

Machine learning applied to natural products

Lessons from nature inspiring the design of new drugs

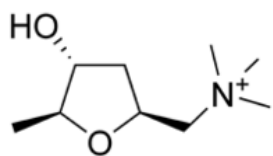
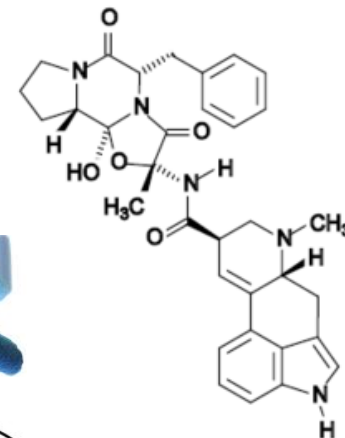
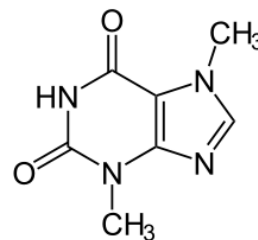
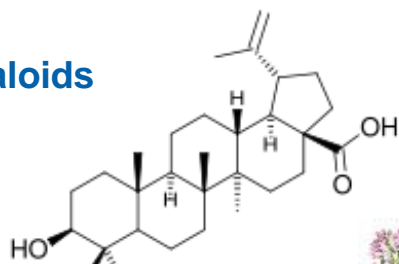
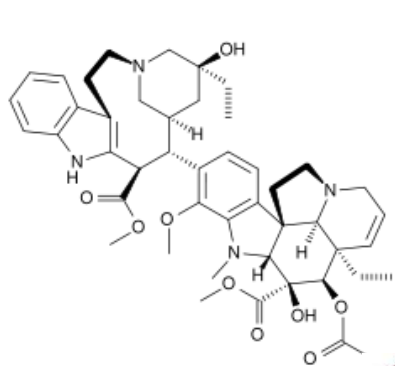
Peter Ertl

Novartis Institutes for BioMedical Research
Basel, Switzerland



Natural products 101

alkaloids



fungi

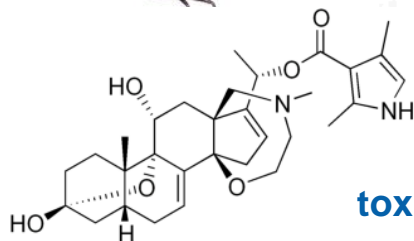
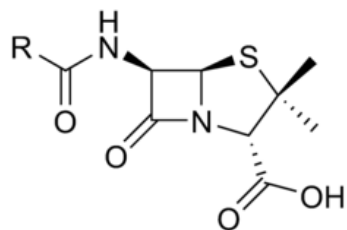
plants

bacteria

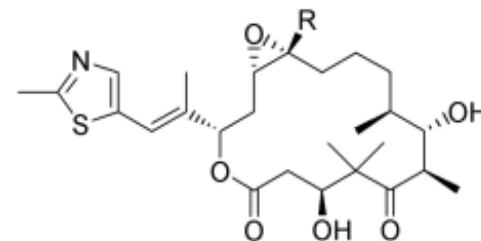
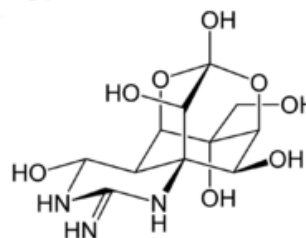


(marine) animals

antibiotics

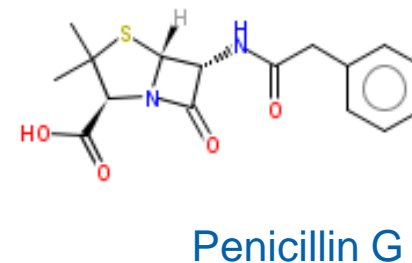
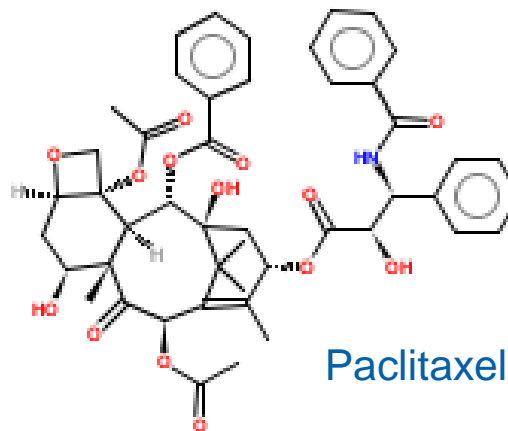
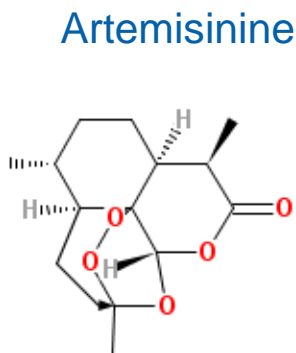
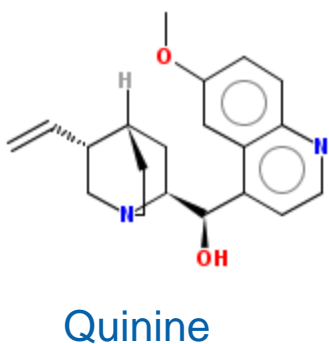


toxins



Natural products as drugs

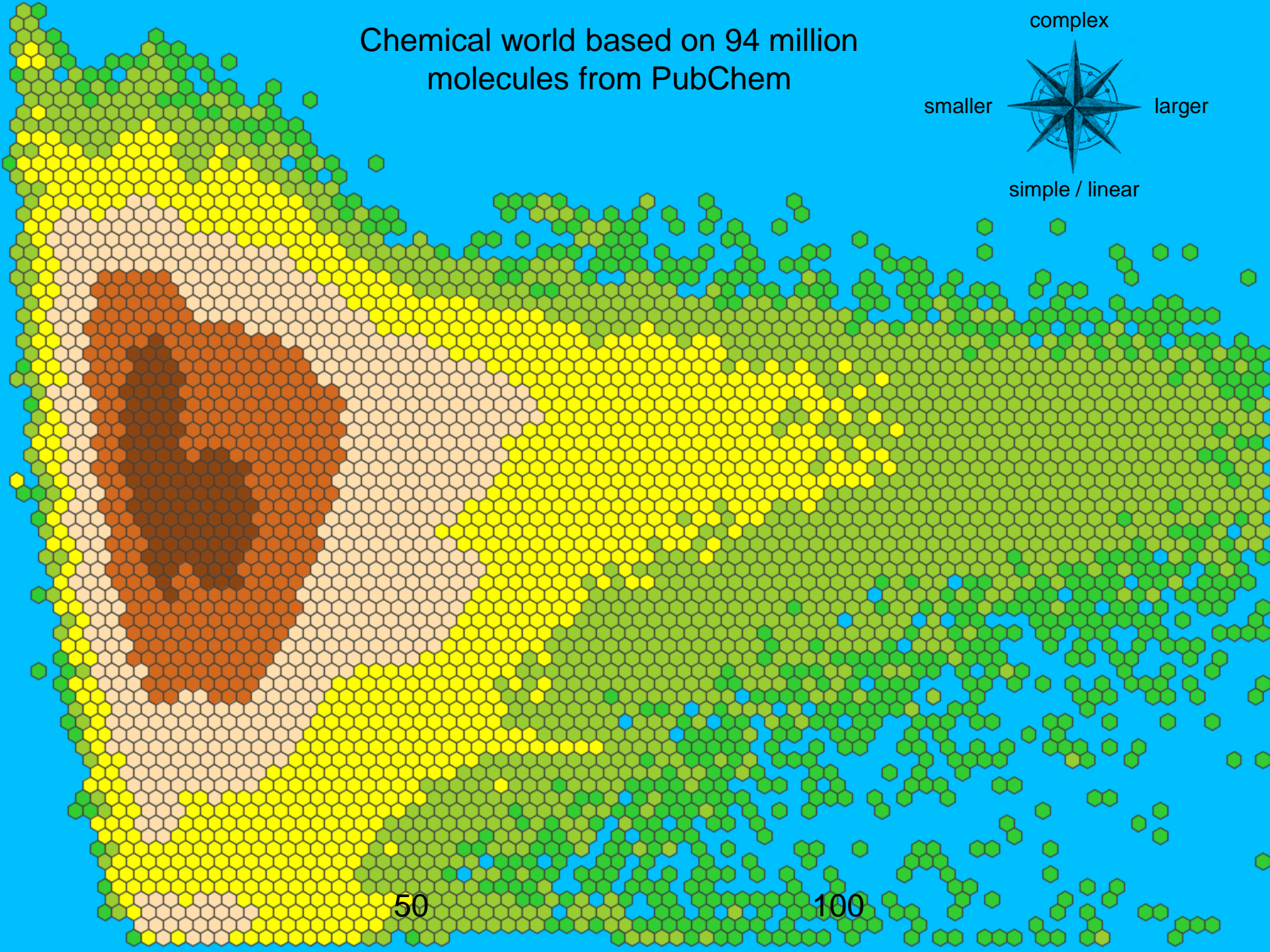
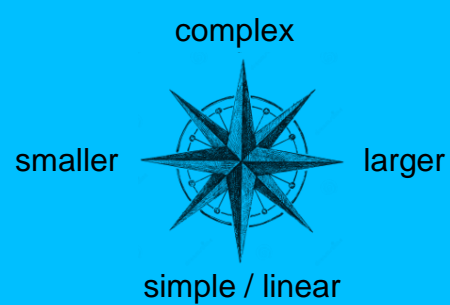
- natural products have been optimized in a very long natural selection process for optimal interaction with biological macromolecules
- NPs are therefore an excellent source of substructures which may be used as a basis in the design of new bioactive compounds
- large portion of drugs on the market are NPs or their derivatives and many other NPs are under development as new drugs



