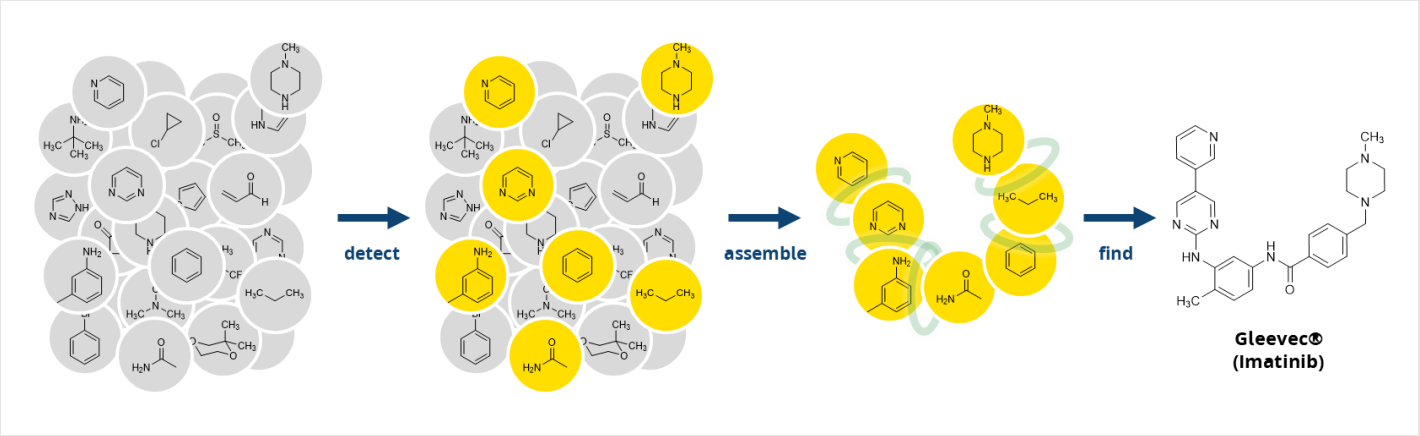
White Paper

BioSolveIT
expect actives!

Deep Dive into Accessible Chemical Spaces

Quickly discover novel IP in huge Chemical Spaces that are set up with your proprietary chemical know-how and in-house building blocks. Find in zillions what will actually land on your bench. Save substantial time and money with Chemical Space navigation technology that is used globally across pharma, crop science, and academia.



Capitalize on your resources and the vast know-how of chemistries developed in your company. There must be hundreds — if not thousands — of protocols for parallel syntheses.

- capturing chemistry and building blocks
- billions of purchasable compounds covered
- > 10^{15} compounds searched in minutes
- proven to work across pharma

This knowledge — probably among your company's most valuable assets — easily covers billions of products, of which only a small fraction has ever been made. The recipes are known, all those compounds are most likely accessible with limited effort, and your chemists will know how to make them. Combining all the information results in a method that mines new leads from your very own Chemical Spaces within minutes. In this white paper we show how easily this can be achieved and provide ample evidence that the procedure does work in practice. It has already led to new active scaffolds in numerous therapeutic projects in Big Pharma.

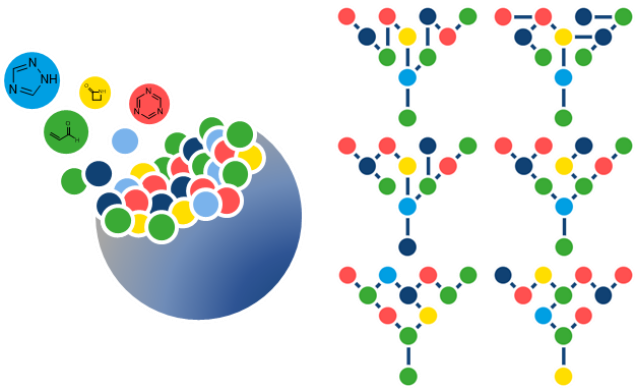
Authors
Christian Lemmen, Marcus Gastreich, Alexander Neumann
BioSolveIT GmbH
An der Ziegelei 79
53757 St. Augustin, Germany

For further information:
biosolveit.de
contact@biosolveit.de
Phone: +49-2241-2525-0
© 2020 BioSolveIT

Unlimited Accessibles

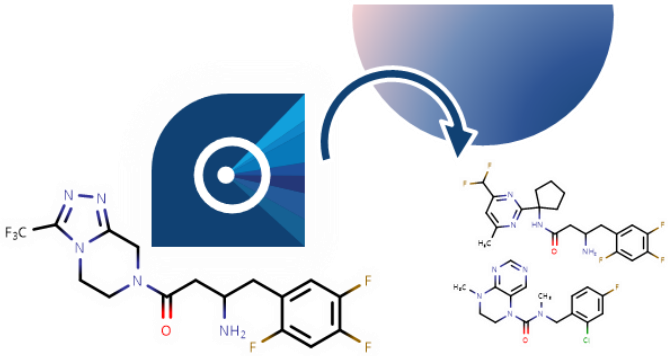
Discover novel IP in your own Chemical Space, based on your proprietary building blocks and unique knowledge, or in spaces from commercial compound vendors offering synthesis-on-demand.

