Integrating data-driven computer-aided synthetic planning with AI-based generative drug design

Hamza Tajmouati (1), Maud Parrot (1), Adam Skiredj (1), Robin Fourcade (1), Nicolas Do-Huu (1), Quentin Perron (1) & Yann Gaston-Mathé (1)
1) Iktos, 65 rue de Prony 75017 Paris, France

Introduction

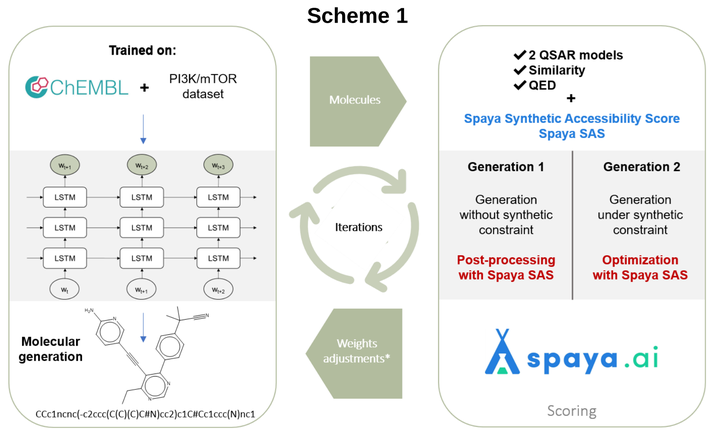
Deep learning approaches have become prevalent in recent years in the field of *de novo* molecular design.[1] Among them, the use of recurrent neural network (RNN) as SMILES generators after training on a large data set of molecules appears to be particularly promising.[2] Iktos is a well-known player in the field with proprietary technologies and know-how, practical experience of applying generative models to many real-life drug discovery projects, and, more recently, the release of Makya™, the first *de novo* design software platform powered by deep generative models. However, synthetic accessibility to the designed molecules remains a practical limitation of such generative approaches. Over the last few years, computer-aided synthetic planning (CASP) has also evolved, driven by progress in data-driven and machine learning (ML) / artificial intelligence (AI) powered approaches.[3]. Iktos has recently released it's own synthetic planning software, Spaya™, built upon Iktos's proprietary developments in the field. In this work, we present the integration of data-driven CASP with generative AI, i.e., molecular generation under the constraint of synthetic accessibility computed by Spaya API, an API (Application Programming Interface) providing a synthetic accessibility score computed by the CASP algorithm implemented in Spaya.

1 Materials & Methods: Generative framework & Spaya Synthetic Accessibility Score

To serve our proof of concept for the integration of AI-based CASP and generative *de novo* drug design, a library of 463 structurally homogeneous PI3K and mTOR inhibitors was selected.[4] This dataset served as a simplified proxy for a real life multi-parametric lead optimization project with four criteria to be optimized: Pi3K inhibition (pKi measured on the Phosphoinositide 3-Kinase), mTOR inhibition (pKi measured on the mechanistic Target Of Rapamycin), Tanimoto similarity (on ECFP) to the initial dataset and Quantitative Estimate of Drug-likeness (QED).[5] All the molecules included in the training set has pKi values ≤ 7 (PI3K) and ≤ 8.5 (mTOR) respectively (**Figure 1**). Compounds with pKi measured above these thresholds were voluntarily excluded as optimizing these parameters is one of the objectives of the subsequent generative process. Two QSAR models (for PI3K and mTOR pKi measures) were built using an extended connectivity fingerprint (ECFP) molecular representation and a ridge regression model. K-fold (K=4) cross validation along with Tree-structured Parzen Estimator was used for model selection (penalty parameter) and ECFP parameters (radius and size). These two QSAR models were used as PI3K and mTOR scorers during the ensuing generative procedure, in addition to QED and similarity objectives (**Generation 1**). The goal of the multi-parametric optimization (MPO) was therefore to generate molecules that satisfy the following specification: Pi3K(mol) >7; mTOR(mol) >8.5; Similarity(mol) >0.5; QED(mol) >0.5. For the second generation (**Generation 2**), a fifth criterion of synthetic accessibility as defined in the dedicated section (Spaya SAS = Spaya Synthetic Accessibility Score > 0.6) was added to the MPO specification to guide the optimization (**Scheme 1**).The molecular generations were performed using a deep Long Short-Term Memory (LSTM) algorithm generating SMILES strings corresponding to unique compounds as described in GuacaMol.[6] The LSTM was first trained on the ChEMBL database, using teacher forcing, to build a character-based language model for generating SMILES strings. Yet, to comply with applicability domain constraints of QSAR models, the generated molecules should have a significant structural similarity to the project dataset (which is a good proxy for applicability domains). Thus, the previous LSTM model was re-trained in teacher forcing on the PI3K/mTOR dataset. This second training allows to focus on the desired chemical space, where our PI3K/mTOR QSAR models can be applied. In the present study, the molecule optimization strategy used in all cases is "Hillclimb-MLE".[2a]

