doi:10.2533/chimia.2019.1018 Chimia 73 (2019) 1018–1023

Exploring Chemical Space with Machine Learning

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

Abstract: Chemical space is a concept to organize molecular diversity by postulating that different molecules occupy different regions of a mathematical space where the position of each molecule is defined by its properties. Our aim is to develop methods to explicitly explore chemical space in the area of drug discovery. Here we review our implementations of machine learning in this project, including our use of deep neural networks to enumerate the GDB13 database from a small sample set, to generate analogs of drugs and natural products after training with fragment-size molecules, and to predict the polypharmacology of molecules after training with known bioactive compounds from ChEMBL. We also discuss visualization methods for big data as means to keep track and learn from machine learning results. Computational tools discussed in this review are freely available at *http://gdb.unibe.ch* and *https://github.com/reymond-group*.

Keywords: Chemical space · Data visualization · Deep learning · Molecular databases · Polypharmacology

Josep Arús-Pous studied Computer Engineering at the Polytechnic University of Catalonia (UPC). After working in the private sector for seven years, he gained a further MSc in Bioinformatics at the Pompeu Fabra University (UPF) in Barcelona. He is currently in the last year of a Marie Skłodowska-Curie European Industrial Doctorate (BIGCHEM) at both the University of Bern (with Prof. Jean-Louis Reymond) and AstraZeneca Gothenburg (with Dr. Hongming Chen and Dr. Ola Engkvist). His main research focus is on using intensive computational tools and deep learning to develop new methods of chemical space exploration.

Mahendra Awale is postdoctoral researcher in the CADD group at F. Hoffmann-La Roche, Basel, Switzerland. He studied pharmaceutical science at AISSMS college of pharmacy in Pune, India and obtained his PhD in Cheminformatics at the University of Bern, Switzerland, in 2015. After postdoc experience at University of Bern, he joined the CADD group at Roche in 2019. His research focuses on exploration of small molecule chemical space for drug discovery projects. To this end, he develops novel computational tools for virtual screening, visualization, target prediction, structure-activity relationship analysis. He also focuses on the integration of various machine learning approaches in drug design projects.

Daniel Probst has a bachelor's degree in computer science from the University of Applied Sciences in Biel with a focus on computer perception, virtual reality, and artificial intelligence and a master's degree in bioinformatics from the University of Bern with a focus on molecular biology. Before starting his tertiary education, he worked for six years as a systems engineer and software developer. He is currently a fourth year PhD student in the research group of Jean-Louis Reymond at the University of Bern. His main research interest is the application of algebraic computation and computer graphics to the exploration and visualization of chemical space.

Jean-Louis Reymond is chemistry professor at the University of Bern, Switzerland. He studied chemistry and biochemistry at the ETH Zürich and obtained his PhD in 1989 at the University of Lausanne on natural products synthesis. After a post-doc and assistant professorship at the Scripps Research Institute, he joined the University of Bern in 1997. His research focuses on expanding the accessible chemical space to novel scaffolds for drug design, including the synthesis of topologically diverse peptides such as dendrimers and innovative small molecules from the chemical universe database GDB (*http://gdb.unibe.ch*).

1. Introduction

The periodic table organizes the known 118 elements as rows and columns from which many of their properties can be understood and predicted. The situation is more complicated when combining elements to form molecules because the possibilities are endless and potentially overwhelming. Chemical space is a concept to organize molecular diversity by postulating that different molecules occupy different regions of a mathematical space, where the position of each molecule is defined by its structural and func-

*Correspondence: Prof. J.-L. Reymond, E-mail: jean-louis.reymond@dcb.unibe.ch
Department of Chemistry and Biochemistry, National Center for Competence in Research NCCR TransCure, University of Bern, Freiestrasse 3, CH-3012 Bern

tional properties.[1,2] One often refers to chemical space to describe a field of inquiry, *e.g.* "we work in that specific (chemical) space".

In our research we aim to develop computational tools to explicitly explore chemical space focusing on molecules relevant for drug discovery.[3] We have used direct enumeration algorithms to list all molecules that are possible up to a certain size following simple rules of chemical stability and synthetic feasibility, resulting in large databases such as GDB17 containing 166.4 billion possible molecules of up to 17 non-hydrogen atoms,[4,5] GDB4c containing 916,130 ring systems (hydrocarbons without acyclic bonds) with up to four saturated or aromatic rings,[6] as well as FDB17[7] and GDBMedChem,[8] which are subsets of GDB17 limited to molecules following fragment-likeness respectively medicinal chemistry criteria. The vast majority of molecules in GDB databases are yet unknown and therefore represent opportunities for discovery.