Challenges

- the structures of NPs are generally quite complex (many fused rings, several stereocenters), require complex separation
- resupply from natural sources is in many cases difficult
- access and benefit sharing is required with the countries of origin (Nagoya protocol)

Chemical world based on 94 million molecules from PubChem



Chemical world based on 94 million molecules from PubChem

natural products

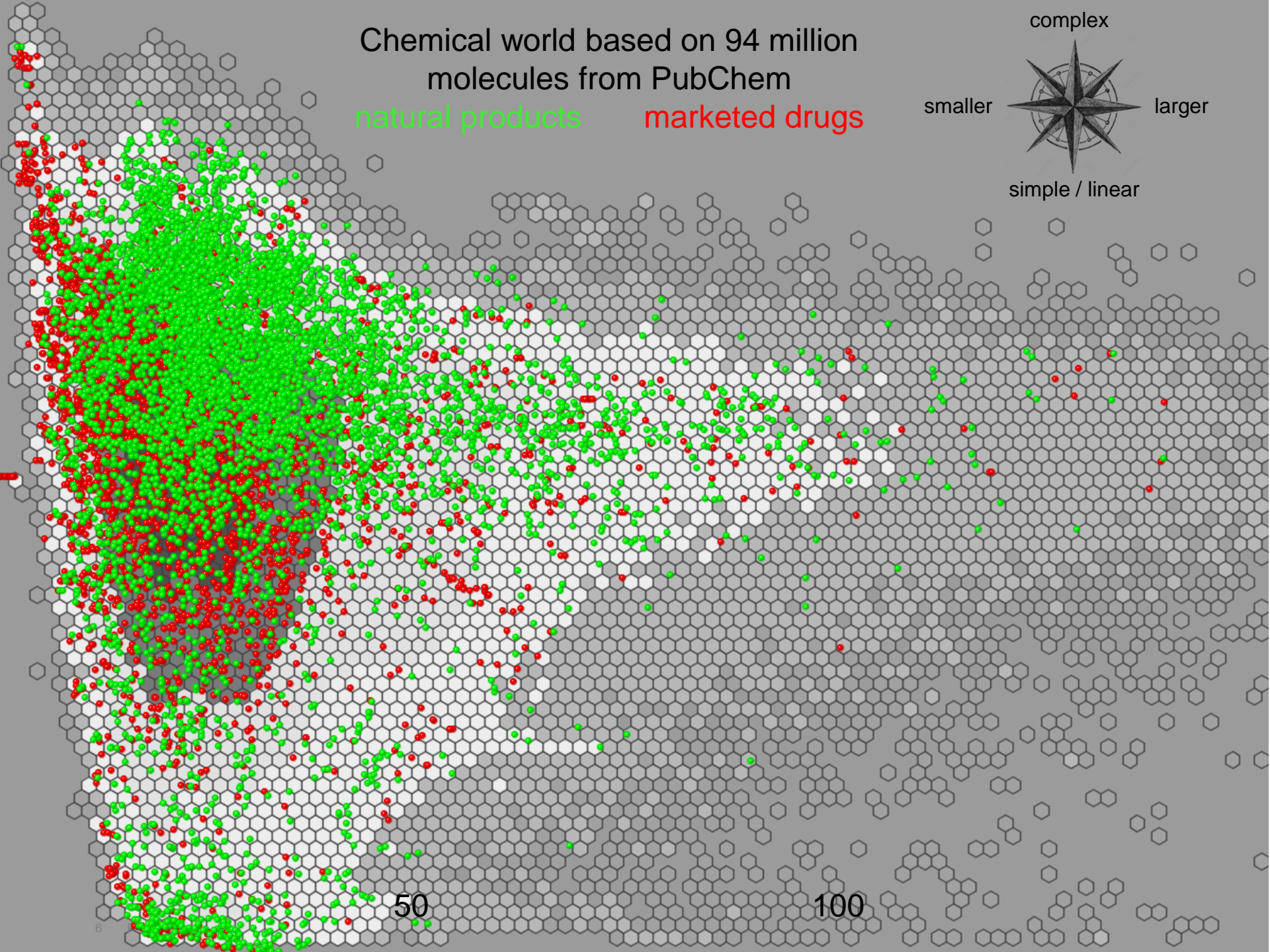
marketed drugs

smaller



simple / linear

larger

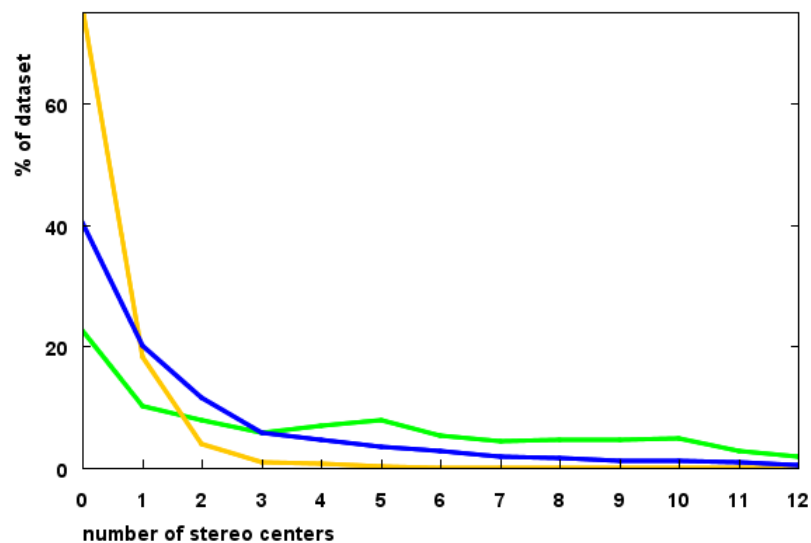
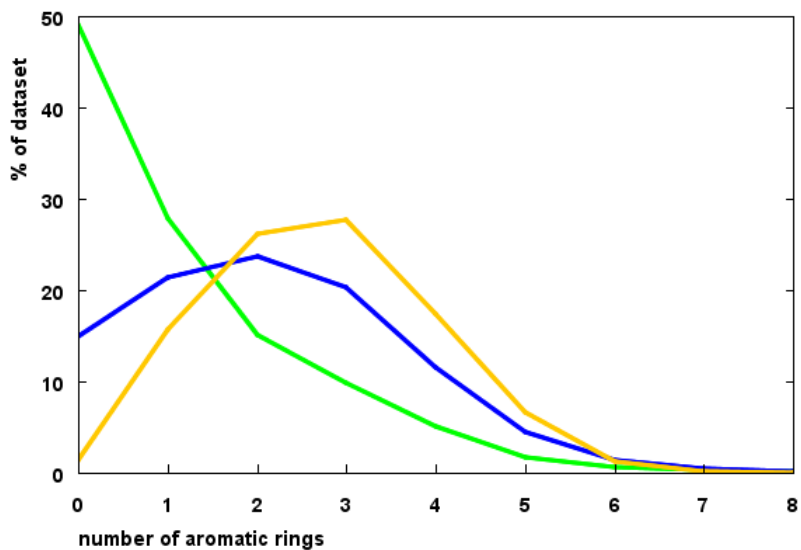
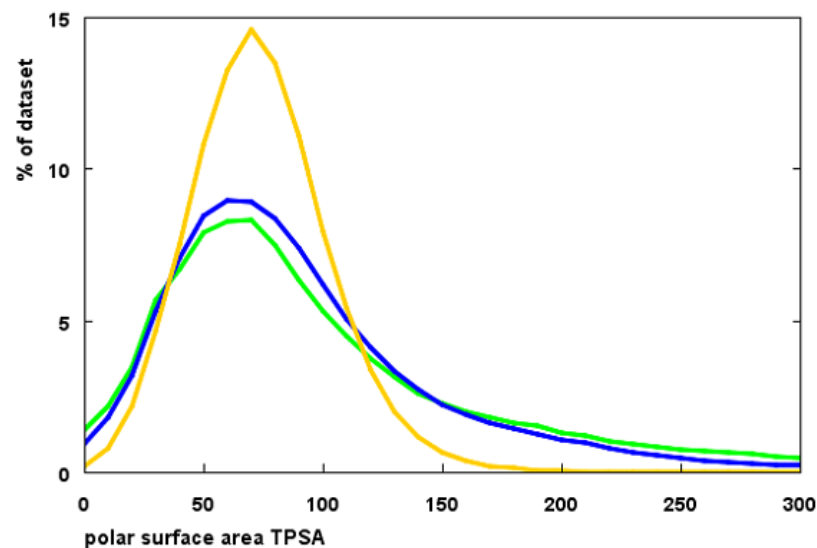
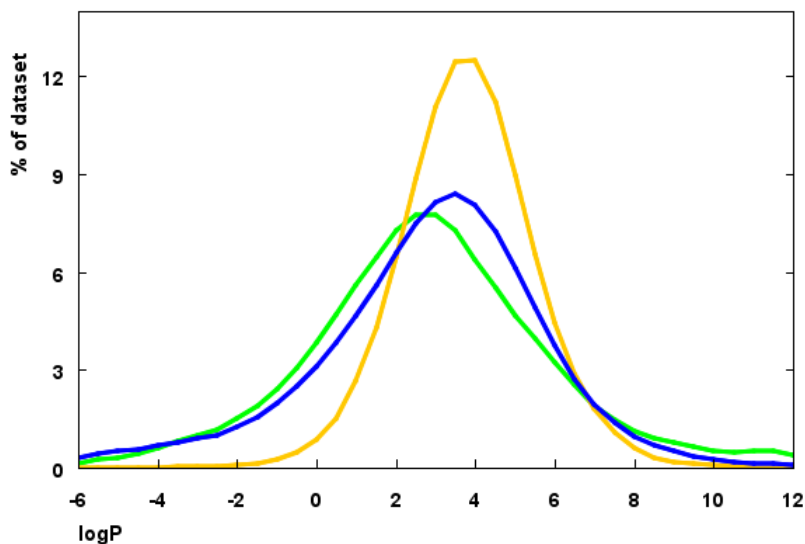