Independently from this work, Iktos has developed Spaya™, [7] a data-driven CASP platform based on AI. This template-based AI retrosynthesis software focuses on prioritizing synthetic routes for individual input molecules. But for chemical libraries, the assessment of synthetic accessibility requires a high-throughput framework. To do so, Spaya API, a recently developed API running on Spaya's algorithmic engine for library scoring purposes, has been used herein to evaluate the synthetic accessibility of newly generated molecules. For a given molecule (m), the Spaya synthetic accessibility score (Spaya SAS) is derived from the scores given to synthetic routes in Spaya but handled in a high throughput manner by Spaya API. To score a molecule (and obtain its Spaya SAS score - between 0 and 1 - the higher the better), Spaya API performs a retrosynthetic analysis with an early stopping process. The early stopping mode stops the Spaya run when a timeout of one minute has elapsed or when a found route reaches a defined threshold (fixed to 0.6 by default). The identified routes are used to compute the Spaya SAS as follows: Spaya SAS(*m*) = max({ score(route(m)); route(m) given by Spaya API with early stopping }). The Spaya SAS is a composite of four scores as follows: *score(route)= f(d,p,c,a): d: number of reactions steps in the route, p: likelihood of the disconnections of the retrosynthesis route, c: convergence of the route, a: applicability domain estimation of the reaction templates used to make the disconnections.* Spaya API also returns the number of steps of the shortest route to the input molecule from a catalogue of ~50M commercially available starting materials provided by MCule (https://mcule.com).