Sizes beyond 10^{25}



Sometimes one cannot see the wood for the trees, although all the components for the solution are at hand. The number of compounds one can principally synthesize is determined by the chemists' know-how and available building blocks. Vast Chemical Spaces capture this knowledge, yet only with efficient screening solutions it is accessible in an almost interactive fashion.

BioSolveIT's novel Chemical Space navigation platform **infiniSee** finds molecules of interest in screening libraries or Chemical Spaces of almost infinite size. Given a template or query molecule, infiniSee searches billions and delivers in seconds. The underlying concept of molecular similarity (**FTrees**) and on-the-fly solution generation (**FTrees-FS**) guarantee accessible, unexpectedly chemically–related molecular htis.

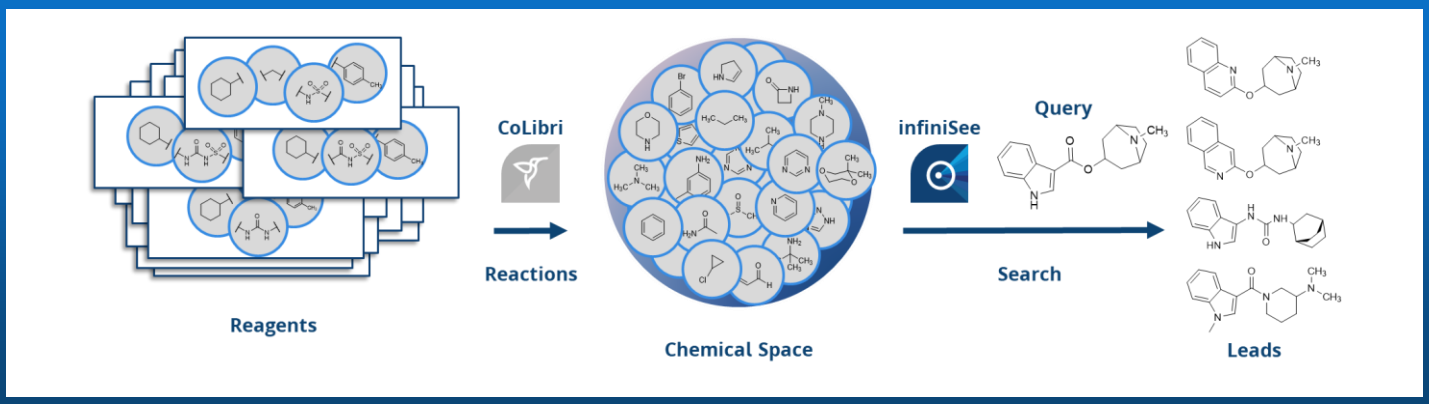


- Commercial Chemical Spaces based on reliable synthetic protocols **complement** in-house libraries with unique building blocks, knowledge, and methods.
- Synthesis requires manpower, building blocks, and in many cases several attempts. **Save time and resources** by cost-effective purchase of desired compounds from our collaborating vendors.
- More than **16 billion of tangible** molecules to mine and **order straight away**. Enamine guarantees delivery rates of >80% within 3 weeks time.[1] WuXi's GalaXi compounds extend possibilities by another 2 billion.

The Big Workflow Picture

5 steps to capitalize on your company's most valuable asset

1. All your in-house chemistry know-how is captured and stored in a Chemical Space
2. Employ a similarity search method proven to be outstanding at scaffold hopping
3. Search the Chemical Space using your unique virtual product assembly
4. Hit lists are reported back to the user — providing synthesis protocols and reagents.
5. Visual inspection quickly leads to novel leads from the innumerably large space



Capturing Chemistries

It all starts with capturing chemistries in a computer-readable fashion: Sketch or write down a description of reaction protocols that range from simple two component reactions to multi-step reactions involving four or more reagents and by-products. We have done this for large parts of the literature for you already. In the end the basic principle is the same: A scaffold — potentially with some variations — is formed, and a certain amount of attached side-chains give rise to a combinatorial explosion in the number of different products.

- library protocols stored in computer-readable format
- exponential rise in numbers of products
- multiple input formats supported

Assume you had three sets of reagents — for simplicity say 1000 each. This would result in a library containing 1000 x 1000 x 1000 = 1 billion products.