50

100

Simple properties

Natural Products (130k)
Bioactive molecules (120k)
Synthetic molecules (150k)



What makes natural products the natural products?

- physicochemical properties (logP, TPSA) of NPs do not differ significantly from those of synthetic molecules

NPs differ from synthetic molecules in some simple structural features, they have:

- less aromatic rings
- more stereo centers
- less nitrogens, more oxygens
- NPs are “more complex”

More important is to understand detailed **structural differences** between NPs and synthetic molecules – differences in **functional groups, scaffolds and substituents**. The NPs occupy **different area of structural space** than synthetic molecules.

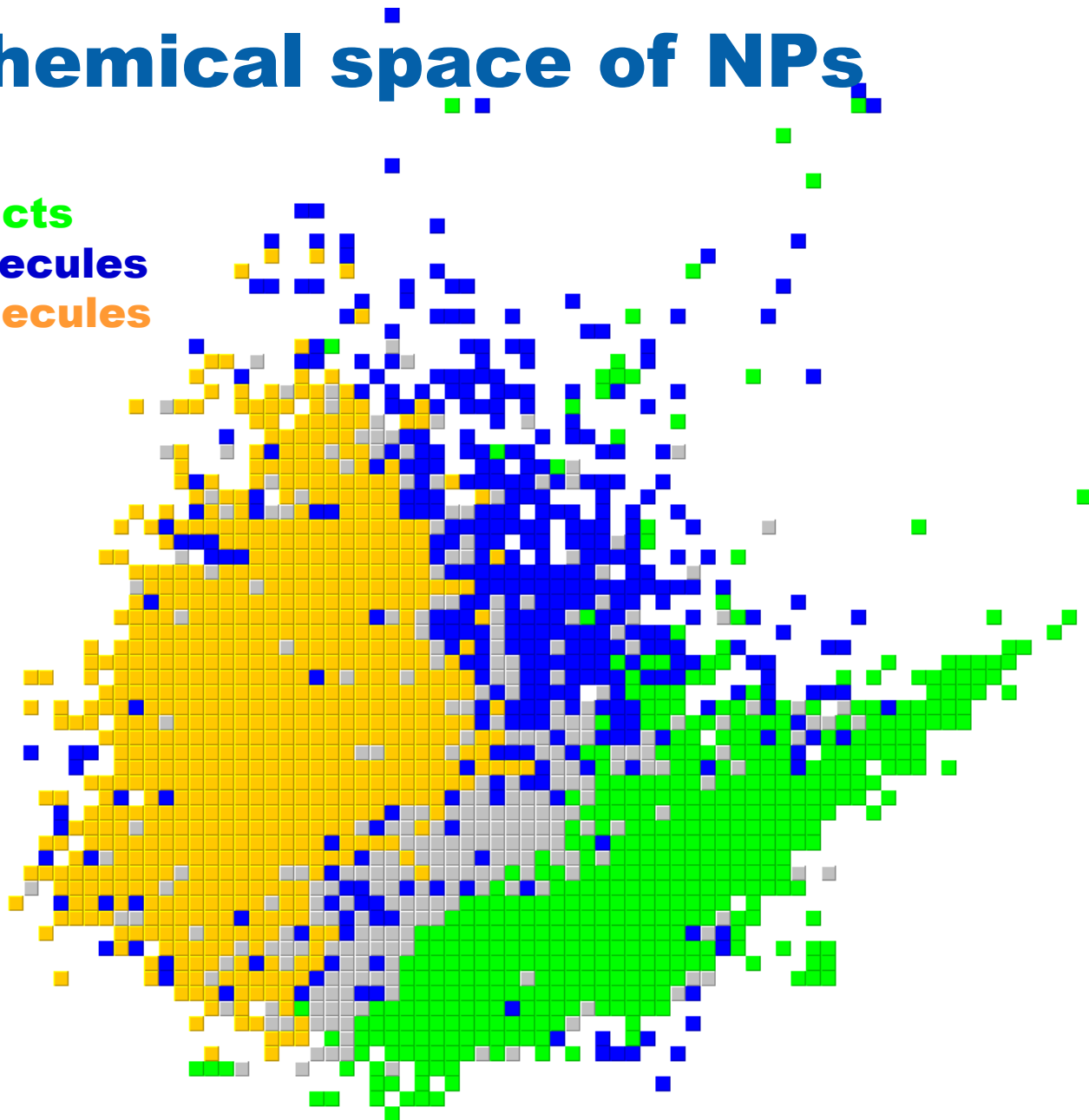
The chemical space of NPs

Natural Products

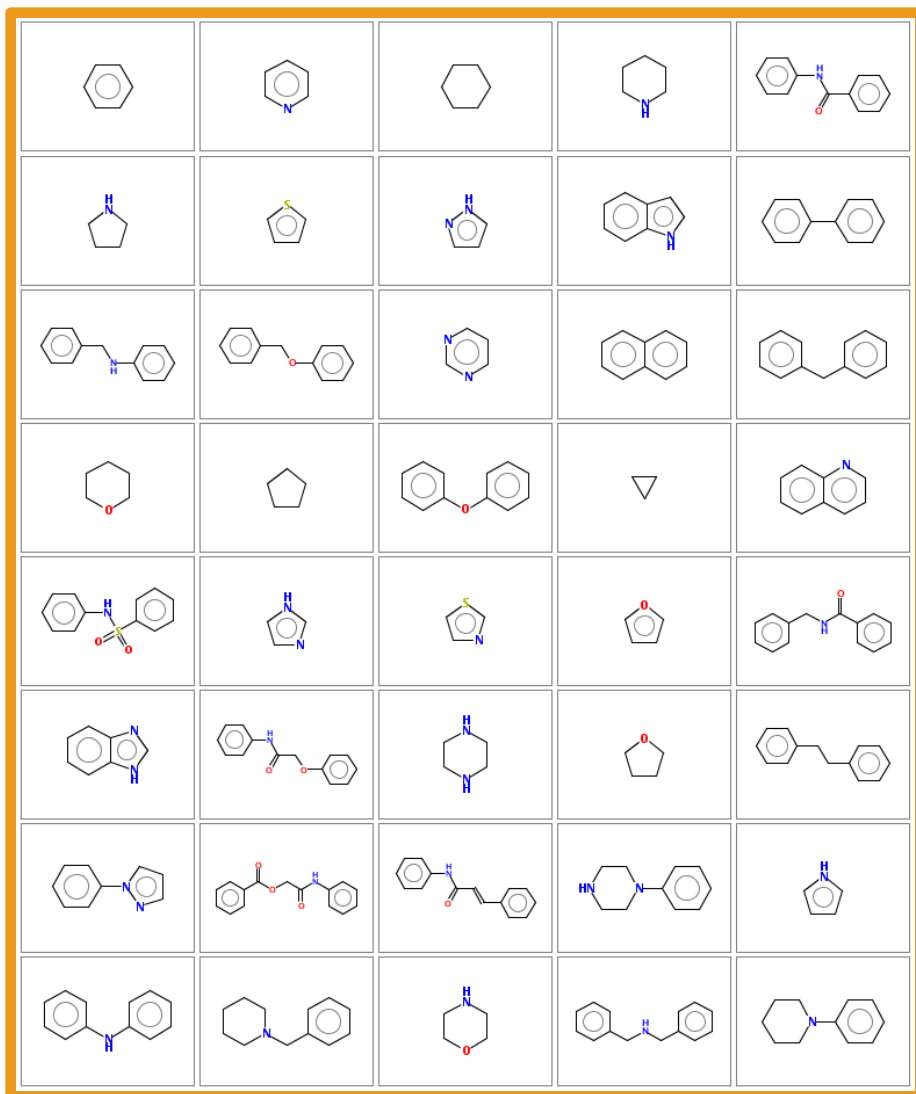
Bioactive molecules

Synthetic molecules

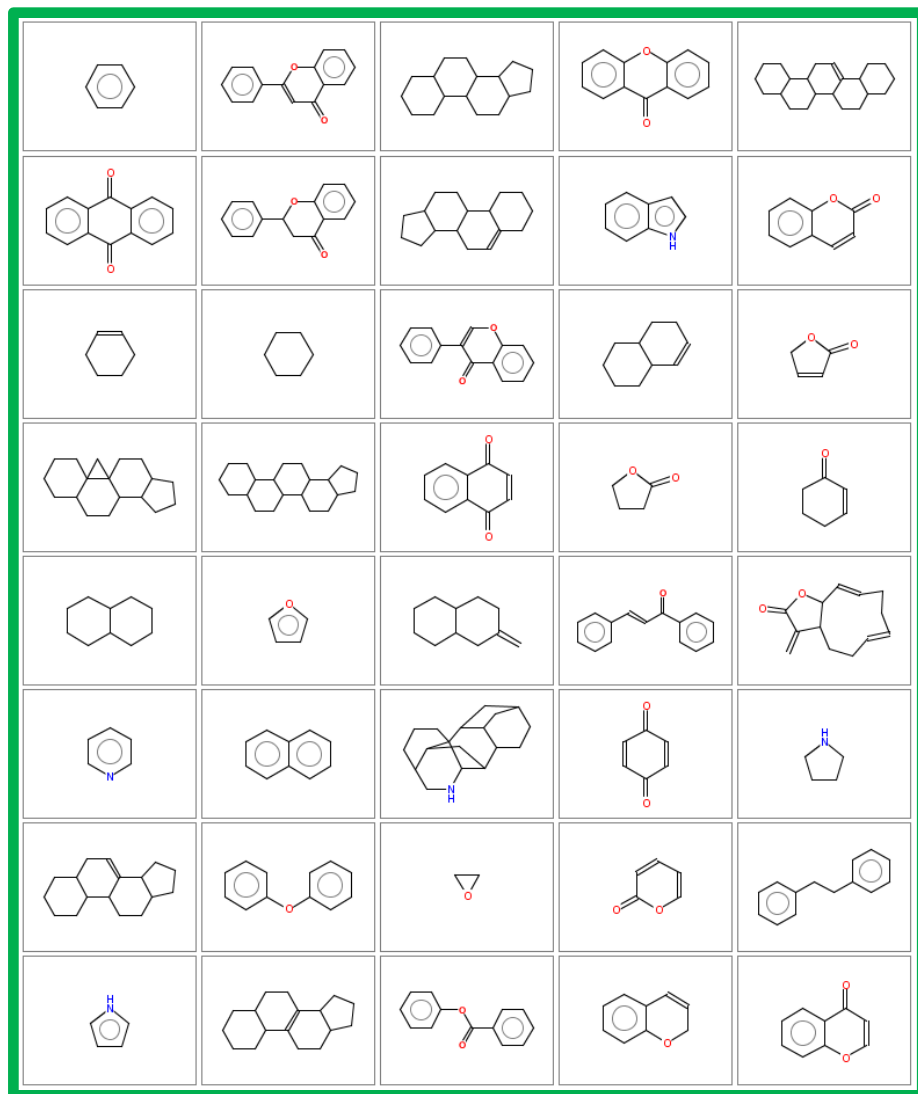
Mixed



NP scaffolds



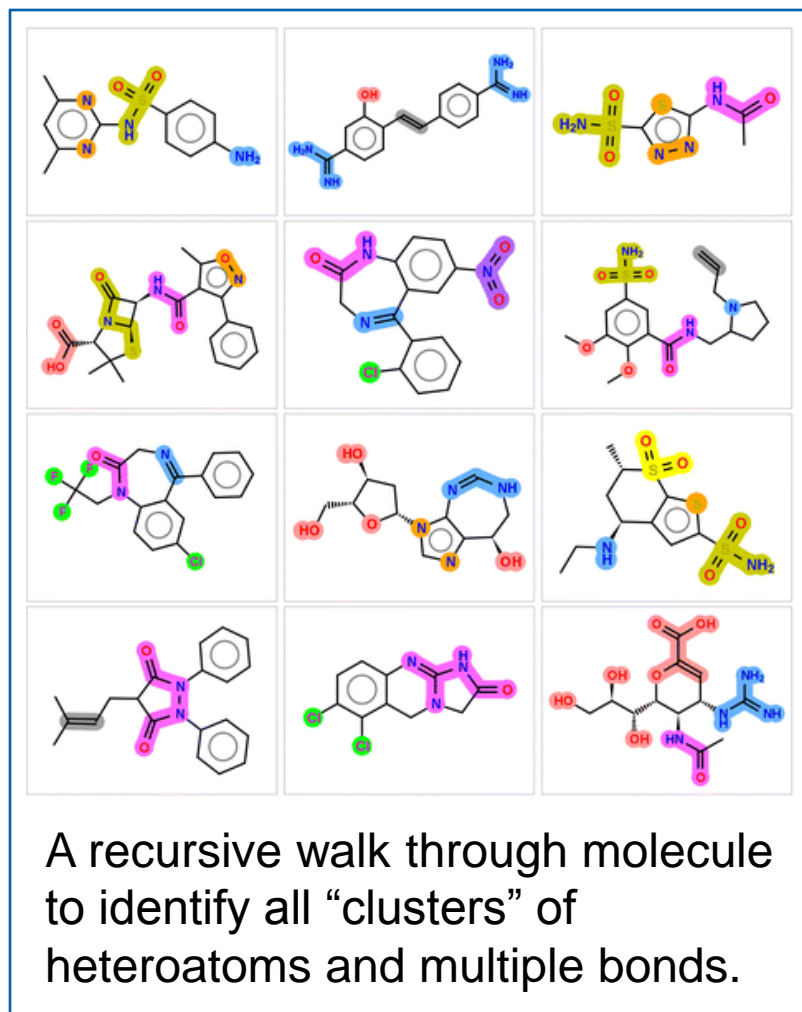
the most common scaffolds from synthetic molecules



the most common scaffolds from natural products

Functional groups

P. Ertl, An algorithm to identify functional groups in organic molecules
J. Cheminformatics 9:36 (2017)



Functional groups in ~260,000 structures of NPs from DNP252.