Scheme 1



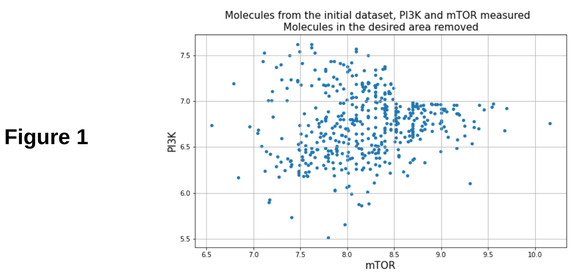


Figure 1

2 Results:

Generation 1: Synthetic accessibility computed in post-processing

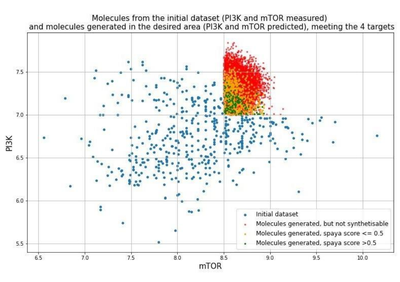


Figure 2a

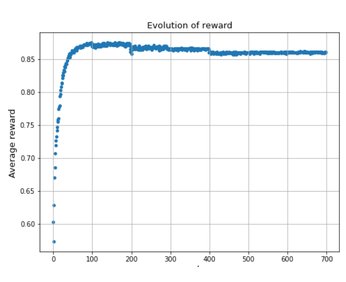


Figure 2b

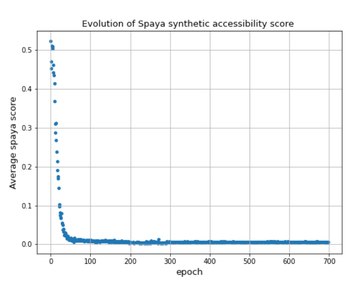


Figure 2c

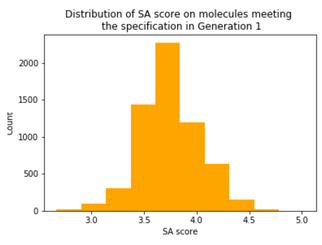


Figure 2f (above) - 2g (below)

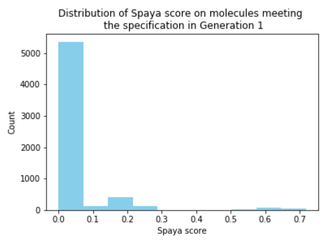


Figure 2e

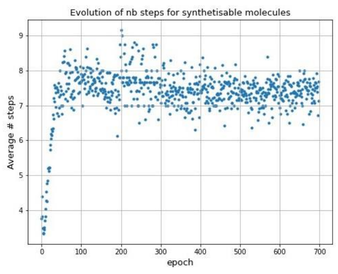
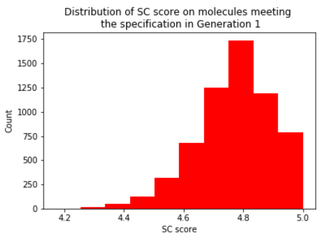


Figure 2d



This first generation was designed to optimize 4 objectives (as described in the materials and methods section) : pKi values for PI3K and mTOR, similarity and QED, without synthetic accessibility constraint. After the generation, 6358 molecules were found to meet the objectives (**Figure 2a**). As a post-processing step, these molecules were scored using Spaya SAS : a great majority (87%) of the molecules satisfying the constraints are not synthesizable (in red figure 2a), and only 2% of those have a good synthetic accessibility score (in green figure 2a). Notably, the higher the predictions of Pi3K and mTOR for the generated molecules, the poorer their synthetic scores. Over time (plotted for 700 epochs), the overall reward increases for the 4 parameters (**Figure 2b**) but the Spaya SAS collapses in 50 epochs (**Figure 2c**) which shows that optimized molecules tend to be in average harder to synthesize. The number of synthetic steps to obtain those compounds (between 7 and 8 which is high) is anticorrelated to our synthetic accessibility scoring (**Figure 2d**). Among the molecules satisfying the target product profile, very few are well scored by the Spaya SAS (**Figure 2e**). This demonstrates again how important it is to use the synthetic accessibility criterion at least to filter the generated molecules, and at best to try integrating it as a new constraint during the optimization process. An analysis of generated molecules synthetic accessibility computed by other synthetic accessibility scores such as the SA score[8] (a score between 1 and 10 - the lower the better - based on estimations of synthetic accessibility of substructures and complexity of the molecule) and SC score[9] (a score between 1 and 5 - the lower the better - inferred by a neural network trained to rank molecules according to their ease of synthesis defined thanks to a database of reactions of the literature: Reaxys) is showed in figures **2f** and **2g** respectively. The SA Score and SC Score appear to be less discriminant than Spaya SAS. Furthermore, the Spearman correlation coefficient between (1-Spaya SAS) and the SA score and SC Score was very low (0.26 and -0.12 respectively).

