Unleashing the potential of natural products libraries by synthetically constrained AI-driven enumeration

Quentin Perron (1), Brice Hoffmann (1), Hamza Tajmouati (1), Robin Fourcade (1), Adam Skiredj (1),Yann Gaston-Mathé (1)
1) Iktos, 65 rue de Prony 75017 Paris, France

Introduction

Natural products (NPs) are known for being a valued source of biologically active molecules but they often prove to be highly difficult to derivatize by medicinal chemist.[1] Indeed, NPs may be sensitive to reaction conditions or not versatile enough to enable sufficient derivatization and chemical space exploration for designing a robust drug candidate. That hurdle prevents their use in most medicinal chemistry projects. In this context, a powerful way to circumvent the problem would be to provide a broad range of NP-derived starting points, easily accessible by organic synthesis (*e.g.* one synthetic step), poised for subsequent virtual screening or pre-mapping of NP derivatization space. This work proposes to address this challenge by a purely data-driven approach. Recent papers have reported successful implementation of data-driven AI approaches to retrosynthetic analysis.[2] We extracted synthetic rules from the USPTO database, as described by Segler *et al.*[3] We curated and selected several hundreds of NPs from the ChEBI database.[4] Then, applying data-driven retrosynthetic rules in the forward direction, we generated a library of NP-derived molecules. Those compounds are intrinsically designed to be druggable, and accessible *via* a limited number of synthetic steps, from the NP starting point and commercially available starting materials. Moreover, our method allows to assess the synthetic potential of NPs and to generate the accessible chemical space for a chosen scaffold. To our knowledge, such data-driven *in silico* synthetic approach is unprecedented.[5] Finally, its application to overcome the synthetic challenge of NP-related compounds reported in this work is of particular interest for the medicinal chemists community.

1 Natural products starting library

To assess the proposed approach, a limited collection of natural products (NPs) was curated from the Chemical Entities of Biological Interest (ChEBI) database.[4] A series of Lipinski-related criteria including MW, number of heavy atoms, number of stereocenters, number of saturated rings, and overall database diversity guided the selection of relevant NPs. From here, 260 natural compounds were selected to serve as starting points for subsequent data-driven enumeration and "library growing". One of the key features of NPs is their three dimensional properties. Hence, a principal moment of inertia (PMI) plot was chosen to measure and compare the shape of compounds throughout the study. The seed population of molecules covers a significant range of the PMI space. Some examples of NPs selected in our starting library are shown here (**Figure 1**) to give some idea of their structural nature and diversity.

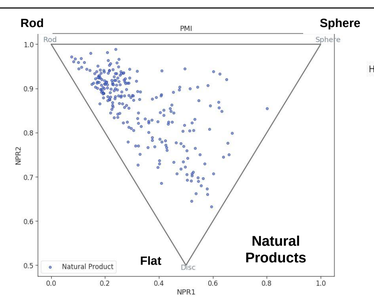


Figure 1

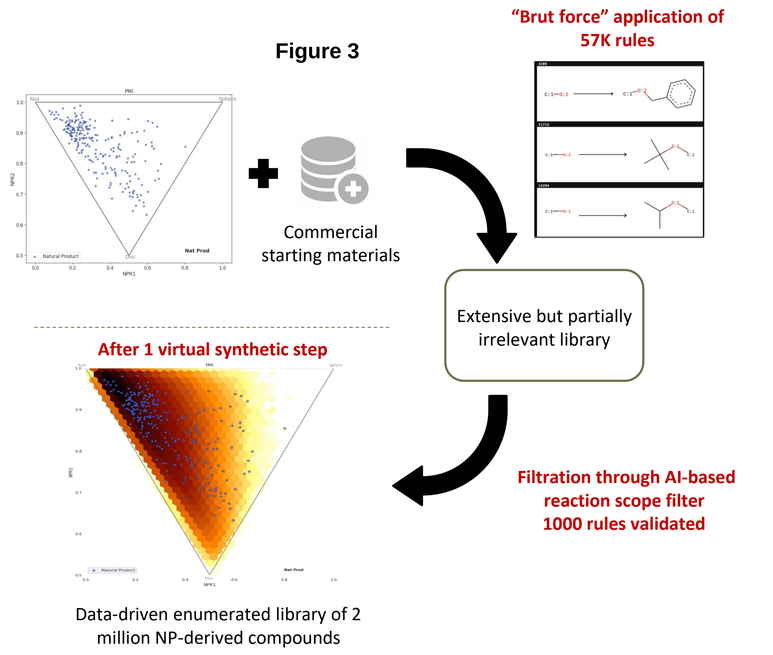
2 Data driven rules extraction & enumeration method