Assembling the GDB databases was based on simple enumeration algorithms. Recently, advancements in hardware and the availability of software libraries have made machine learning (ML) methods such as deep learning applicable as versatile tools to explore chemical space in the context of drug discovery.[9–11] In contrast to enumeration algorithms, ML approaches enable a computer to learn from data and then generate new data in an unsupervised manner by exploiting hidden rules that are present in the training data but cannot be explicitly formulated and labelled. Here, we review our contributions to this rapidly growing field of inquiry, focusing on molecule generation as an alternative to GDB enumeration and on bioactivity prediction. We also illustrate how data visualization allows one to keep track and retain control of ML output.

2. Enumerating Chemical Space with Molecular Generative Models

2.1 *GDB13 as a Benchmark*

Molecular generative models trained with a set of molecules represented as SMILES strings acquire the capability of generating SMILES strings representing new molecules that are close analogs of the molecules in the training set. Various neural network architectures have been shown to possess this capability, including recurrent neural networks (RNNs),[12–14] variational auto-encoders (VAE)[15–17] and generative adversarial networks (GAN).[18,19] Other approaches not using SMILES strings but the entire molecular graph have also been developed and showed promising results.[20–22] These examples focus on generating new and potentially bioactive molecules by training a neural network with molecules from databases such as ChEMBL,[23] which contains compounds known to be bioactive, and generate molecules around the known drug-like chemical space.

In collaboration with Ola Engkvist and coworkers at Astra Zeneca within the BIGCHEM project,[24] we have recently used RNN for SMILES generation for our GDB project to answer the question whether a RNN might be able to cover a complete section of chemical space when trained with a small subset of that chemical space, and thereby substitute exhaustive enumeration.[25] Given that generative models are sampled with replacement (*i.e.* any given molecule can be repeatedly sampled), a mathematical model based on the 'Coupon collector problem' was used to assess whether the model was able to create the whole database in a way that is complete (all molecules in it sampled), closed (no molecules outside of it sampled) and uniform (all molecules have the same probability of being sampled). We tested the approach with our database GDB13, which has a very large yet manageable size of 975 million molecules. To our surprise, we found that a RNN trained with a random sample of 1 million molecules (only 0.1% of GDB13) was capable of generating 69% of the database by sampling two billion molecules (Fig. 1). This first approach used canonical SMILES (the same representation for each molecule) and resulted in an uneven sampling of GDB molecules across molecular properties, having problems to sample highly cyclic molecules. Later, we found that RNN deliver a much more even sampling when trained with randomized SMILES (different representations at each epoch), yielding for the same experiment 83% of GDB-13 and a much more uniform distribution.[26] Moreover, models trained with smaller training set sizes (10,000 or even 1,000 molecules) were able to generate 63% and 34% of GDB-13 when sampled 2 billion times respectively, further highlighting the data augmentation capabilities of randomized SMILES. These studies show for the first time that deep generative models have the potential to generalize an entire chemical space from a limited subset and might therefore offer a viable alternative to exhaustive enumeration as an exploration tool.

2.2 *LSTM Neural Networks for Fragment-based Drug Analog Generation*

One of the defining features of our GDB databases is the very large number and high diversity of molecules at the scale of fragments, defined as molecules below 300 Daltons with only up to