Generation 2: Generation under constraint of synthetic accessibility

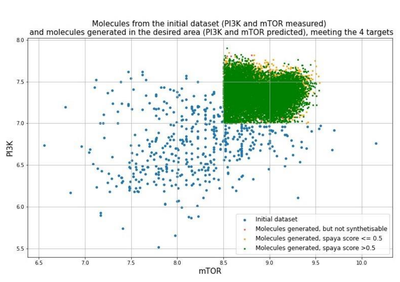


Figure 3a

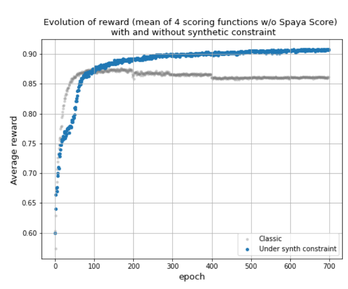


Figure 3b

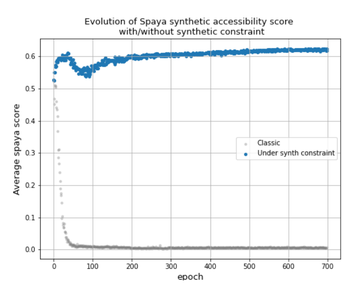


Figure 3c

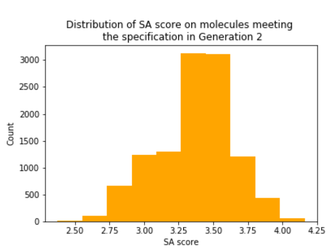


Figure 3f (above) - 3g (below)

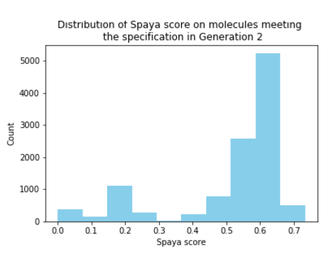


Figure 3e

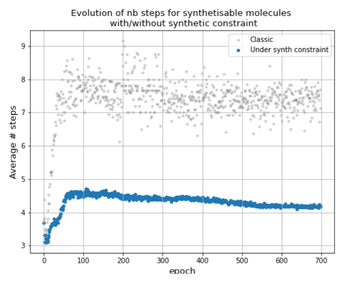
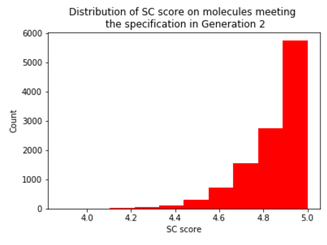


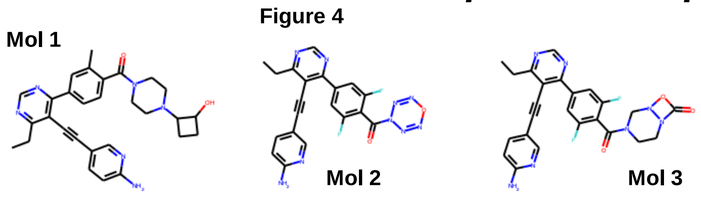
Figure 3d



This second generation, carried out under synthetic accessibility constraint (as a fifth objective), resulted in twice as many molecules satisfying the 4 other constraints : 11729 instead of 6358. But most importantly, a majority (76%) of molecules now had a good Spaya SAS (**Figure 3a**), and only 3% were found not be synthesizable by Spaya API. Finally, in this experiment, we notice that a proportion of easily synthesizable molecules are higher above the PI3K and mTOR thresholds. Those three observations highlight that adding Spaya SAS to the reward function significantly improves the latter score, driving the generator toward a chemical space fulfilling the 4 constraints with easier to make molecules. Once again, the overall reward of the algorithm increases over time (**Figure 3b**): the geometric mean of the 4 scoring functions (without Spaya SAS) are comparable, with a slight advantage for the second generation. Hence, adding a synthetic constraint did not penalize the generator on the other targets. Contrary to **Generation 1** for which the average Spaya SAS collapsed in 50 epochs to almost 0, in the **Generation 2**, carried out under synthetic accessibility constraint, the Spaya SAS remained steadily above 0.55 (**Figure 2c**). As expected, this is correlated to a lower (slightly above 4) number of steps required to synthesize these compounds (**Figure 3d**). The distribution of Spaya SAS of the molecules that meet the specification in **Generation 2** shifted clearly to the right (decrease in complexity) (**Figure 3e**). For the same molecules, we observe a slight decrease in complexity according to the SA score (**Figure 3f)**, and, unexpectedly an increase in complexity as measured by the SC score (**Figure 3g**).