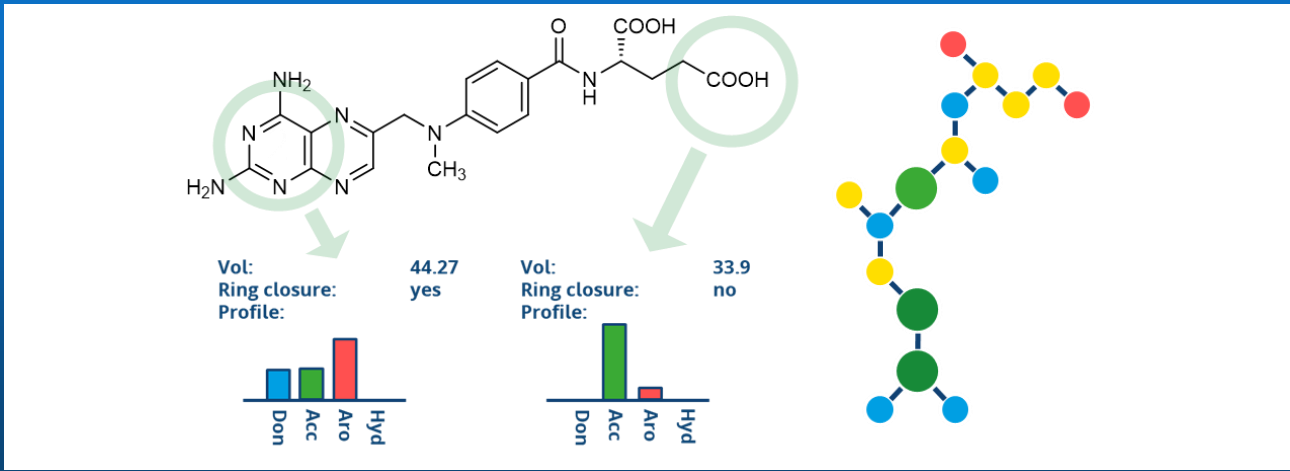
Our app **CoLibri** is capable of taking reagent lists in various formats and reactions as either RXN or SMIRKS. Based on this input it creates your virtual Chemical Space.

Scaffold Hopping: FTrees

Molecular similarity by alignment

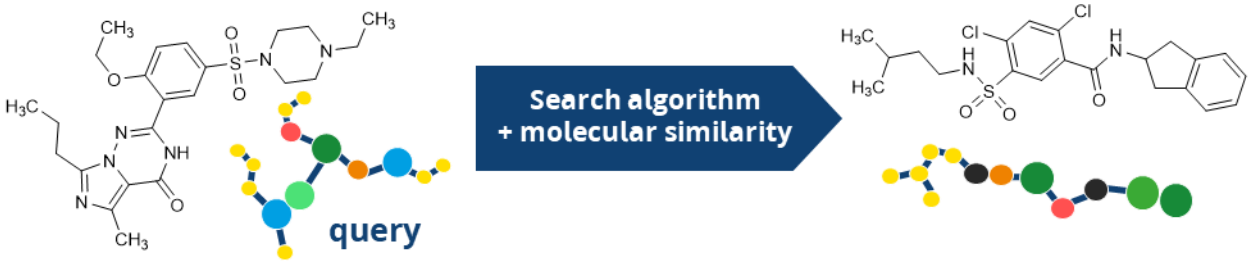
With your in-house Chemical Space at hand, the Feature Tree software (**FTrees**) performs similarity searches in this space.[2-4]

A Feature Tree represents the molecule as a so-called reduced graph with physico-chemical properties. This makes it detect distant similarities and thus scaffold hop.



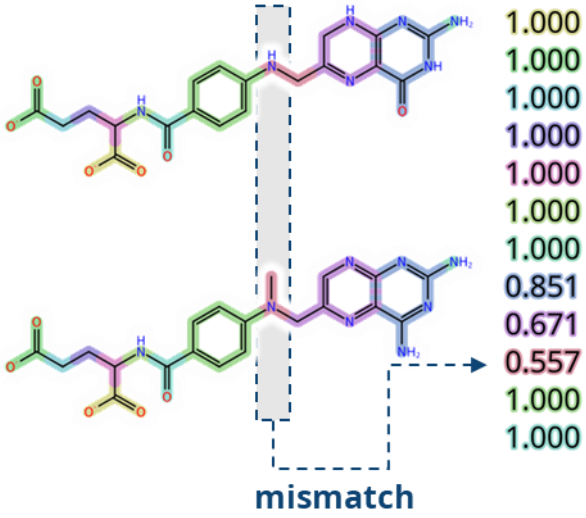
It represents functional groups as well as rings as single nodes. The physico-chemical properties of the substructure represented by a node are stored in a chemistry property profile for that node.

- different than traditional similarity, yet chemistry-aware
- proven to hop scaffolds



The **overall topology is preserved** in the Feature Tree: Nodes representing substructures that are connected in the molecule are also connected in the Feature Tree. Now, if two molecules are both represented by their respective Feature Tree, FTrees is able to calculate from among all the topology preserving mappings of the two Feature Trees the one that gives the highest possible similarity value. How is the similarity value calculated? By an alignment of trees: If two nodes are mapped then the difference of their respective property profiles gives a local similarity, and the overall similarity is just the normalized sum of these local similarities.

See the similarity, stay in the driver's seat



An FTrees mapping of dihydrofolate (top) onto methotrexate (bottom). Sub-structures of the same color are mapped onto each other and their given FTrees similarity is presented on the right.
FTrees is able to recognize the similarity of the electron pair donor (highlighted in blue) of the heterocycle as well as to differentiate between the non-methylated and methylated nitrogen atom. Good to know: This is irrespective of the protonation states.