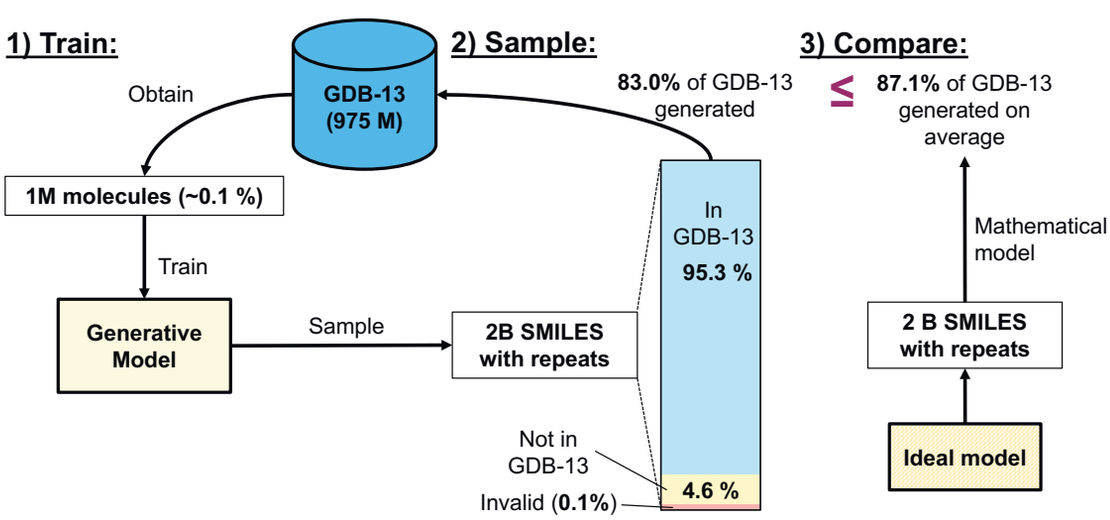

Fig. 1. Process of enumerating and benchmarking GDB-13 (975 million molecules without stereochemistry) using an RNN trained with SMILES strings. First, a sample of GDB13 is obtained and a generative model trained. After training the model is sampled with replacement 2 billion times and the resulting SMILES sorted between those that represent molecules in GDB13, those that do not, and those that are invalid. Lastly, the percent of GDB13 is compared with the upper bound obtained from an ideal model that evenly samples all molecules from GDB13.

three hydrogen bond donor and acceptor atoms.[27] For instance, the complete fragment subset of GDB17 contains 4.5 billion fragment-like molecules covering a broad range of heterocyclic, carbocyclic, aromatic and heteroaromatic compounds, which is much larger and more diverse than the approximately 100,000 fragments that can be collected from databases of known molecules and mostly comprise aromatic molecules.[7]

The long short-term memory generative neural network architecture (LSTM) is a type of RNN capable of processing sequential data and thus generating analogs of specific subclasses of bioactive compounds based on SMILES. To achieve this goal, one first trains the LSTM with SMILES from a drug-like database such as ChEMBL,[23] and then performs transfer learning with a specific subset, typically a family of inhibitors of a given enzyme or receptor.[12,13,22,28,29] In collaboration with the Novartis Institute of Biomedical Research in Basel, we recently investigated the effect of training such LSTMs with fragments from FDB17 and from other databases rather than with the entire ChEMBL database on the outcome of transfer learning.[30] To our surprise, we found that LSTMs trained with as little as 40,000 fragments, comprising fragment-like[27] molecules up to 17 atoms collected from various online catalogs, can generate new drug analogs similarly or better than LSTMs trained with ChEMBL itself, even when generating analogs of complex natural products such as macrocycles (Fig. 2). The best results were, however, not obtained with FDB17 but with commercially available fragments as training set. Strikingly, molecules generated after training with commercial fragments had a better synthetic accessibility and a higher predicted bioactivity than molecules generated after training with FDB17 fragments. Interestingly, we found that fragment-based LSTMs can generate a large number of close analogs of any drug irrespective of the set of fragments used from primary training. Furthermore, fragment-based LSTMs generate known analogs documented to be active, illustrating the validity of the approach.

3. Deep Neural Networks for Target Prediction

Enumerating chemical space exhaustively as in GDB or by sampling as with generative neural networks opens unlimited possibilities for innovation. To select molecules for synthesis and testing from this large diversity, one must however first perform a virtual screening (VS) step. VS attempts to predict which molecules have the highest probability to display a given biological activity based on a scoring function, which is often the similarity to a reference active molecule (ligand-based VS) or a docking score to a given protein binding pocket whose structure is known (structure-based VS).[31,32] In addition to VS, one must then also predict if the selected molecules might have additional off-target effects, a phenomenon known as polypharmacology and which is often undesirable.[33]