Eloquent discrepancies between the Spaya SAS and other synthetic accessibility scores

Figure 4



It is important to note that the SC score scored all the considered molecules above 4, which means that they are all considered as difficult to access. The SA score seems less stringent but scores all molecules in a limited range around the mean. Hence, they appear to be ill-suited to discriminate easily synthesizable molecules from complex molecules in the present case. A selection of three molecules (**Figure 4**) obtained after **Generation 1** was made to illustrate the discrepancies found between Spaya SAS and SA score and/or SC Score. The first molecule (**Mol1**, Spaya SAS score = 0.72, SC score = 4.98 and SA score = 3.66) is well scored by the Spaya SAS, but displays a high complexity according to the SC score and SA score. Spaya was indeed able to find a very short route for **Mol1** (4 reaction steps). Conversely, the second compound (**Mol2**, Spaya SAS score = 0, SC score = 4.3 and SA score = 3.8) and **Mol3** (Spaya SAS score = 0, SC score = 4.75 and SA score = 3.6) was scored at zero by Spaya but were scored better than **Mol1** by the SA Score and SC Score, despite the presence of irrelevant motives: obviously not synthesizable and probably unstable.

3 Conclusion

While existing synthetic accessibility scores such as SA score and SC score are used to rank molecules from a library designed by human, they may not be suited to assess molecules emanating from deep generative processes where molecules can contain exotic moieties (not synthesizable - unstable - unrealistic). Conversely, the Spaya SAS seems to discriminate between easily accessible compounds and the others, since it is the result of a real retrosynthetic analysis of the generated molecules. Herein, we show that generating molecules without synthetic accessibility constraint requires a subsequent filtering, enabled by Spaya SAS, to identify synthesizable compounds. Moreover, we demonstrate that integration and maximization of Spaya SAS to the generative process solves the problem of synthetic complexity of molecules designed by generative models, and leads to synthesizable molecules without penalizing the other objectives of the MPO blueprint. From our perspective, integration of synthetic accessibility is key to take advantage of the full potential of generative models in real life, *i.e.*: obtaining synthesizable optimized molecules. To our knowledge, existing scores (*e.g.* QED, similarity, SC score and SA score) are not sufficient to achieve the level of compound quality expected by the chemists.

[1] Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., Fisher, J., Jansen, J., Duca, J., Rush, T., Zentgraf, M., Hill, J. E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebkemann, C. and Schneider, G. Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug. Discov.* **2020**, *19*, 353–364. [2] (a) Neil, D., Segler, M.H., Guasch, L., Ahmed, M., Plumbley, D., Sellwood, M., & Brown, N. Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design. *ICLR* **2018**. (b) Segler, M.H.S., Kogej, T., Tyrchan, C., Waller, M.P.: Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120-131. [3] (a) Segler, M., Preuss, M. and Waller, M. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604-610. (b) Struble, T. J, Alvarez, J. C., Brown, S. P., Chytil, M., Cisar, J., DesJarlais, R. L, Engkvist, O., Frank, S. A., Greve, D. R., Griffin, D. J., Hou, X., Johannes, J. W., Kreatsoulas, C., Lahue, B., Mathea, M., Mogk, G., Nicolaou, C. A., Palmer, A. D., Price, D. J., Robinson, R. I., Salentin, S., Xing, L., Jaakkola, T., Green, W. H., Barzilay, R., Coley, C. W. and Jensen, K. F.. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020**, *63*, 16, 8667–8682. [4] (a) Liu, P., Cheng, H., Roberts, T. and Zhao, J. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov* **2009**, *8*, 627-644. (b) Engelman, J. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat. Rev. Cancer.* **2009**, *9*, 550-562. [5] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat Chem.* **2012**, *4*, 90-98. [6] Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 3, 1096–1108. [7] Spaya is available online at: https://spaya.ai [8] Ertl, P. and Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8. [9] Coley, C. W., Rogers, L., Green, W. H. and Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58*, 252−261.