3051 FGs identified




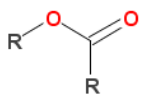

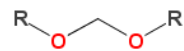

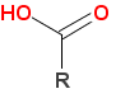


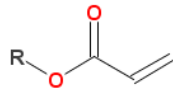

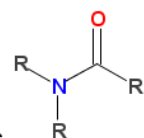
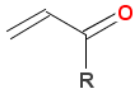
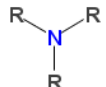

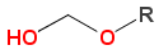
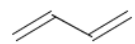
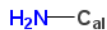
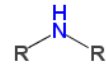

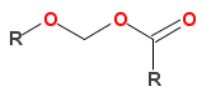

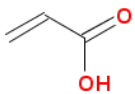

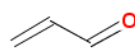
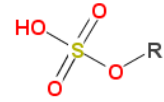
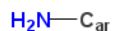
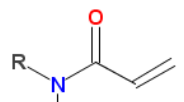

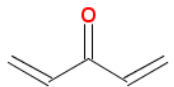

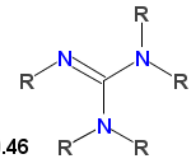
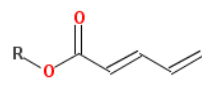

few common FGs and large number of rare FGs

11 in > 10% of molecules

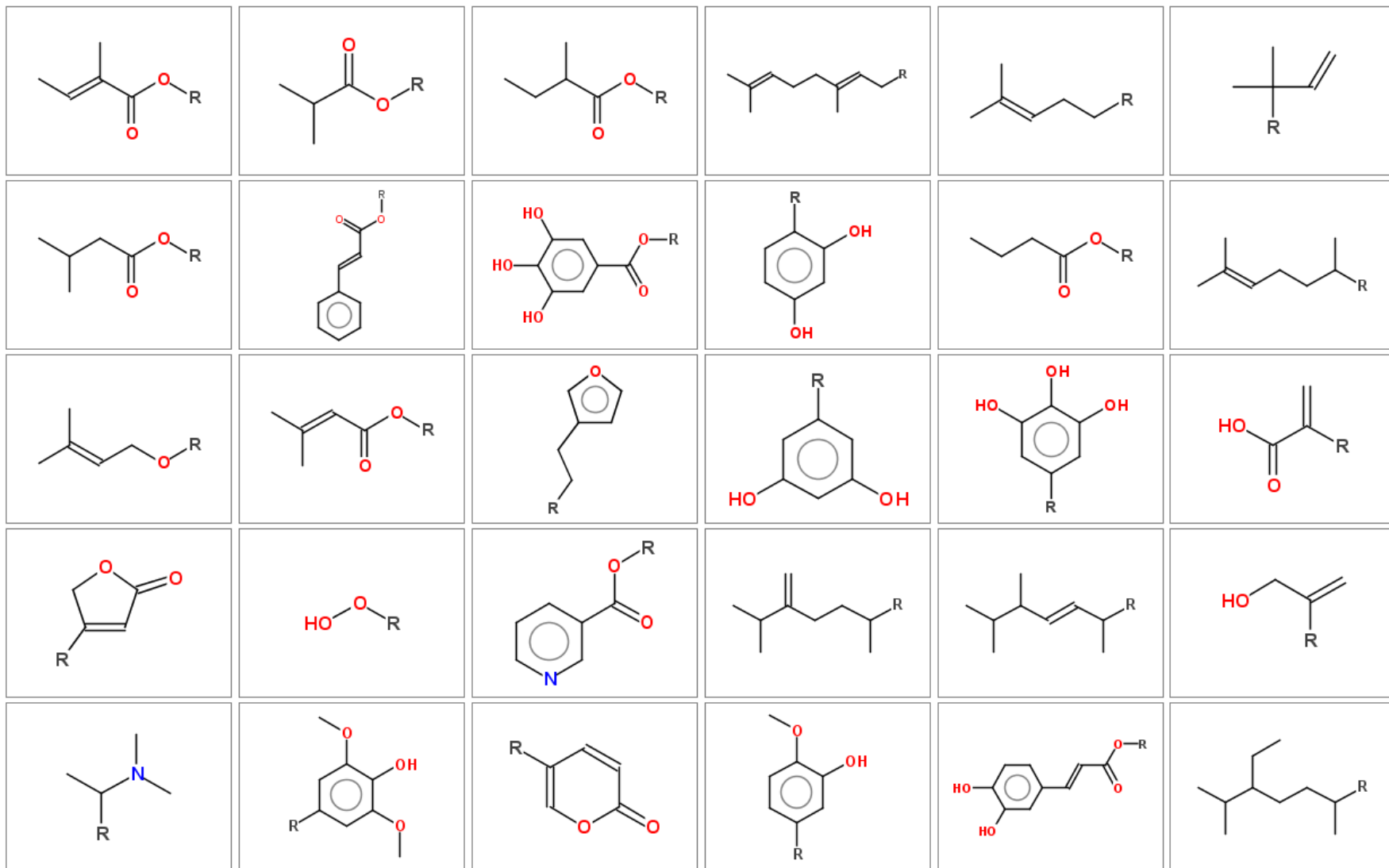
25 in > 1% of molecules

37% of FGs are singletons

The most common FGs in NPs

 60.58	 38.05	 29.57	 25.01	 24.10	 17.93	 14.43
 12.09	 11.81	 11.37	 10.28	 6.73	 6.16	 5.81
 4.62	 3.39	 3.16	 2.93	 2.79	 2.21	 2.16
 1.98	 1.69	 1.57	 1.29	 0.84	 0.76	 0.66
 0.58	 0.54	 0.53	 0.48	 0.46	 0.46	 0.44

Substituents typical for NPs



Natural product-likeness

P. Ertl, A. Schuffenhauer, S. Roggo, J. Chem. Inf. Model. 48, 68 (2008)

NP-likeness is a measure how the molecule is similar to the chemical space occupied by NPs. The method, developed at Novartis, is based on fragment contributions and naive Bayesian statistics.

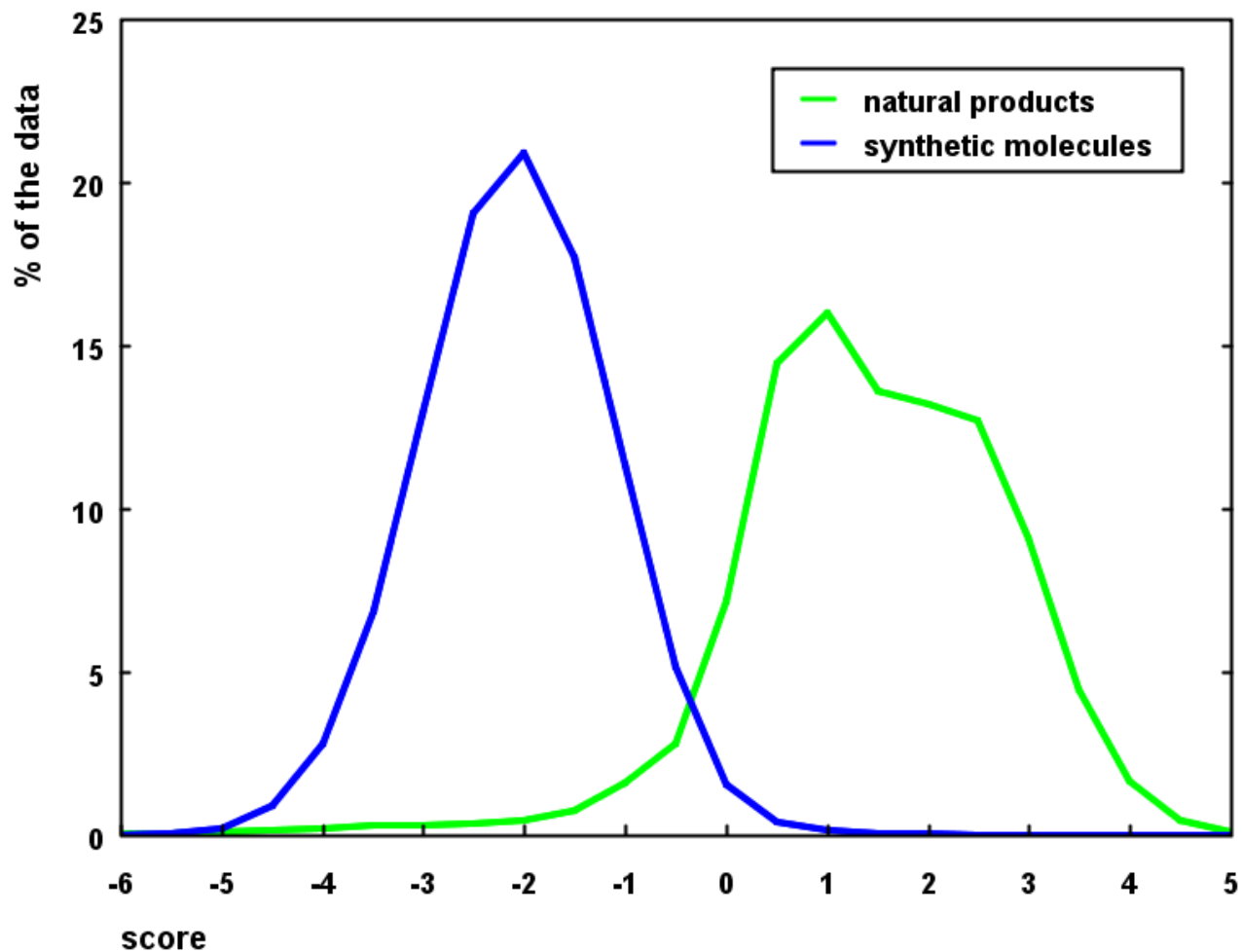
1. molecules in 2 large training sets (NPs and “average” synthetic molecules) are fragmented into atom centered fragments with up to 2 neighbor levels
2. for each fragment a fragment NP-likeness contribution is calculated using the following formula

