Design hub for early phase drug discovery

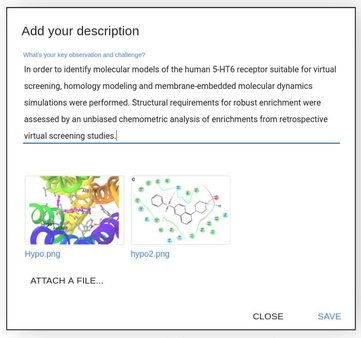
ÁKOS TARCSAY, IVÁN SOLT, ANDRÁS STRÁCZ

Abstract Drug discovery is an iterative process of hypothesis construction relying on observations and validation through triggering new observations mainly by synthesis of new chemical entities. During the evolution of an idea to reach selection for synthesis, evidence and prediction results are collected and assessed and scrutinized by the project group. Therefore, the success of recent drug design depends on how data is turned into information and how much knowledge is extracted out of it. Accordingly, attempts toward connecting data sources or making an even broader spectrum of data available in centralized data lakes with corresponding access engines operating on top drive contemporary development and represent a key trend. Powerful data analysis (like matched molecular pairs (MMP)) or instant search over large chemical datasets are highly demanded. Depending on the volume and quality of the raw data, model building approaches may play crucial role in the preprocessing steps. Data analytics platforms with supervised or unsupervised methods are applied like linear fitting, clustering, pattern recognition or neural networks. These models are moving beyond the raw information and the extracted correlations can be exploited on novel, hypothetical structures to judge them in a triaging phase, before deciding on synthesis.

Effective coordination of the hypotheses and compound series in projects where multiple groups are collaborating requires access to optimized and dynamically changing information. Accordingly, the major problem is the collection, grouping, management, and overview of the relevant information (ideas, calculated properties, related data from databases, graphics, comments, attachments, etc.) within a single application.

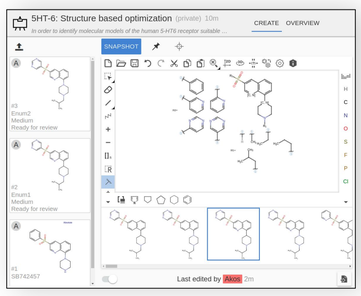
The goal of this presentation is to introduce the Marvin Live [1] platform for integration of a wide variety of data sources and services to augment real-time design. Marvin Live offers a vendor agnostic, real-time plugin system that can be configured to the current information needs. This allows the seamless integration of in-house databases, local models and workflow tools (KNIME, Pipeline Pilot). We are presenting two use cases: first, we will show how an MMP analysis based on ChEMBL data can support designing out hERG liability. Second, we will exemplify the simultaneous and instant searching in various databases like ChEMBL, SureChEMBL, PubChem and vendor catalogs (eMolecules, Mcule, Molport, Enamine). Utilising novel search engines, this service provides results within seconds to a compound collection with a total size of >800M molecules. It supports estimation of freedom to operate, novelty and provides a quick insight to reagent and purchasable compound availabilities.