A variety of methods exist to predict polypharmacology by performing multiple comparisons with molecules of known bioactivity such as those in the ChEMBL database.[34] In our own implementation of such methods we reported the Polypharmacology Browser (PPB) as an online tool for off-target prediction based on ChEMBL data.[35] The defining feature of PPB was the combination of multiple molecular fingerprints for evaluation by a k-nearest neighbor classifier (k-NN), a well-known approach for similarity searches[36–38] which, however, had not been implemented for target prediction online tools. We recently further improved PPB and released PPB2, which uses a subset of ChEMBL containing molecules with high confidence activity datapoints against single protein targets.[39] In addition to k-NN classifiers, PPB2 uses other ML methods to improve predictions, considering naive Bayes classification as well as deep neural networks (DNN), which allows a direct performance comparison between these different methods. We find that combining k-NN with naive Bayes classification using the ECFP4 fingerprint[40] gives the highest target prediction performance across a broad range of targets, in particular in terms of precision. By contrast a k-NN classifier with ECFP4 performs best in terms of recall, while a DNN trained with ECFP4 performs well but not best across all methods (Fig. 3a). It should be noted that different ML methods have different numbers of parameters and hyper-parameters to be fitted and therefore different requirements for the size of training sets. It's known that the performance of classical machine learning methods improves with the increasing size of data and plateaus at some point, while the performance of deep learning models keeps improving steadily. In the presented target prediction study, the performance of the DNN model could have been affected due to the relatively small size of the training set.

We have used PPB2 to identify the target of a triazine designed as kinase inhibitor and identified as a nanomolar cytotoxic compound, but which turned out to be inactive on kinases by whole kinome profiling. PPB2 predicted that this triazine might in fact inhibit the enzyme LPAAT-β (lysophosphatidic acid acyl transferase β), a prediction which was later verified experimentally.[41] The correct prediction was based on a combination of a k-NN classifier with the Xfp pharmacophore fingerprint[42] and a naive Bayes model based on ECFP4[40] (Fig. 3b,c). These studies illustrate that DNN, despite being computationally more complex,

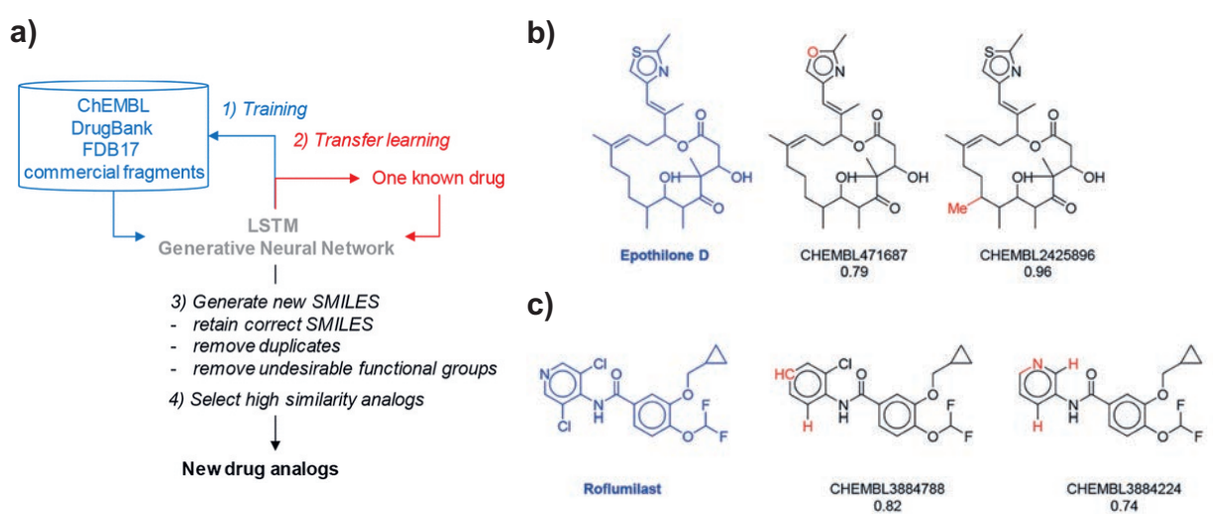

Fig. 2. **a**) LSTM workflow for generation of drug analogs. **b**,**c**) generated analogs for Epothilone D and Roflumilast. For each generated analog (black) the Avalon fingerprint Tanimoto similarity (*https://Sourceforge.Net/P/Avalontoolkit/Wiki/Home/*) with respect to the parent drug (blue) is shown and structural changes are highlighted in red.