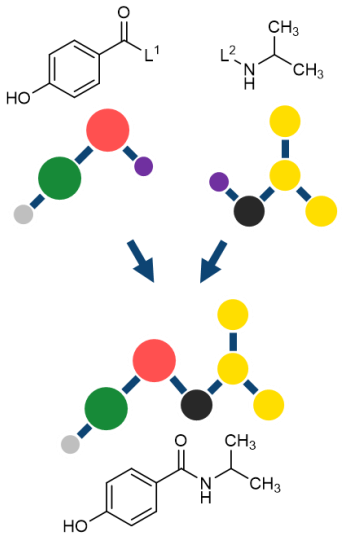
More than your typical screening platform

- **lightning-fast screening**: screen millions of molecules within seconds
- Other than classical Tanimoto descriptors Feature Trees capture molecular topology, allowing to discover **distant neighbors** that surprise and inspire you
- technology among the top performers when it comes to enrichment rates
- best in class in terms of scaffold hopping

Search for neighbors

It gets even better: FTrees not only has all these fantastic attributes that make it the perfect choice for similarity searching across traditional libraries, but it can also be used to search across Chemical Spaces. In the same way as for a whole molecule, you can represent the building blocks in a Chemical Space as Feature Tree fragments.

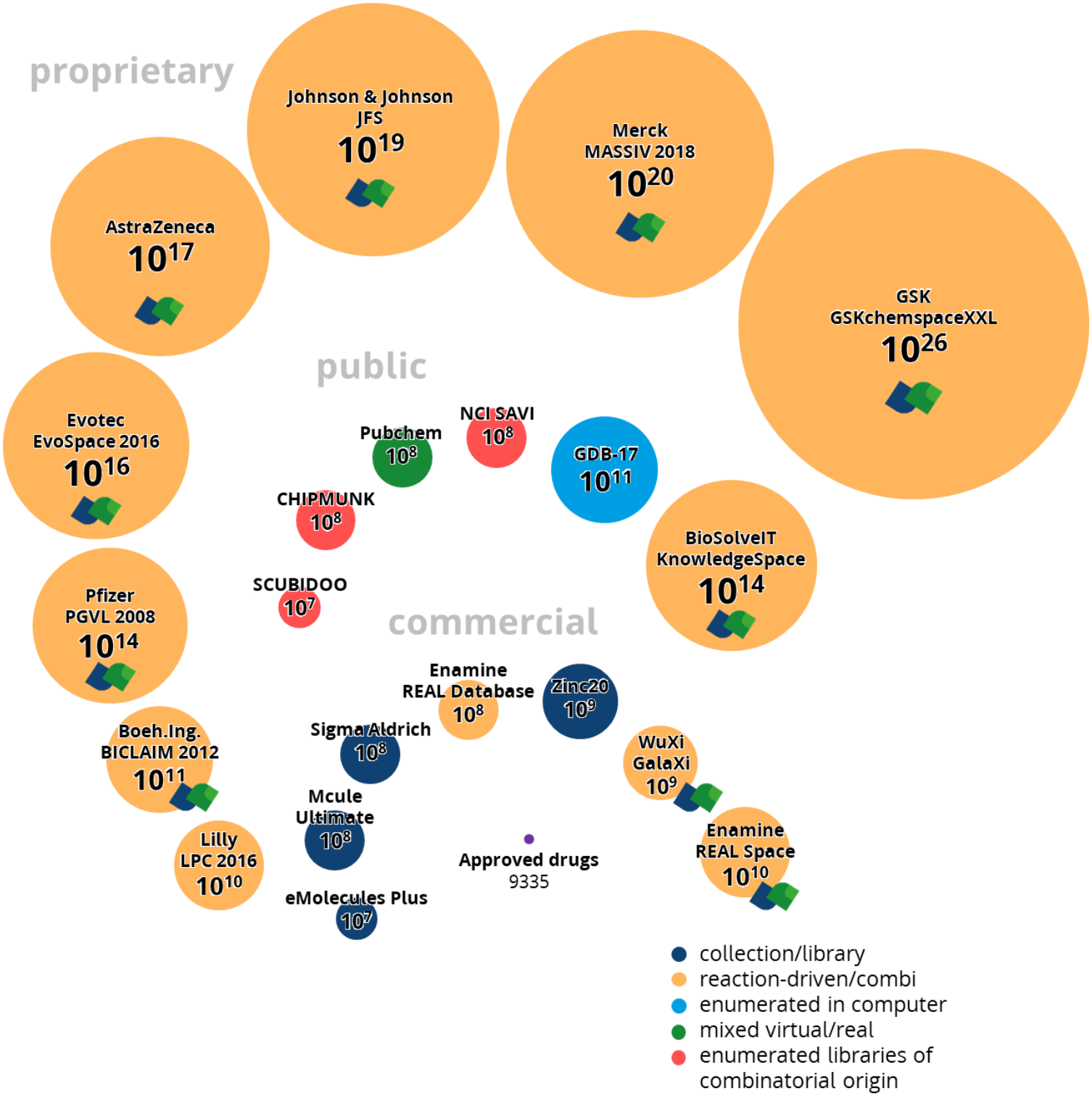
If you provide a query molecule as input, an extension module called **FTrees-FS** is capable of searching the Chemical Space; the result will be a list of the most similar product molecules generated. This is done by recursively detecting most similar substructures and assembling multiple building blocks to virtually grow a set of molecules from the Chemical Space.



Two building blocks are represented by two Feature Trees. Forming a "bond" between the link atoms leads to a bigger Feature Tree, which translates back to a new molecule. The history behind its formation is preserved.