Observation



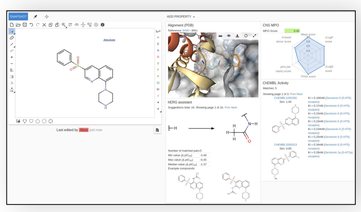
Capture and share rationale

Hypothesis



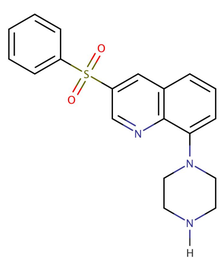
Expand into a design set

Brainstorm

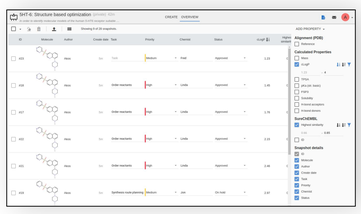


Gather information

In-depth look

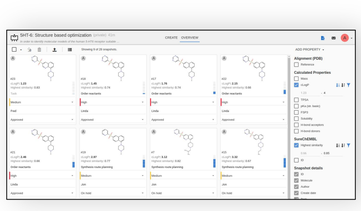


Weigh the evidence



Balance attributes, set priorities

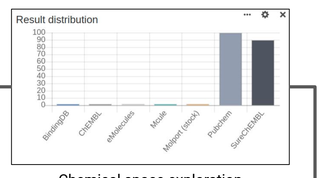
Choose



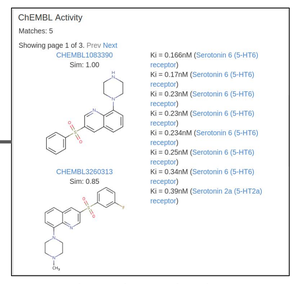
Review and decide

Synthesize

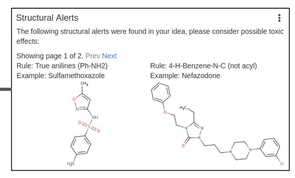
Plugins



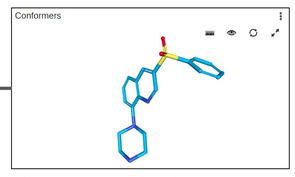
Chemical space exploration



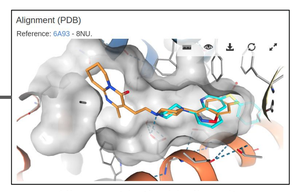
Assay data exploration



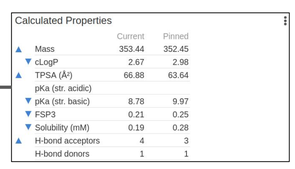
SMARTS pattern based toxicophore checks



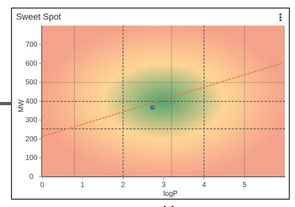
Low energy conformations on blue-red color scale



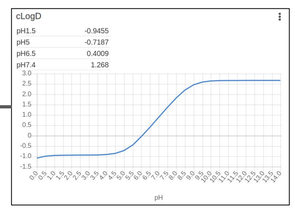
Alignment in the binding site



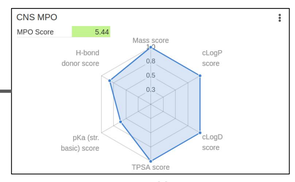
Phys-chem attributes



Sweet Spot[6] view



Predicted logD profile

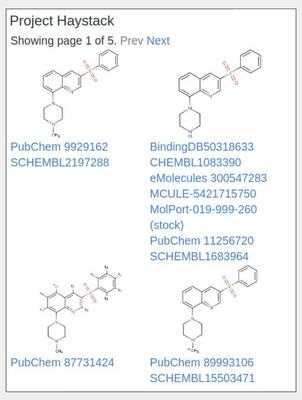


CNS MPO[7] score

MMPA based assistant

Gathering information from the available chemical space

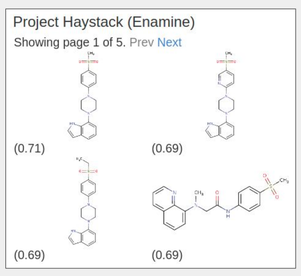
Chemical space exploration



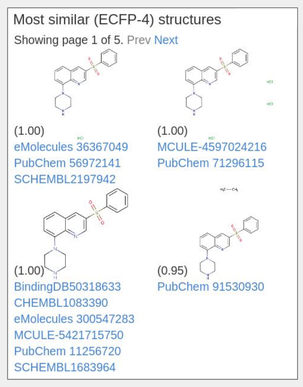
Substructure hits from the combined dataset

Grasping the relevant chemical information

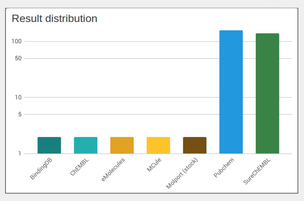
- As the structure is drawn, search results come back continuously from relevant data sources
- Quick overview of the coverage of analogs in public datasets
- Get information about available assay results - explore results of analogs
- See vendor availability - avoid the availability bias
- Indication on patent coverage - understand the risks
- Dig deeper through the linked data sheets for each hit
- Make the right decision for your design