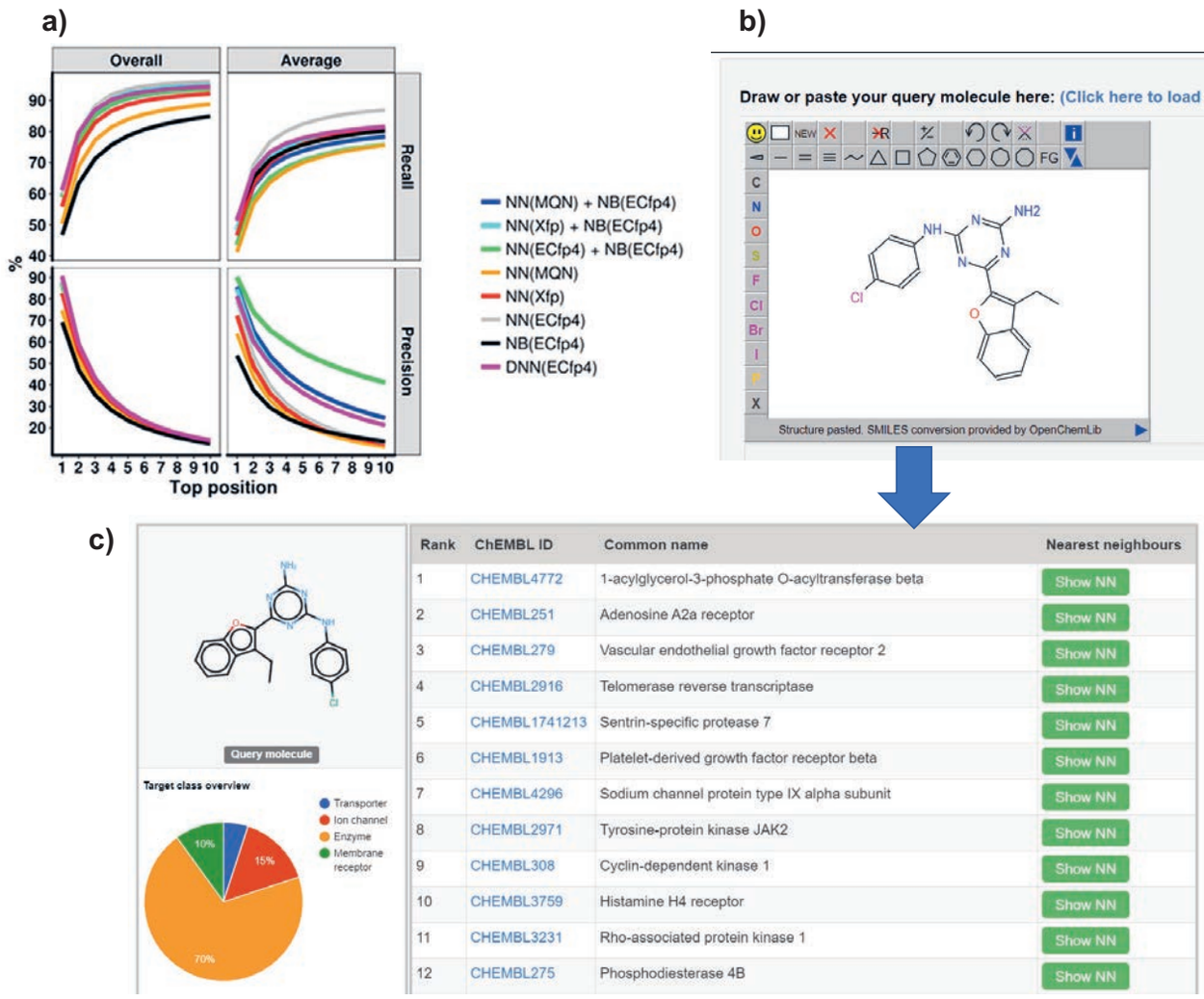

Fig. 3. **a**) Target prediction performance in a 10-fold cross-validation study for eight different methods available in the polypharmacology browser PPB2. **b**) PPB2 entry page with a triazine later identified to act on LPAAT-β as a query compound. **c**) PPB2 target prediction result for the triazine using k-NN(Xfp) + NB(ECfp4) method. k-NN = nearest neighbor, NB = Naïve Bayes model.

do not necessarily perform better for target prediction than more classical ML methods such as naive Bayes or k-NN classifiers.

4. Visualizing Chemical Space

Due to its extremely large size, chemical space can only be explored with the help of powerful computers and methods such as ML. ML represents an unprecedented opportunity to approach such big data problems, but carries with itself the risk of losing track of the results due to the complexity of the computation and the large amount of data generated. We believe that data visualization methods can play an essential role in ensuring that chemical space exploration produces interpretable results.