The Advent of Chemical Spaces

Over the past decade, global pharmaceutical companies have discovered the possibilities behind Chemical Spaces.[5] The biggest and renowned proprietary compound libraries were created with CoLibri reaching colossal sizes of 10^{26} (with rising tendency). Novel IP was highly successfully mined, actives found in previously uncharted Chemical Space (see refs. at the end).



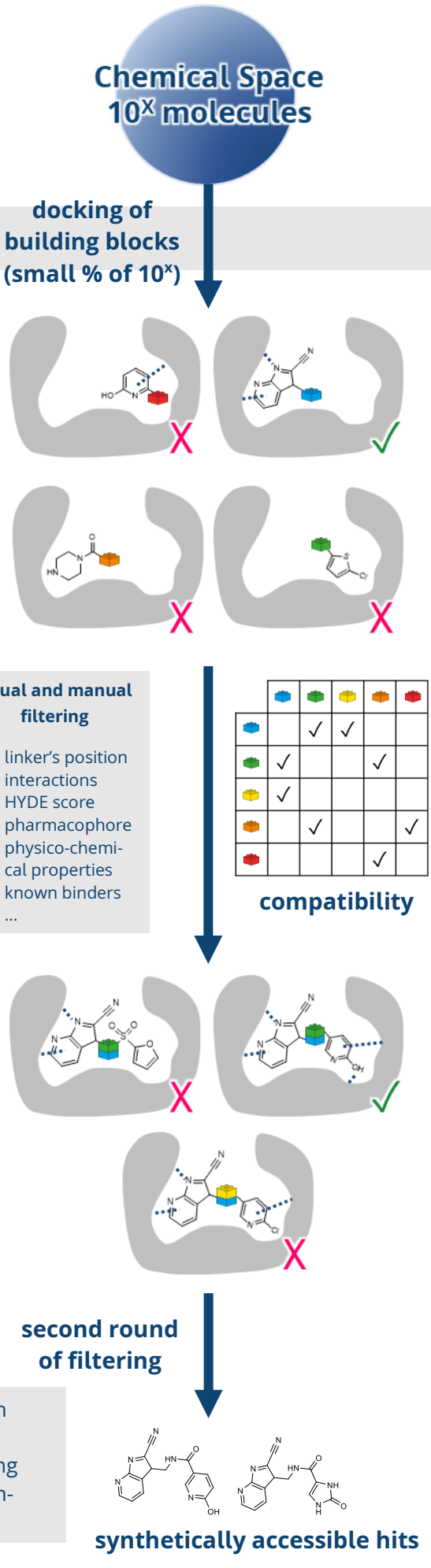
3D: Chemical Space Docking

Step by step solution generation

The colossal size of Chemical spaces poses a challenge for traditional docking methods due to their sheer size and the associated computations. Docking calculations of 10^{15} sized spaces would take thousands of years, even on modern hardware for vast Chemical Space, disqualifying them for practical applications.

Chemical Space Docking takes a different approach how to handle these massive numbers. In the first step you **dock your building blocks which represent only a miniscule percentage of your actual Chemical Space**. Then we assess their binding mode with desolvation-aware scoring. Automated filtering removes binding modes of building blocks with unwanted linker positions, low scores, and few interactions. Optionally include known binders to fine-tune the docking process with pharmacophore constraints or template docking. Bump checks and more are performed. In the next step you let your selection "evolve" based on which synthons/linking partners are compatible. The whole procedure can be done by a single person in short time with a standard work station.

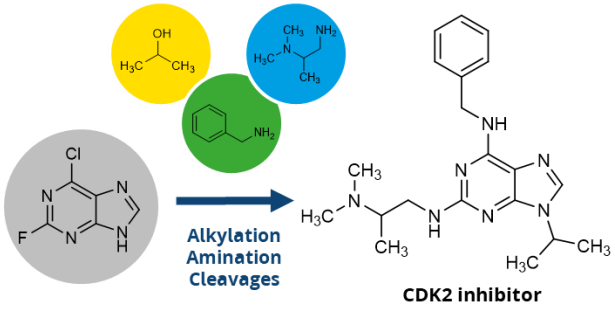
Subsequently, you can **cherry pick** the most promising and interesting candidates for further growing. Do this manually, by checking for specific pharmacophores/(un)wanted substructures, or remove those with unspecific binding. Your expertise during the visual inspection process maximizes the likelihood of success.



- combination of the vastness of Chemical Spaces with SBDD
- explores more compounds than conventional docking
- manual and automated filtering for desirable und unwanted properties

Real-Life Application

Capturing CDK2 inhibitors as published in Science



- Example from an article published by Gray et al. represented virtually using BioSolveIT technology [6]
- 2,6,9-tri-substituted purines synthesized from a much purine scaffold
- Involved methods: solid-phase amination, alkylation reactions, and a subsequent acidic cleavage

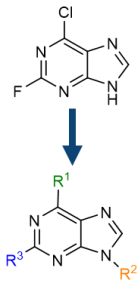
For the expert: The genesis of a 10^9 molecules Chemical Space

Representing this protocol in the computer is a simple procedure. What is essential are **educts**, **products**, and a **formalized description** of the newly formed bonds. The purine scaffold will be termed as a "core", whereas the reagents for substitution will supply the "R-groups" (i.e., the residues in the resulting products — not to be confused with the *R*-notation in chemical formulas). Where exactly a bond is created is defined on the basis of dummy or "linker atoms". The core and educts therefore need to be equipped with these.