Results from the Enamine REAL dataset

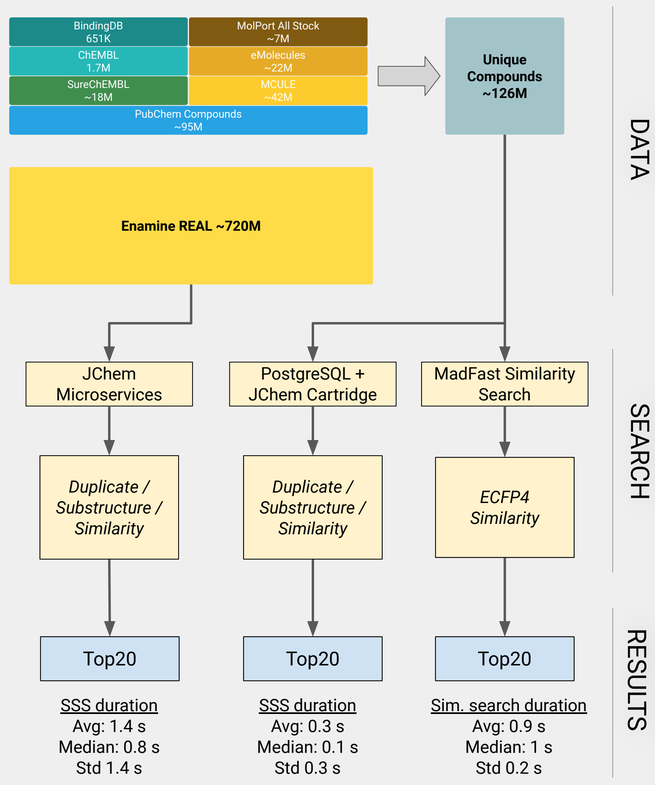


Similarity hits from the combined dataset



Estimated distribution of hits

Searching in a collection of databases



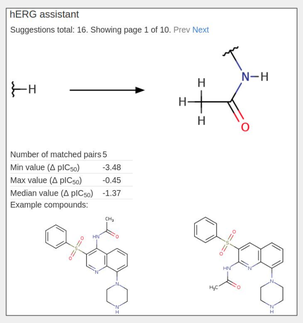
The Haystack project aims to allow exploring the relevant chemical space during a design process without the need to rely on a large number of external services with querying or standardization differences. To achieve this, eight databases were chosen to create a prototype of a single service within Marvin Live:

- Enamine REAL library (~720 million virtual compounds) was indexed without any modification using the JChem Microservices.
- Seven databases (see figure above) were standardized and deduplicated. Unique structures (~126 M) were collected in a single table and indexed using the JChem PostgreSQL Cartridge.
- ECFP4 fingerprints were calculated for the unique compound set to allow similarity searching using MadFast Similarity Search.

Used hardware: 24 CPU (Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz), 125 GB RAM (95 GB used by the three services), 3.6 TB disk (970 GB used for the data storage)

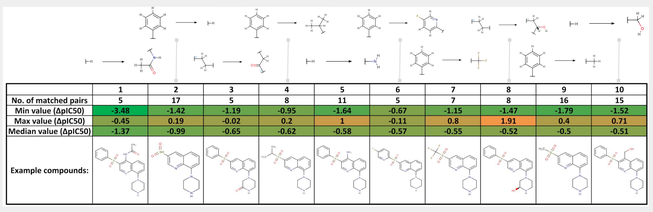
Designing out hERG liability

From MMPA to Assistant

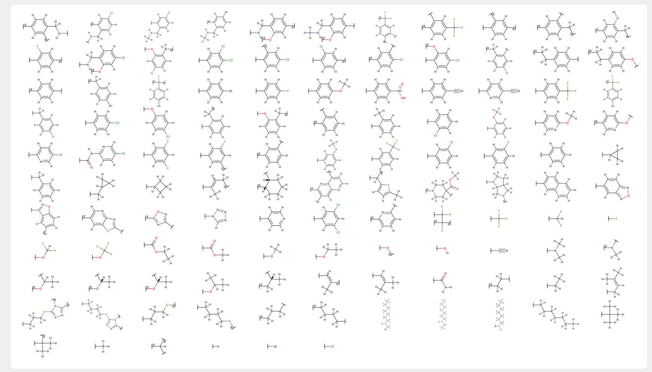


The plugin exploits the extracted knowledge of matched pairs with the following steps