Our main approach to visualize chemical space[43–45] consists in generating interactive 2D or 3D maps featuring clouds of color-coded points, each point representing a molecule whose structure is made visible by pointing to it, and whose color encodes a particular property such as the similarity to a reference molecule or a molecular property such as size or polarity. The coordinates of each point on the map are calculated by applying dimensionality reduction to data points in a high-dimensional mathematical space defined by a molecular fingerprint. Our chemical space maps are available in the form of Java applets (Mapplets),[46,47] web applications (WebMolCS,[48,49] Faerun,[50,51] TMAP)[52] and virtual reality[53] applications enabling the visualization of up to several millions of datapoints simultaneously.

Our chemical space maps provide remarkable overviews and insights into the results of ML-based computations. For example, visualization of GDB13 as a MQN map[54] illustrates how the RNN generated database overlaps with the exhaustively enumerated one and in which regions the RNN lacks in performance (Fig. 4a/b).[25] In the context of generative models, interactive 3D-maps generated by WebMolCS play a key role in enabling a rapid inspection of the generated drug analogs to appreciate the types of structures generated (Fig. 4c).[30] Furthermore, in the context of our target prediction tool PPB2, a 3D visualization of target prediction using WebMolCS illustrates how k-NN predictions based on a pharmacophore fingerprint Xfp[42] were able to correctly predict LPAAT-β as the correct target of a cytotoxic triazine. This visualization shows that the close chemical space vicinity was also populated with reference molecules pointing to A2aR and VEGFR-2 as other possible targets, against which the compound was, however, inactive (Fig. 4d).[41]

5. Conclusion and Outlook

As illustrated in this review, ML can usefully complement and extend more classical algorithms for exploring chemical space by helping in generating molecules and predicting their bioactivity. Multiple ongoing efforts are rapidly expanding the range of questions that can be asked about chemical space using ML. We anticipate that implementing computer-assisted synthesis planning (CASP)[55,56] in the framework of our GDB project will soon allow us to focus our enumeration on readily synthesizable compounds and reach beyond our current limit of 17 atoms to discover new drugs with innovative chemical structures. Chemical space visual-