$$f_i = \log (nact_i / ninact_i * ninact_{total} / nact_{total})$$

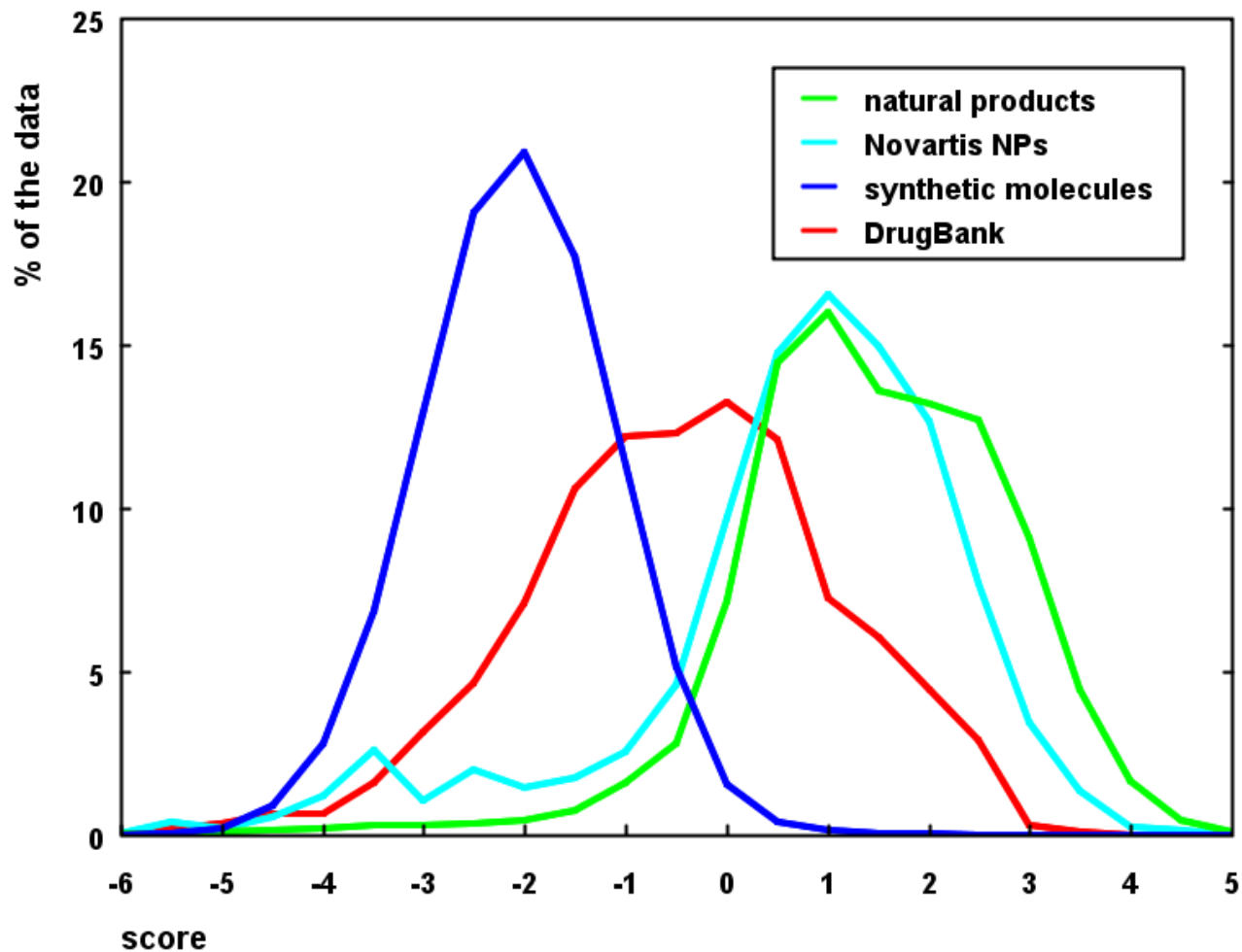
this creates a large database of fragments with their contributions

3. NP-likeness of a new molecule is calculated simply as a sum of its fragment contributions, normalized to the size of the molecule

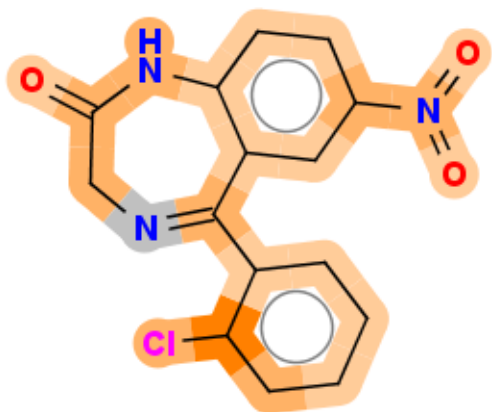
NP-likeness score



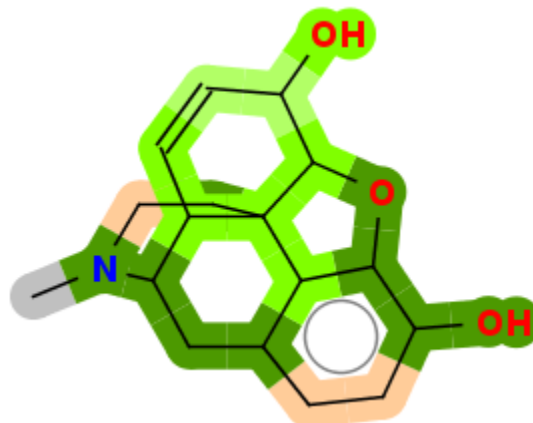
NP-likeness score



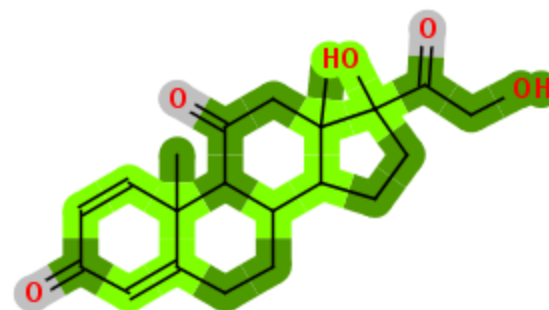
Visualization of NP-likeness



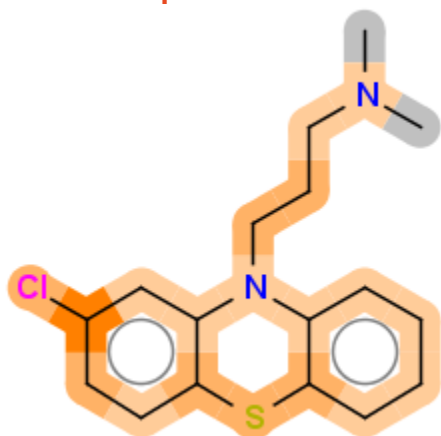
clonazepam



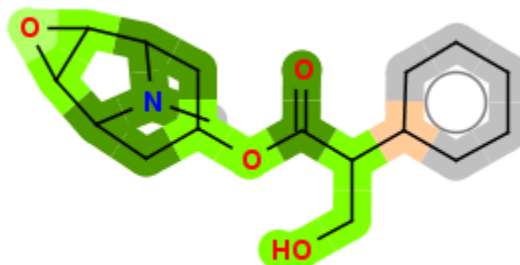
morphine



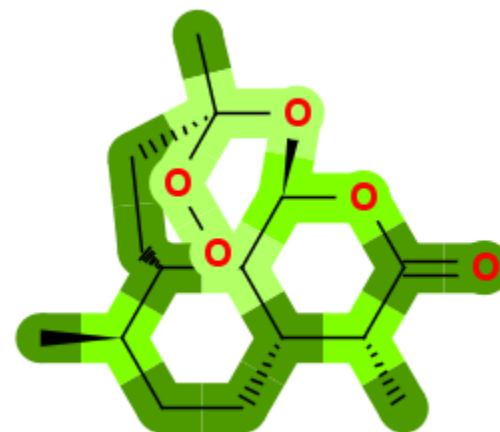
prednisone



chlorpromazine



scopolamine



artemisinin

Application of NP-likeness

virtual screening

selection of molecules for screening and enhancement of the company molecular collection

library design

selection of NP-like scaffolds and substituents for synthesis of NP-like combinatorial libraries; a balance between NP-likeness and complexity need to be considered

