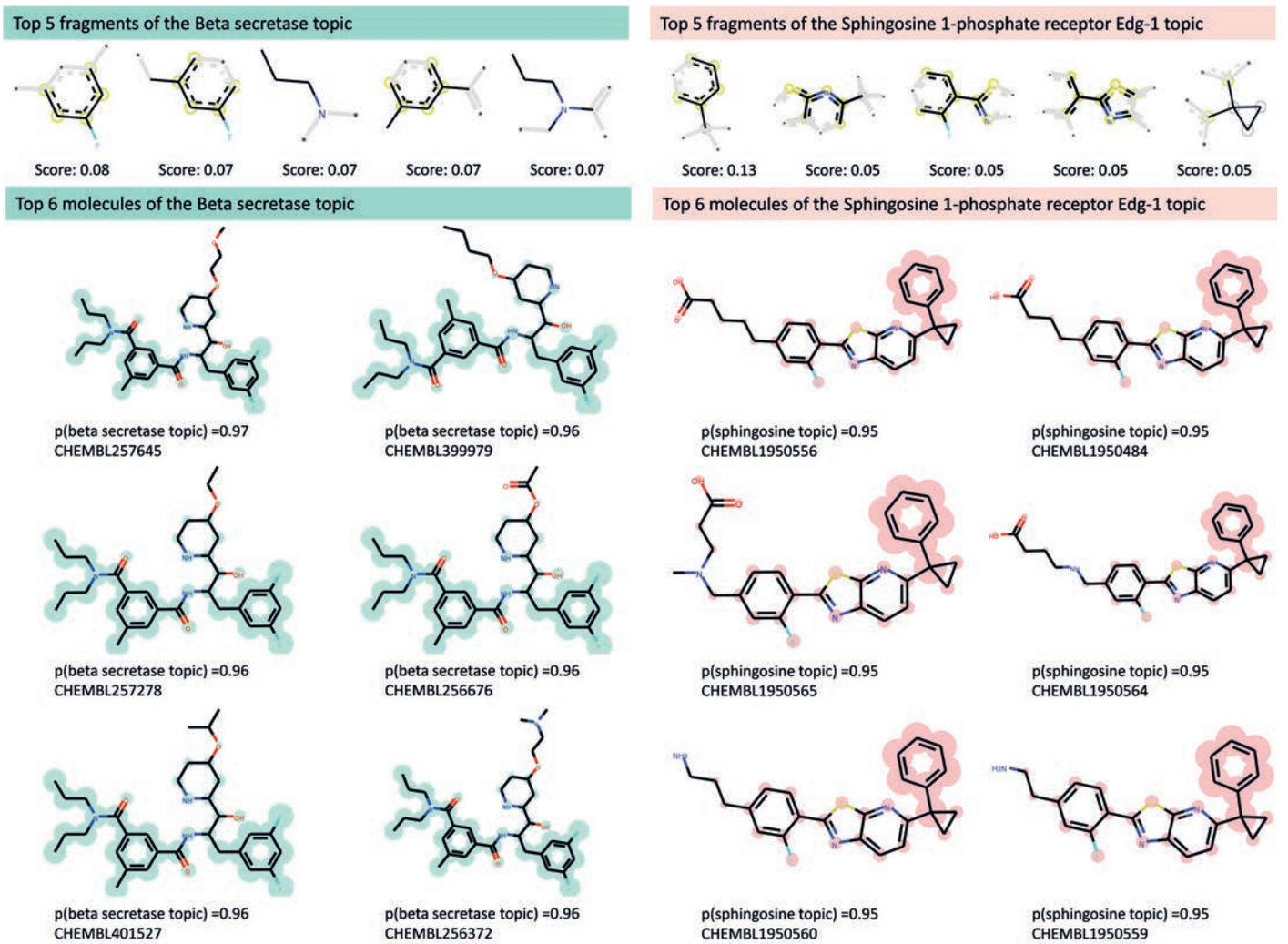

Fig. 4. Beta secretase and Sphingosine 1-phosphate receptor Edg-1 topics. (top) Five most probable fragments of both topics along with their probabilities. (bottom) Top six molecules of both topics. All of those have a probability of more than 90% for their topics. The topic is directly highlighted in turquoise/light orange within the compound structures. Reprinted with permission from Schneider *et al. JCIM* **2017**.[19] Copyright 2017 American Chemical Society.

An example might be the interactive web tool for navigating in property space (Fig. 5). One of the key principles used in tools such as FOCUS is modularity (Fig. 6). Over time key needs for project teams change (for example moving from mostly structure-enabled projects to more data-driven projects) and only if systems are able to add and remove functionality without having to retrain end users, will they be accepted. Also, recent internal interviews show that it is not feature completeness that is wanted to by end users but much more a tight integration of features to enable typical workflows.