In addition to the NPs, two other “dictionaries” were required to initiate the enumeration work. First, commercial starting materials relevant to synthetic chemists working in medicinal chemistry. For this purpose, a repository of 1 million building blocks was assembled from various sources and curated based on their properties as number of rotatable bond, HBA, HBD and also the MW since the total weight of the final product should not be higher than 500. Second, an extensive and reliable panel of synthetic reaction rules. To extract the chemical transformations that we intended to apply to our library of NPs, we developed our own pipeline of reaction cleaning, starting from the USPTO database which allows access to a database of 1.6 million reactions extracted from patent literature from 1976-2016.[6] Since the database was extracted using machine learning (ML), an important work of data cleaning was necessary (**Figure 2**). Starting from raw reactions, the first task is to perform atom-mapping[7], *i.e.* labelling all the atoms of reactants and products in order to identify the ones which are involved into the reaction. We remove all non-reacting species by filtering the reactions against a list of known solvents, catalysts and commercially available reagents. Then, we applied the Indigo open source software to atom map the reaction, and the Coley procedure[8] to extract the reactive center, *i.e.* atoms & bonds of reactants and products which are modified during the course of the reaction. A rule is defined as the reactive centers of reactants and products. We chose to extract rules of rank 2, meaning that we not only extract the modified atoms but also their immediate neighbours.

Figure 2

With NPs, commercial compounds and reactivity rules in hands, the enumeration phase was carried out to generate a vast *in silico* library constituted by all relevant products emanating from reactions (one synthetic step) between commercial starting materials and NPs, as explained below. Notably, the synthetic relevance of each generated product is ensured through the use of a reaction scope filter identical to the one proposed by Segler *et al*. (see **Figure 3**, right).[3] First, a “brute force” application of the clean, data driven synthetic rules to all the NPs of our initial collection led to exhaustive but partially irrelevant library of *in silico* compounds. Indeed, the purely data-driven enumeration phase does not take into account the reaction contexts and incompatibilities between functions. Therefore, in a second phase, that vast reaction bank was filtered by the scope validation tool, built to assess the reactions feasibility. Basically, the scope filter can identify and remove reactions which are not feasible in a specific context. This is the ML-based step of the approach, allowing contextual assessment of generated products. Finally, 2 million NP-derived *in silico* compounds, grown from 260 NPs, were generated and this library is characterized in the next section.