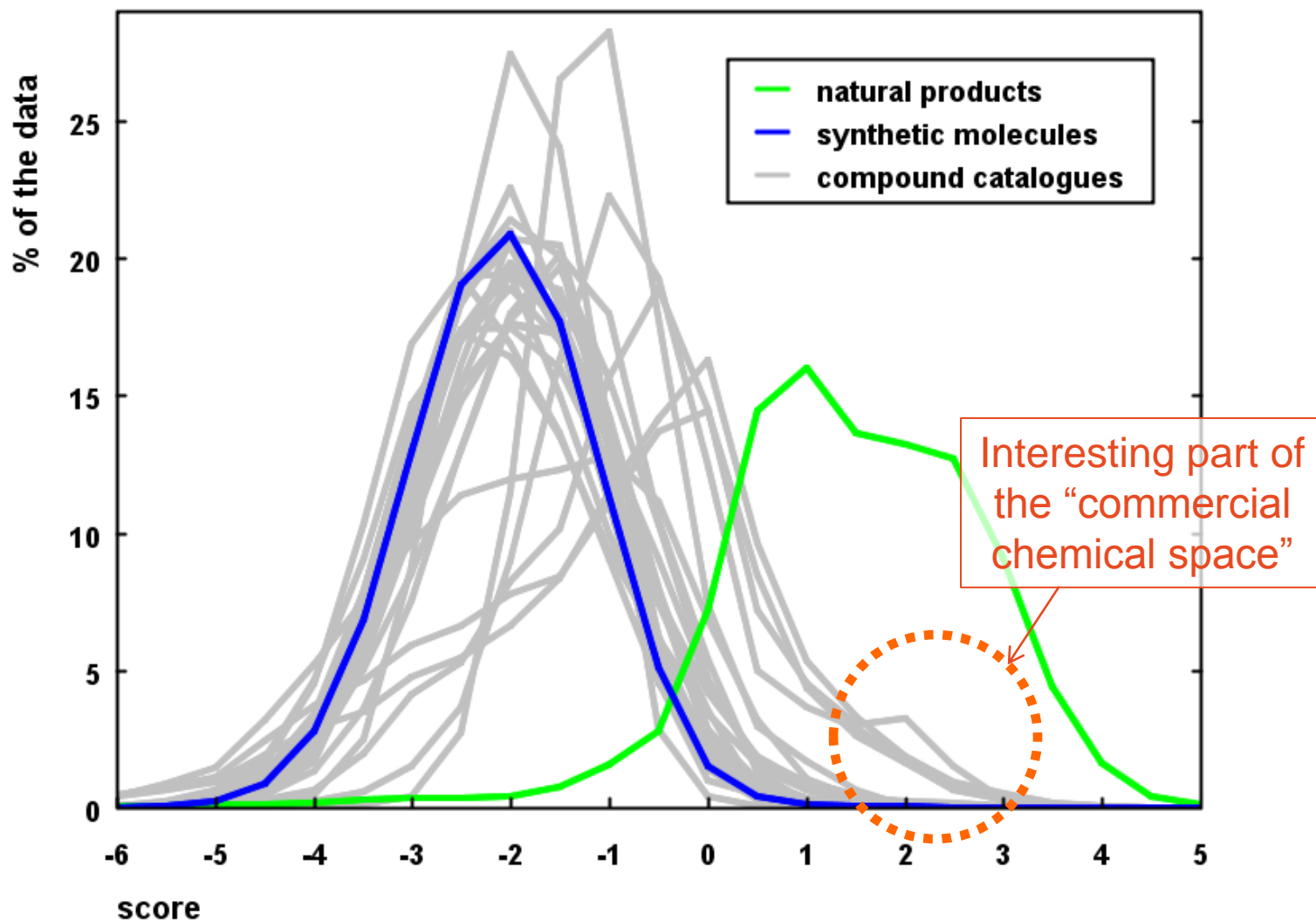
selection of fragments for fragment-based screening

identification of fragments with high NP-likeness for fragment-based screening

de novo molecule design

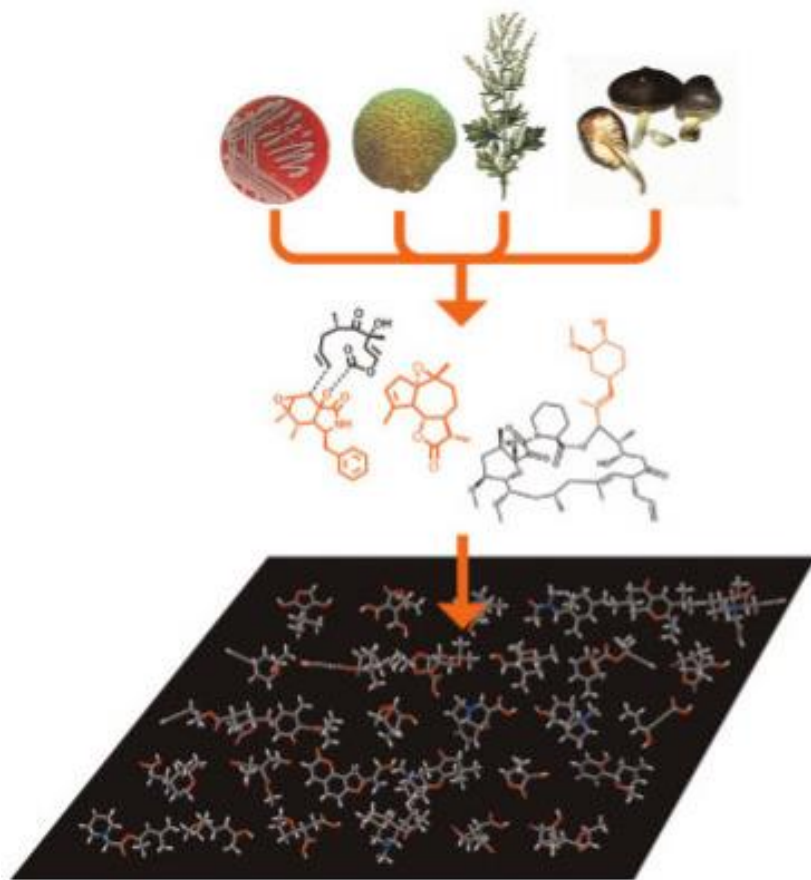
automatic design of molecules, optimizing at the same time multiple properties, including the NP-likeness