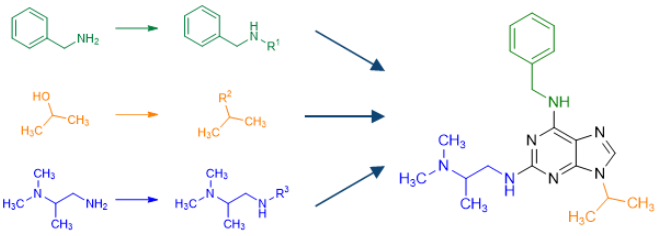
In Step 1, CoLibri is used to replace the amine-H at position 9 and halogens at positions 2 and 6 at the core for linker atoms. The linker atoms are denoted R^1-R^3. The corresponding SMIRKS-like CoLibri rules for the core is a simple three-liner in which a dot denotes the cleavage of a bond and the [n*] notation introduces the linker atoms.

```
[Cl][*]     >>  [Cl].[*][1*]
[H:1][N:2]  >>  [H:1].[N:2][2*]
[F][*]      >>  [F].[*][3*]
```

Three lines of code to prepare the purine core.



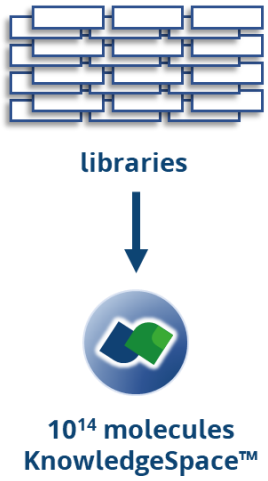
The purine core, after applying these transformations, is represented on the left. With 3 more rules, CoLibri prepares the educts for the substituents: It clips the amine and alcohol H-atoms to form the "naked" R-groups. Finally, you need to define how clipped building blocks may be recombined ("linker compatibility"). Certainly, we can also do all this for you in a Service setup. Please get in touch!



Combinatorial build-up

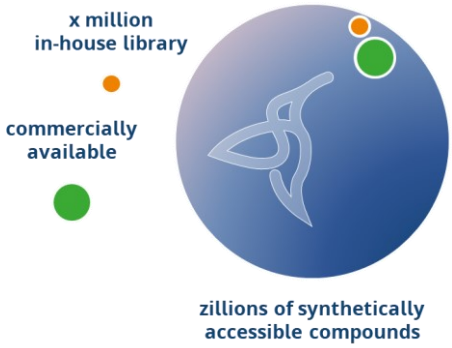
The original Science publication reports 19 reagents for R^1, 7 for R^2, and 10 educts for R^3 resulting in 19 x 7 x 10 = 1,330 products. Reagents lists were combined for R^1 and R^3 and added 309 additional primary amines. Furthermore, R^2 was extended by 274 alcohols from the same source. All this was done with fine medicinal chemistry expertise. Our resulting search space consists of **70 million compounds** (499 x 281 x 499 = 69,969,281) for this single reaction protocol.

Now, how about more reactions?
We processed more than 120 combinatorial libraries with three or four R-groups. Our virtual Chemical Space (**KnowledgeSpace™**) today consists of **more than 10^{14} virtual products,** and this is nowhere near the limit as you will see below. The KnowledgeSpace™ comes free of charge with our software.



Search for neighbors

Everything is based on the easy-to-read chemistry description standards RXN/SMIRKS which allow all kinds of substructure detection and replacement. Using the CoLibri procedure, the raw input data will transform into a Chemical Space.



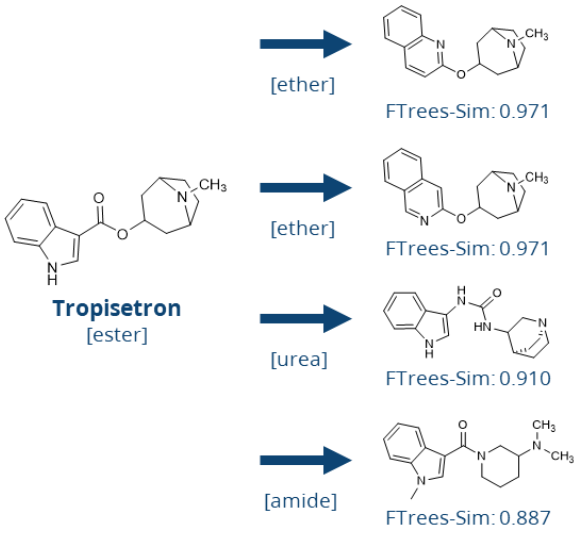
No matter how big your in-house library and no matter how many compounds you acquire to add to it – it will only be a tiny fraction of what your chemists are capable of synthesizing.