- Select transformations relevant for the idea
- Select only transformations reliable enough (min. 5 evidence)
- Suggest transformations sorted by median effect size in terms of ΔpAct
- Display related statistical results to provide reliability
- Run transformations and suggest structures

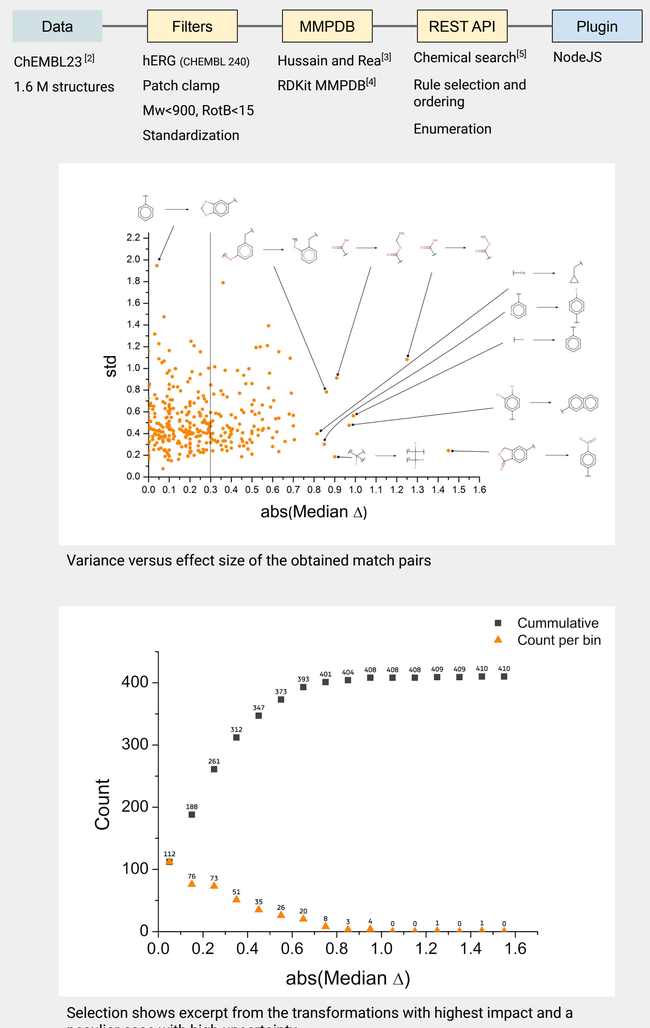


Top 10 suggested transformations with example molecules sorted based on median value



Structural mosaic of all the chemical moieties with relevant transformation in the MMP rules

Matched Molecular Pair Analysis



Variance versus effect size of the obtained match pairs

Selection shows excerpt from the transformations with highest impact and a peculiar case with high uncertainty.

Conclusion The design hub fosters integration of knowledge and predictive models available in various forms into a design space to aid both the ideation and the decision making on synthesis targets. The first detailed example described in this presentation is the collection of analogues from vast amount of chemical space for freedom to operate analysis, supporting information collection based on similar structures or expansion by purchasing. Second, the preprocessed and transformed data relaying on assay results in the form of matched molecular pair analysis is shown to facilitate design towards reduced hERG liability analogues.

[1] Marvin Live, https://chemaxon.com/products/marvin-live
[2] https://www.ebi.ac.uk/chembl/downloads
[3] J. Hussain, C. Rea, J. Chem. Inf. Model., 2010, 50, 339–348
[4] A. Dalke, C. Kramer, J. Hert, J. Chem. Inf. Model., 2018, https://github.com/rdkit/mmpdb
[5] JChem Web Services, https://chemaxon.com/products/jchem-engines
[6] M. Hann, G. Keserű, Nat. Rev. Drug Discov., 2012, 11, 355–365
[7] T. Wager, X. Hou, P. Verhoest, A. Villalobos, ACS Chem Neurosci., 2010, 1, 435-449

atarcsay@chemaxon.com, isolt@chemaxon.com, astracz@chemaxon.com

ChemAxon