As the trend moves into data analysis, a new way of presenting data is needed, with the facility for the users to manipulate the data themselves and develop their own local models, based on a deep understanding of their SAR and goals. Internally this is done *via* preprocessed data incorporated into Spotfire sessions that are easy to setup by non-experts but other solutions are available.

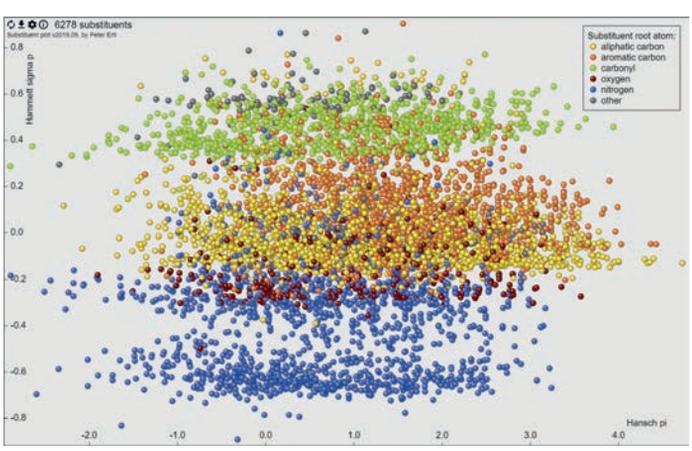

Fig. 5. Interactive web tool enabling navigation in the property space of organic substituents supporting bioisosteric design.

3.1 *Retrosynthetic Analysis*

An early goal in the field of Artificial Intelligence has been the retrosynthetic analysis of compounds.[23] Today, this goal seems to be within reach. Novartis is part of an industrial/academic partnership[24] to explore the application of machine learning to the corpus of reactions in the literature and to encode this into a system that can predict retrosynthetic routes and learn from new data. Other approaches are also being explored.[25]

3.2 *Science about Data Science*

Dohono defines the effectiveness of a tool (model) is related to the probability of deployment times the probability of effective results once deployed.[2] In this context, the probability of deployment should depend on the quality of the model, however that is defined, rather than the phenomenon being modelled. It is not the view here that inferior models should be accepted because they

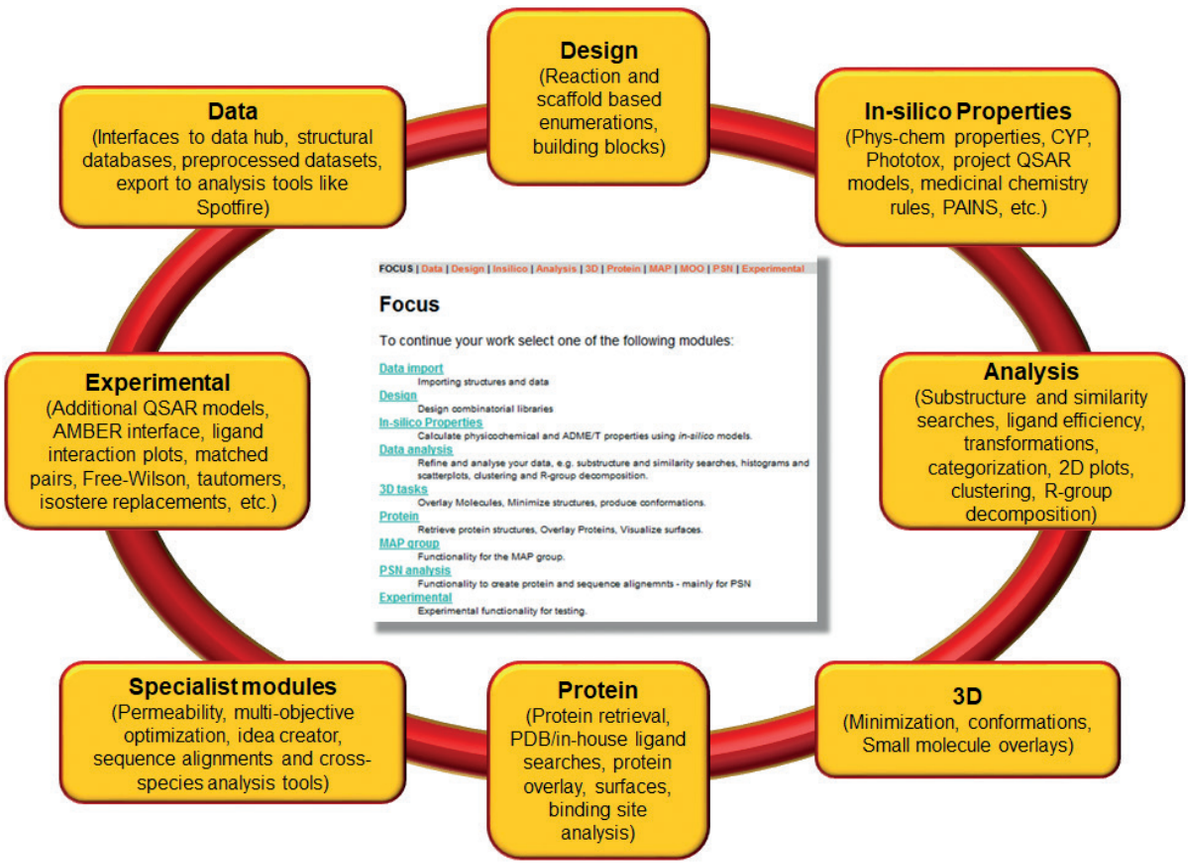

Fig. 6. FOCUS modules and their interconnections. Typical project tasks such as patent analysis or property filtering can be easily performed through working with multiple modules. Reprinted with permission from Stiefl *et al. JCIM* **2015**.[22] Copyright 2015 American Chemical Society.