The entire CoLibri procedure is **scriptable and supported by 2D visualization** of substructure matches. The real power of the mechanism obviously comes from its ability to process hundreds or even more protocols as above and make them accessible as a single enclosed Chemical Space. Here it is important to note that CoLibri is able to **remove the redundancy** from a dataset by representing duplicate building blocks in the input using only one representative instance and maintaining a lightning fast, hash key-based lookup table to map any results data back onto the original input. This way CoLibri reports not only virtual products to the user but is also able to **annotate these results with the chemistry library protocol and the particular reagents** that form a product.

Already in 2008, Pfizer mined from trillions of compounds within a few minutes

Amongst the pioneers in proprietary Chemical Space set-up, Pfizer, in 2008, created the PGVL combinatorial chemistry protocols.[4] A total of **358 combinatorial libraries** were converted into a single, concise Feature Tree Chemical Space comprising a total of 3,000,000,000,000 (3 trillion) virtual products. This Chemical Space was then validated in a variety of ways. In summary, with a sample set of 1,790 query building blocks (5 randomly chosen for each protocol), it was possible to retrieve three or more queries in the top 100 ranks for 99% of the protocols. Considering the vast number of products in the space, this is literally akin to finding a needle in a gigantic haystack.

When applied to searching a sample set of 1,661 compounds from the WDI, 91% retrieved a compound with similarity of 0.9 or higher, demonstrating that the **Chemical Space covers a broad range of drug-like molecules**. 90% of these searches had a search time of less than 20 minutes on a single CPU — back in the days! Also, the results covered a broad range of different chemistries in that 50% of all protocols were employed at least once to form the top-ranking product for a search.



Most interesting is of course the **ability to scaffold-hop** from one active hit to another attractive series. The most interesting hits were found in the range of similarities between 0.90 and 0.95. In other searches at Pfizer, a central pyrrolo-indole scaffold was replaced by an indanyl piperazine ring, a central ketone group was substituted by an amide bond linker, or a phenothiazine heterocycle was replaced by a phenyl-indole scaffold — to mention just a few scaffold hops.

Other methods incapable of finding FTrees-FS hits

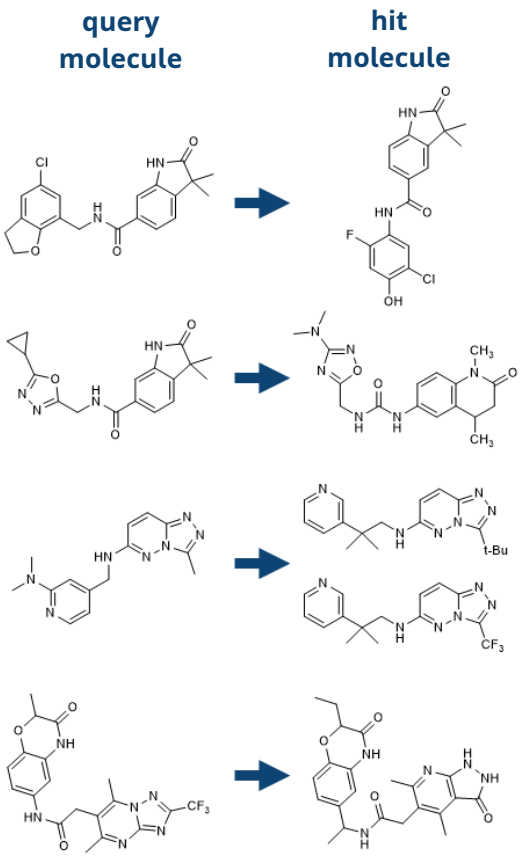
In the same work, two marketed drugs for the serotonin 5-HT3 receptor were used as queries which produced active hits originating from a variety of chemistries such as ether and amide bond formation or a urea reaction. Unfortunately, some of the more exciting hits could not be disclosed by the authors. Interestingly, for a sizeable number of cases the hits produced by FTrees-FS had quite low Daylight or Pipeline Pilot (FCFP4) fingerprint similarities, which further underlines the uniqueness of these results. Not only are these methods easily accessible for virtual libraries of this size (10^{12}) due to sheer size, but also these other methods would only have retrieved these solutions ranked worse than a few billion others because of the low similarity scores that do not capture distant, non-obvious similarities.

Nanomolar inhibitors with novel scaffolds by Boehringer Ingelheim

In another study, Uta Lessel of Boehringer Ingelheim presented at the 8th International Conference on Chemical Structures[7] **two successful applications** of Feature Tree Chemical Space searches based on combinatorial library protocols. Based on a sizeable number of combichem protocols, the so-called BI-CLAIM Chemical Space was generated on the basis of roughly 1,600 scaffolds and about 30,000 unique reagents. Thousands of compounds were actually synthesized for each of these protocols, which however amounts to only a tiny fraction of the 500,000,000,000 (500 billion) virtual products covered by BI-CLAIM then. The typical workflow described in this presentation has two parts. First a literature active is taken to search the Chemical Space and yield in the order of a few thousand hits. Then a shape filter is applied in order to provide a first pass validation of the hits, and finally the results are grouped by scaffold and visually inspected. Part two of the workflow is to manually select the most promising scaffolds and design focused libraries around them, or purchase prototypes of those compounds if commercially available. If activity is found and confirmed in these series, then one or more rounds of refinement based on traditional medicinal chemistry are applied. The researchers from Boehringer Ingelheim reported on a **GPCR and a proteinase project** where these procedures quickly led to nanomolar inhibitors in novel series.