Figure 3



3 Metrics of the enumerated library

Figure 4a

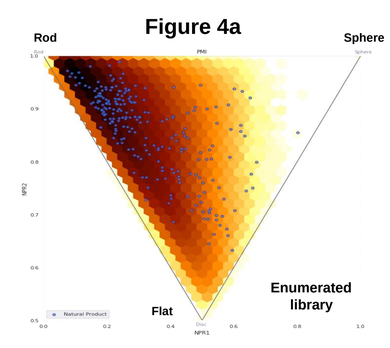


Figure 4b

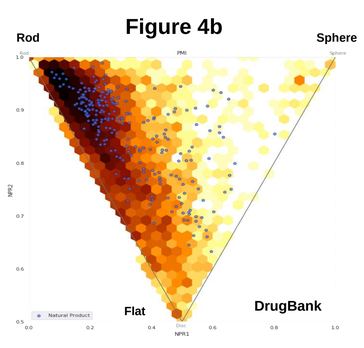


Figure 4c

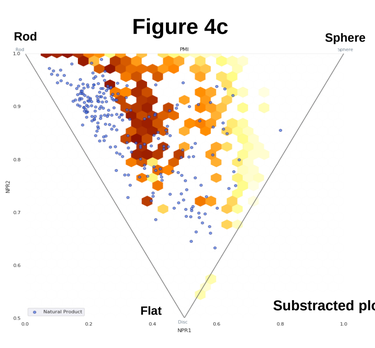


Figure 5a

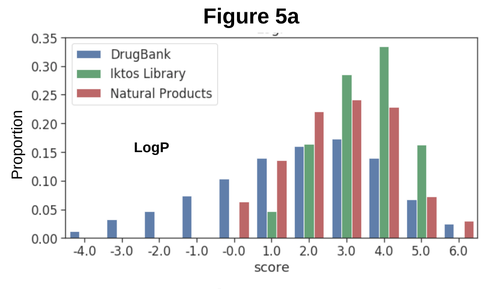


Figure 5b

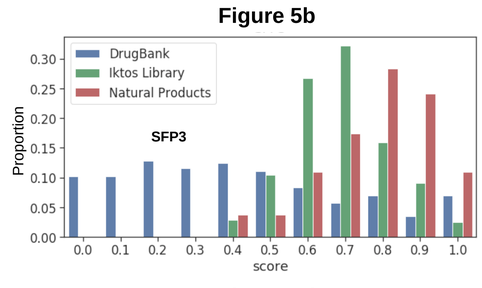


Figure 5c

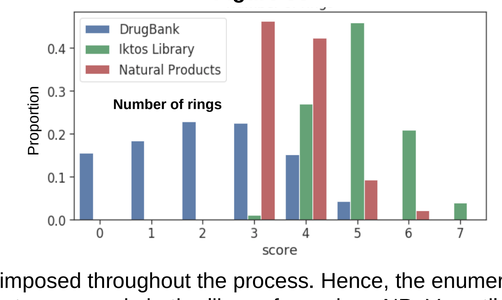
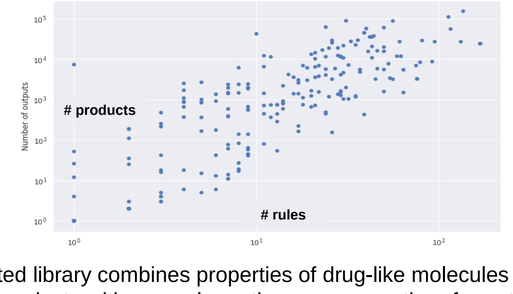


Figure 5d

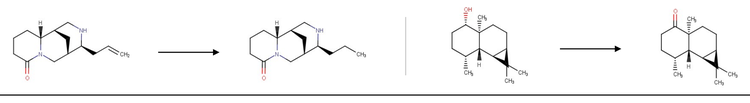


As previously explained, the 2 million NP-derived library results from 260 NPs reacting with commercial compounds in one synthetic step. The resulting chemical space enrichment and expansion is shown on the three PMI plots provided above. From left to right, our *in silico* library (**Figure 4a**) populates most of the chemical space. When compared to the PMI plot of DrugBank (**Figure 4b**), the similarity of the profiles illustrates the power of our method to generate drug like molecules. Moreover, by subtraction of DrugBank to our library, the resulting PMI plot (**Figure 4c**) shows that we are able to achieve a significantly higher population in the central and top-right areas of the plot which represent 3D molecular shapes, as opposed to flat and linear ones. A focus on the molecular features of the library such as LogP, fraction of sp3 carbons and the number of rings is presented in **Figure 5** (right). Interestingly, the Iktos library (in green) is somewhat similar to the NP starting points on those three parameters. NP-seeded library generation therefore allows to transfer NP properties to the library, while fostering differentiation with DrugBank. It is important to note that conformity to Lipinski rule of 5 was imposed throughout the process. Hence, the enumerated library combines properties of drug-like molecules and NPs. Moreover, there is a correlation between the number of applicable rules and the number of *in silico* products effectively obtained as output compounds in the library for a given NP. Versatile products with several reactive groups are therefore at the origin of a large number of molecules (**Figure 5d**, bottom-right plot). Notably, the 63 most promising NPs are at the origin of 90% of the final library population (top right of the plot of **Figure 5d**). NPs with less output molecules but generated through the application of numerous rules with several output vectors are also interesting. On the contrary, NPs with a low number of applicable templates but leading to a large number of products are not necessarily valuable since the diversity of output will be limited.