model an important observation, unless the inferiority is clearly pointed out and understood by the user. The probability of effective results in pharma is more about how many decisions were influenced by the results of a model, which means capturing data on both compounds that are made and those that are deprioritised. This implies tracking of all parts of the design-make-test-analyse cycle. This is not an easy undertaking, but it is necessary to demonstrate the role that data science and modelling can play in making the drug discovery process more efficient.

4. Conclusions

Machine Learning and Data Science will undoubtedly have an increasing influence on the daily work of the bench chemist. The immaturity of the tools and techniques is such that care is still needed and the provision of data science skills within discovery chemistry teams will be vital. A thoughtful partnership is being developed in Novartis between these two roles, so that models and tools are used appropriately and lessons from their usage are fed back into the next cycle of data analysis and tool building. The future medicinal chemist will take robust tools as a given, but at present there is still much be learnt.

Received: October 23, 2019

[1] D. Cook, D. Sutherland Brown, R. Alexander, R. Macclesfield March, P. Morgan, G. Satterthwaite, M. N. Pangalos, *Nature Rev. Drug Discov*. **2014**, *13*, 419.
[2] D. Donoho, *J. Comput. Graph. Stat.* **2017**, *26*, 745.
[3] T. C. Redman, *Harvard Business Review*, **2018**, *https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless.*
[4] C. Kramer, T. Kalliokoski, P. Gedeck, A. Vulpetti, *J. Med. Chem*. **2012**, *55*, 5165.
[5] J.C. Dearden, M.T. Cronin, K.L. Kaiser, *SAR QSAR Environ. Res*. **2009**, *20*, 241.
[6] RDkit: *www.rdkit.org*
[7] N. Schneider, R. A. Lewis, N. Fechner, P. Ertl, *ChemMedChem* **2018**, *6*, 1315.
[8] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.
[9] P. Ertl, R. Lewis, E. Martin, V. Polyakov, *arXiv.* **2017**, 1717:07449.
[10] M. Awale, F. Sirockin, N. Stiefl, J.-L. Reymond, doi:10.26434/chemrxiv.7277354.v1.
[11] P. Ertl, A. Schuffenhauer, *J. Cheminf.* **2009**, *1*. doi: 10.1186/1758-2946-1-8.
[12] R. P. Sheridan, *J. Chem. Inf. Model.* **2013**, *53*, 783.
[13] R. A. Lewis, *J. Med. Chem.* **2005**, *48*, 1638.
[14] E. J Martin, V. R. Polyakov, X. Zhu, P. Mukherjee, L. Tian, X. Liu, doi: https://doi.org/10.1101/620864.
[15] *www.imi.europa.eu/projects-results/project-factsheets/melloddy*
[16] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, *Nucleic Acids Res.* **2014**, *42*, D1083.
[17] T. J. Ritchie, P. Ertl, R. A. Lewis, *Drug Discov. Today* **2011**, *16*, 65.
[18] S. Riniker, G. A. Landrum, *J. Cheminf.* **2013**, *5*, 43.
[19] N. Schneider, N. Fechner, G. A. Landrum, N. Stiefl, *J. Chem. Inf. Model.* **2017**, *57*, 1816.
[20] P. Ertl, P. Selzer, J. Mühlbacher, *Drug Discov. Today* **2004**, *2*, 201.
[21] R. Lewis, P. Ertl, E. Jacoby, M. Tintelnot-Blomley, P. Gedeck, R.M. Wolf, M.C. Peitsch, *Chimia* **2005**, *59*, 545.
[22] N. Stiefl, P. Gedeck, D. Chin, P. Hunt, M. Lindvall, K. Spiegel, C. Springer, S. Biller, C. Buenemann, T. Kanazawa, M. Kato, R. A. Lewis, E. Martin, V. Polyakov, R. Tommasi, J. van Drie, B. Vash, L. Whitehead, Y. Xu, R. Abagyan, E. Raush, M. Totrov, *J. Chem. Inf. Model.* **2015**, 55, 896.
[23] D. A. Pensak, E. J. Corey, 'Computer-Assisted Organic Synthesis', American Chemical Society, **1977**, DOI: 10.1021/bk-1977-0061.ch001
[24] *http://mlpds.mit.edu/*
[25] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604.