NP-likeness in commercial libraries



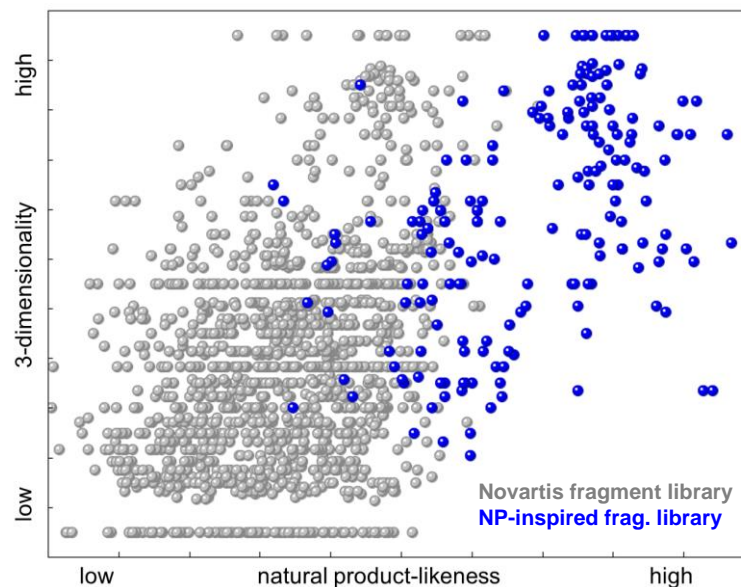
NP-inspired fragment library

H. Prescher, P. Ertl *et al.* *Bioorg & Med Chem* 25, 921 (2017)



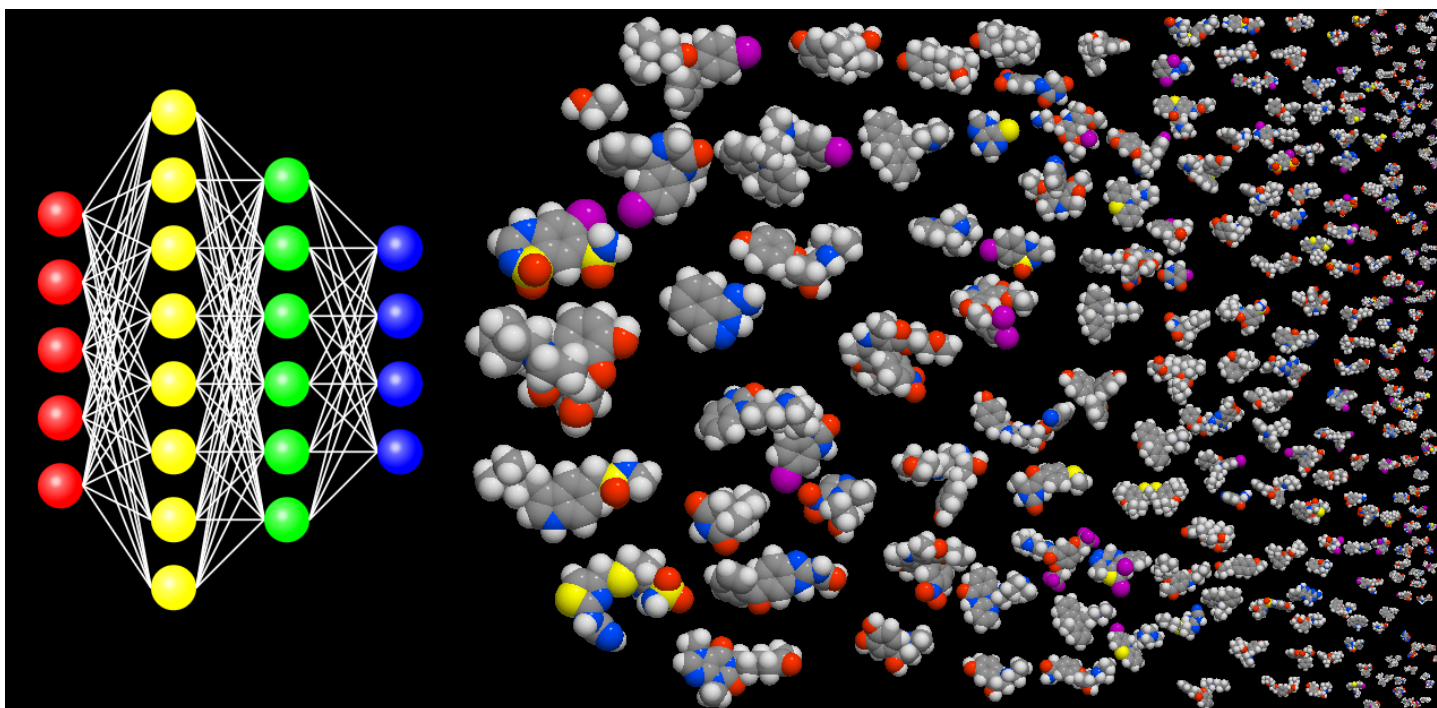
A fragment library consisting of ~250 3D-shaped, NP-like fragments was assembled by:

- NP degradation and diversification
- identification of NP-like fragments from commercial sources



Generation of novel molecules by deep neural networks

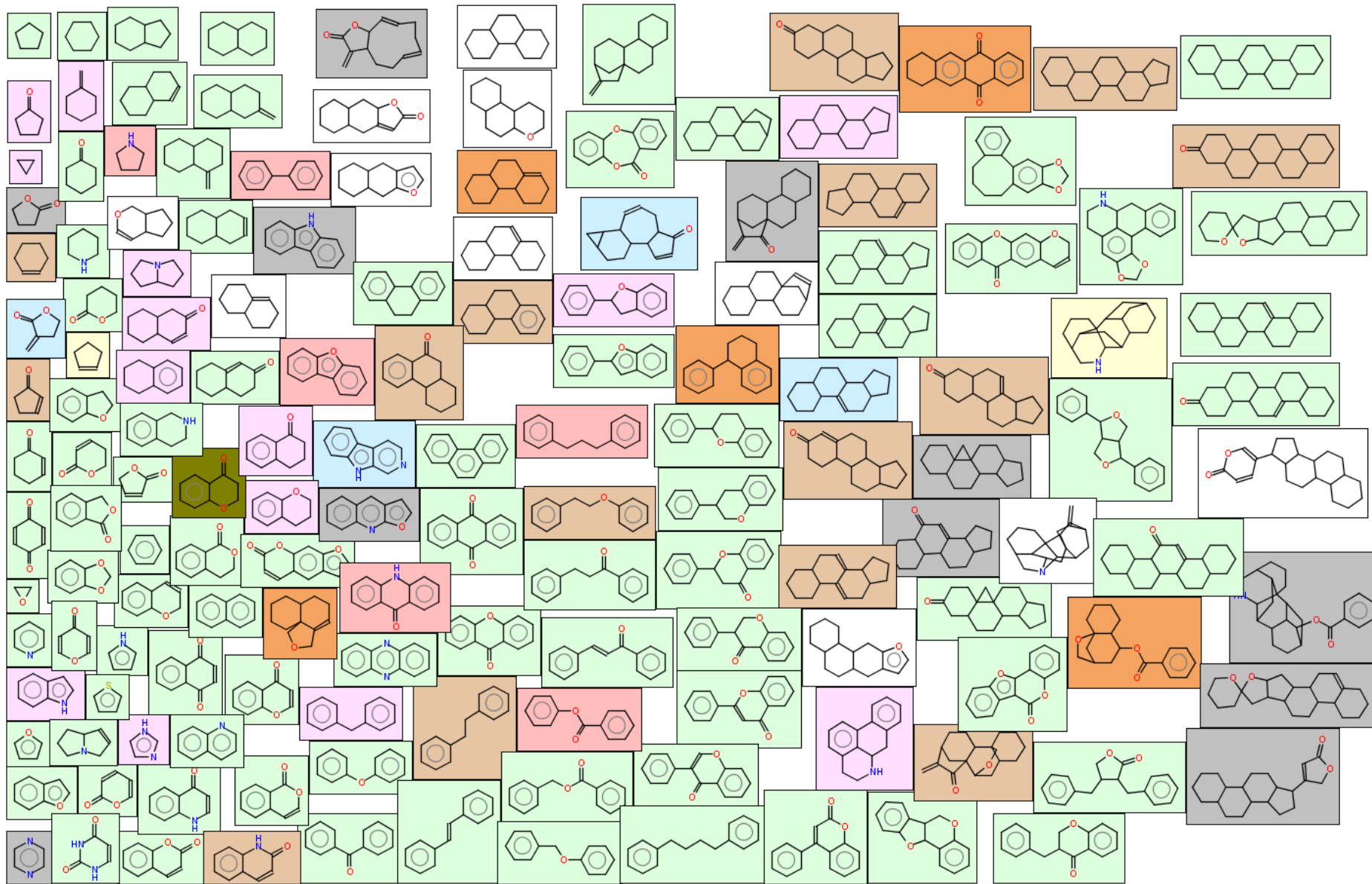
P. Ertl *et al.* arXiv: 1712.07449 (2017)



LSTM deep recurrent neural network may be trained to generate molecules with desired characteristics. Trained on a large collection of NPs the network will generate “non-natural” natural products – i.e. novel, diverse molecules covering well the NP chemical space.

Target preference

GPCR, kinase, protease, other enzyme, ion channel,
nuclear receptor, transporter, epigenetic, multiple targets



Summary

- natural products are very promising class of molecules to be used in drug discovery
- they have been optimized in a very long natural selection process for optimal interactions with biological macromolecules
- the high bioactivity potential of NPs is encoded in their structural features, that are distinctly different from those of synthetic molecules
- sophisticated cheminformatics methods are needed to master the challenge of analyzing the NPs and learning from their structures: substructure analysis to identify scaffold, substituents or functional groups typical for NPs, NP-likeness, target analysis
- application in drug discovery: virtual screening, library design, fragment screening or generation of “non-natural” NPs by deep neural networks