4 Selected samples of in silico NP-derived compounds

As previously stated, a significant number of NP compounds are promising (sometimes albeit being at the origin of a smaller number of enumerated molecules). They possess several output vectors, *i.e.* reactive functional groups which are successfully involved in the enumeration process (**Figure 6a**). Hence, the library emanating from them is more diverse because it is based on various reaction rules and it results from growing the starting NP at distinct positions (*i.e.* chemo- and regio-diversity). Yet, a vast number of valuable analogs with very different structural and physico-chemical properties are generated with less versatile NPs (*e.g.* only one functional group involved in coupling reactions, **Figure 6b**). Moreover, very simple transformations such as oxidations or hydrogenation are also generated in the process (**Figure 7** below). Removal or creation of a new functional group drastically changes the manner the product could bind to a target. Those addition of functional groups, albeit simple, can pave the way for generating numerous and highly functionalized candidates.

Figure 7



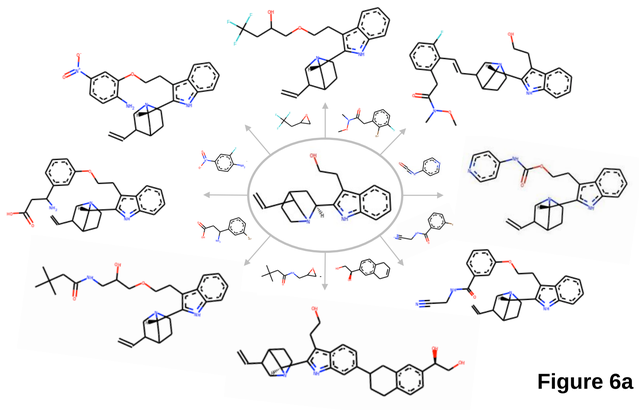


Figure 6a

Figure 6b

5 Conclusion

Starting from 260 NPs, an expanded library of 2 million NP-derived *in silico* compounds was generated using a fully data driven approach of synthetic enumeration in the forward direction. The library expansion corresponds to NPs reacting with commercial compounds in one synthetic step. The resulting chemical space enrichment is significant, especially towards molecules with 3D shapes. Moreover, the approach permits to combine properties of NPs and drug like molecules providing a large number of readily accessible compounds for subsequent virtual screening. In the meantime, the present work allows to make the link between structure, number of applicable rules/templates and number of *in silico* output products. Hence, the training of a predictor of synthetic vectors is ongoing and will be of high value to assess the derivatization potential (critical to carry-out SAR and design robust drug candidates) of chosen scaffolds or advanced synthetic intermediates.

1) Newman D.J. and Cragg G. M. NPs as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **2016**, *79*, 629-661. 2) a) Coley C.W., Rogers L., Green W.H. and Jensen K.F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent Sci.* **2017**, *3*, 1237-1245. b) Coley C.W., Green W.H. and Jensen K.F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289. c) Plehiers P.P., Marin G.B., Stevens C.V. and Van Geem K.M. Automated reaction database and reaction network analysis: extraction of reaction templates using cheminformatics. *J Cheminformatics* **2018**, 10:11. 3) Segler M.H., Preuss M. and Waller M.P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604-610. 4) Hastings J., Owen G., Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P and Steinbeck C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. **2016**, *44*, D1214–D1219. 5) a) NP libraries are numerous but rarely derivatized into larger libraries : Barnes E.C., Kumar R. and Davis R.A. The use of isolated NPs as scaffolds for the generation of chemically diverse screening libraries for drug discovery. *Nat Prod Rep* **2016**, *33*, 372-381. b) No universal tool allows to do it, *in silico*, under synthetic constraint and chemists are bound to limited, proprietary NP-like or NP-inspired libraries. An attempt to advance this topic with a systematic approach being : Yu M. J. NPt-like virtual libraries: recursive atom-based enumeration. *J Chem Inf Model.* **2011**, *51*, 541-557. 6) PhD work of Daniel M. Lowe : Lowe, D. M. Extraction of chemical structures and reactions from the literature, **2012**, Doctoral thesis, University of Cambridge, DOI : 10.17863/CAM.16293. 7) For more information about the indigo software: https://lifescience.opensource.epam.com/indigo/ 8) Coley C. W., Rogers L., Green W. H., Jensen K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity *ACS Cent. Sci.* **2017**, *3*, 1237-1245.