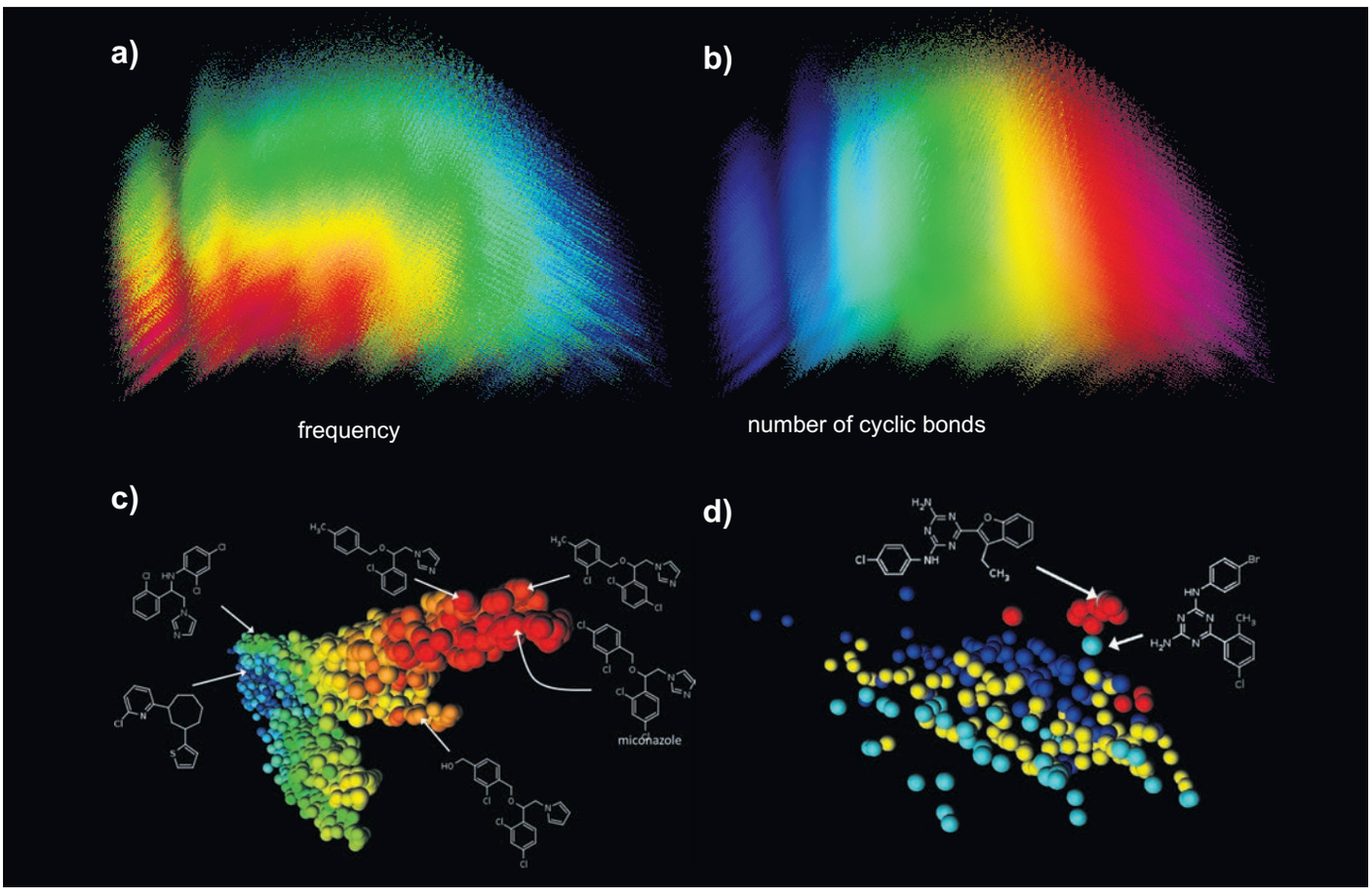

Fig. 4. **a**) Map of GDB13 generated by RNN colored by frequency (*i.e.* how often is a molecule sampled by an RNN). The colors range from blue (lowest value) to cyan, yellow read and magenta (highest value). The map was obtained by calculating the 42-D MQN (Molecular Quantum Numbers) fingerprint for each molecule, performing dimensionality reduction by principal component analysis, and plotting the (PC1, PC2) plane. **b**) MQN map of GDB13 generated by RNN colored by the number of cyclic bonds per molecule. **c**) 3D-similarity map (using substructure fingerprint) of miconazole analogs produced using fragment-based LSTM, with selected examples of analogs. **d**) 3D-similarity map (using Xfp pharmacophore fingerprint) of the triazine and its nearest neighbors from ChEMBL. Red: triazines investigated in the study, Cyan: LPAAT-β compounds, blue: A2aR compounds, yellow: VEGFR-2 compounds.

ization tools will play a decisive role in this project by allowing us to keep track and learn from the results produced by ML.

Acknowledgements

This work was supported financially by the NCCR TransCure and Novartis Pharma Basel, Switzerland. J.A.P. is supported financially by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 676434, 'Big Data in Chemistry' ('BIGCHEM,' *http://bigchem.eu*).

Received: September 28, 2019

[1] P. Kirkpatrick, C. Ellis, *Nature* **200**4, *432*, 823.
[2] M. Awale, R. Visini, D. Probst, J. Arus-Pous, J. L. Reymond, *Chimia* **2017**, *71*, 661.
[3] J. L. Reymond, L. Ruddigkeit, L. C. Blum, R. Van Deursen, W*IREs comput. Mol. Sci.* **2012**, *2*, 713.
[4] L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864.
[5] L. Ruddigkeit, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 56.
[6] R. Visini, J. Arus-Pous, M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 2707.
[7] R. Visini, M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 700.
[8] M. Awale, F. Sirockin, N. Stiefl, J.-L. Reymond, *Mol. Inf.* **2019**, *38*, 1900031.
[9] E. Gawehn, J. A. Hiss, G. Schneider, *Mol. Inf.* **2016**, *35*, 3.
[10] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241.
[11] G. Schneider, *Nat. Rev. Drug Discov*. **2018**, *17*, 97.
[12] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminf.* **2017**, *9*, 48.
[13] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci*. **2018**, *4*, 120.
[14] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700111.
[15] T. Blaschke, M. Olivecrona, O. Engkvist, J. Bajorath, H. Chen, *Mol. Inf.* **2018**, *37*, 1700123.
[16] 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.
[17] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharm.* **2017**, *14*, 3098.
[18] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, *ChemRXiv* **2017**, 10.26434/chemrxiv.5309668.v3.
[19] O. Prykhodko, S. Johansson, P.-C. Kotsias, J. Arús-Pous, E. J. Bjerrum, O. Engkvist, H. Chen, *10.26434/chemrxiv.8299544.v3* **2019**, DOI 10.26434/chemrxiv.8299544.v3.
[20] W. Jin, R. Barzilay, T. Jaakkola, *arXiv:1802.04364 [cs, stat]* **2018**.
[21] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, P. Battaglia, *arXiv:1803.03324 [cs, stat]* **2018**.
[22] Y. Li, L. Zhang, Z. Liu, *J. Cheminf.* **2018**, *10*, 33.
[23] 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. Magarinos, J. P. Overington, G. Papadatos, I. Smit, A. R. Leach, *Nucleic Acids Res*. **2017**, *45*, D945.
[24] I. V. Tetko, O. Engkvist, U. Koch, J. L. Reymond, H. Chen, *Mol. Inf.* **2016**, *35*, 615.
[25] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, O. Engkvist, *J. Cheminf*. **2019**, *11*, 20.
[26] J. Arús-Pous, S. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, O. Engkvist, *ChemRXiv* **2019**, 10.26434/chemrxiv.8639942.v2.
[27] M. Congreve, R. Carr, C. Murray, H. Jhoti, *Drug Discov. Today* **2003**, *8*, 876.
[28] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.
[29] P. Ertl, R. Lewis, E. Martin, V. Polyakov, *arXiv:1712.07449v2* **2017**.