SAR by Space

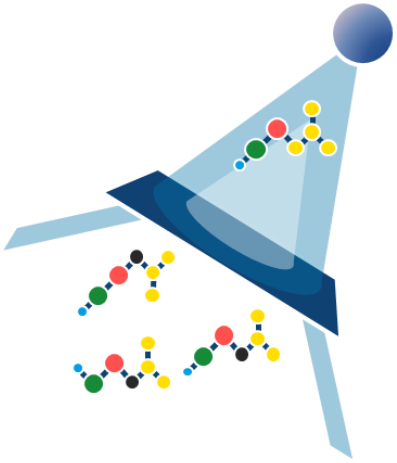
In a recent publication, Enamine's **REAL Space™** was mined to identify novel bromodomain-containing protein 4 (BRD4) inhibitors using FTrees-FS.[8] Commercially available neighbors to query molecules with drug-like properties were retrieved from the fragment-based Chemical Space using the similarity search of FTrees-FS. Starting from very, very weakly actives, 5 micromolar hits have been identified and verified within less than 6 weeks, including the pharmacological assessment of IC_{50} values. This unconventional approach was more efficient for hit expansion compared to the straightforward fragment-based discovery which required synthesis and biological evaluation of thousands of analogs for the initially discovered active fragments. The retrieved hit molecules exhibited **similar pharmacophoric properties**. The strategy required 100-fold fewer compounds to be synthesized and screened, it was **faster and therefore way more cost-efficient**.



Future Perspectives

Chemical Space Docking

The ability to screen vast Chemical Spaces as a **primary source for novel intellectual property** is becoming more and more important in the modern drug discovery process. However, the main obstacle of exploiting the uncharted territory is the computational effort and speed behind the process: docking 10^{12} compounds with one second calculations per molecule, would take over 15 million years to screen. Accessing such colossal numbers requires ground-breaking methods but bears unlimited potential.



SpaceLight — Close neighbor collection also for 10^{15} and more

This **topological similarity search algorithm** of this novel method is based on an efficient recombination approach.[9] Similar to FTrees-FS, SpaceLight calculates similarities for query molecules based on fragments and retrieves 'close neighbors'. Using Extended-Connectivity Fingerprint (ECFP) and Connected Subgraph Fingerprint (CSFP), SpaceLight **exploits the combinatorial character** of fragment-based Chemical Spaces resulting in unrivaled performance. Billions of compounds can be searched within seconds with SpaceLight.

Even bigger – even more diverse

In the past decade, Big Pharma and compound makers have recognized the **true potential behind Chemical Spaces**. The underlying concept of maximizing in-house resources, be it building blocks, experience, or knowledge, to create huge and diverse compound libraries has already brought forth several drug candidates where classical approaches failed. The field of Chemical Spaces is continuously expanding as researchers realize the hidden possibilities.

When will you go and create yours?

References

[1] Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681. https://doi.org/10.1016/j.isci.2020.101681.

[2] Briem, H.; Lessel, U. F. In Vitro and in Silico Affinity Fingerprints: Finding Similarities beyond Structural Classes. *Perspect. Drug Discov. Des.* **2000**, *20* (1), 231–244. https://doi.org/10.1023/A:1008793325522.

[3] Kogej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint Based Similarity Searches for Targeted Class Compound Selection. *J. Chem. Inf. Model.* **2006**, *46* (3), 1201–1213. https://doi.org/10.1021/ci0504723.

[4] Boehm, M.; Wu, T.-Y.; Claussen, H.; Lemmen, C. Similarity Searching and Scaffold Hopping in Synthetically Accessible Combinatorial Chemistry Spaces. *J. Med. Chem.* **2008**, *51* (8), 2468–2480. https://doi.org/10.1021/jm0707727.

[5] Hoffmann, T.; Gastreich, M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discovery Today*. Elsevier B.V. **2019**, pp 1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013.

[6] Gray, N. S.; Wodicka, L.; Thunnissen, A. M.; Norman, T. C.; Kwon, S.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S. H.; Lockhart, D. J.; Schultz, P. G. Exploiting Chemical Libraries, Structure, and Genomics in the Search for Kinase Inhibitors. *Science* **1998**, *281* (5376), 533–538. https://doi.org/10.1126/science.281.5376.533

[7] Lessel, U. at ICCS 2008, Noordwijkerhout, Netherlands, **2008**

[8] Klingler, F. M.; Gastreich, M.; Grygorenko, O. O.; Savych, O.; Borysko, P.; Griniukova, A.; Gubina, K. E.; Lemmen, C.; Moroz, Y. S. SAR by Space: Enriching Hit Sets from the Chemical Space. *Molecules* **2019**, *24* (17), 3096–3106. https://doi.org/10.3390/molecules24173096.

[9] Bellmann, L.; Penner, P.; Rarey, M. Topological Similarity Search in Large Combinatorial Fragment Spaces. *J. Chem. Inf. Model.* **2020**. https://doi.org/10.1021/acs.jcim.0c00850.