[30] M. Awale, F. Sirockin, N. Stiefl, J.-L. Reymond, *J. Chem. Inf. Model*. **2019**, *59*, 1347.
[31] T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martinez-Mayorga, T. Langer, K. Cuanalo-Contreras, D. K. Agrafiotis, *J. Chem. Inf. Model*. **2012**, *52*, 867.
[32] J. Lyu, S. Wang, T. E. Balius, I. Singh, A. Levit, Y. S. Moroz, M. J. O'Meara, T. Che, E. Algaa, K. Tolmachova, A. A. Tolmachev, B. K. Shoichet, B. L. Roth, J. J. Irwin, *Nature* **2019**, *566*, 224.
[33] A. Anighoro, J. Bajorath, G. Rastelli, *J. Med. Chem.* **2014**, *57*, 7874.
[34] A. Lavecchia, C. Cerchia, *Drug Discov. Today* **2016**, *21*, 288.
[35] M. Awale, J. L. Reymond, *J. Cheminf*. **2017**, *9*, 11.
[36] N. Salim, J. Holliday, P. Willett, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435.
[37] J. Chen, J. Holliday, J. Bradshaw, *J. Chem. Inf. Model.* **2009**, *49*, 185.
[38] G. M. Sastry, V. S. S. Inakollu, W. Sherman, *J. Chem. Inf. Model.* **2013**, *53*, 1531.
[39] M. Awale, J.-L. Reymond, *J. Chem. Inf. Model.* **2019**, *59*, 10.
[40] D. Rogers, M. Hahn, *J. Chem. Inf. Model*. **2010**, *50*, 742.
[41] M. Poirier, M. Awale, M. A. Roelli, G. T. Giuffredi, L. Ruddigkeit, L. Evensen, A. Stooss, S. Calarco, J. B. Lorens, R.-P. Charles, J.-L. Reymond, *ChemMedChem* **2019**, *14*, 224.
[42] M. Awale, J. L. Reymond, *J. Chem. Inf. Model*. **2014**, *54*, 1892.
[43] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157.
[44] H. Geppert, M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 205.
[45] J. L. Medina-Franco, R. Aguayo-Ortiz, *Mol. Inf.* **2013**, *32*, 942.
[46] M. Awale, R. van Deursen, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 509.
[47] M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2015**, *55*, 1509.
[48] M. Awale, D. Probst, J. L. Reymond, *J. Chem. Inf. Model*. **2017**, *57*, 643.
[49] C. Delalande, M. Awale, M. Rubin, D. Probst, L. C. Ozhathil, J. Gertsch, H. Abriel, J.-L. Reymond, *Eur. J. Med. Chem*. **2019**, *166*, 167.
[50] D. Probst, J. L. Reymond, *Bioinformatics* **2018**, *34*, 1433.
[51] A. Capecchi, M. Awale, D. Probst, J. L. Reymond, *Mol. Inf*. **2019**, *38*, 1900016.
[52] D. Probst, J.-L. Reymond, *ChemRXiv* **2019**, 10.26434/chemrxiv.9698861.v1.
[53] D. Probst, J. L. Reymond, *J. Chem. Inf. Model.* **2018**, *58*, 1731.
[54] R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2010**, *50*, 1924.
[55] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281.
[56] A. Thakker, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Chem. Sci.* **2019**, *https://doi.org/10.1039/C9SC04944D*.