SPROUTING OUT NEW IDEAS FROM KNOWLEDGE GROUND

Anna Tomin, Ákos Tarcsay, András Strácz
ChemAxon Kft. Záhony u.7 H-1031 Budapest

A computational chemistry approach, Matched Molecular Pair analysis (MMP) has been selected to reveal the importance of mining and analyzing in-house knowledge-base to support early phase drug discovery research cycle and decision making. [1]
MMP method and publicly available data set from ChEMBL [2] were combined to

- **Increase our understanding of the relation between biology and chemistry on large dataset**
- **Extend current capabilities of ChemAxon toolset via suggesting what compound to make next and identify structural changes that alter on key properties within the framework of Marvin Live.**

MMP: The matched molecular pair is a concept to study the correlation in changes of compound properties that are associated with a single localized structural change. In other words, MMP method provides an automated and systematic compilation of medicinal chemistry rules using compound/property data sets.
In current work a fragmenting - indexing algorithm method by Hussain and Rea [3] was selected and RDKit 'mmpdb' implementation was used for analysis [4]. The entire ChEMBL data set was subjected to MMP analysis, besides, a focused set of data (hERG) responsible in cardiac ion channel regulation were investigated as well.

MMP on ChEMBL

Widespread analysis of millions of chemical and related data from a medicinal chemists perspective is demonstrated here,.

Dataset

The entire **ChEMBL** data set has been subjected to MMP analysis.
Chemical content was pre-processed via the followings:

- Structure standardization, data correction and validation was done using ChemAxon's **Standardizer and Structure Checker** to match the requirements of the MMP algorithm
- Content was limited to **drug-like structures**: molecular weight between 250-900, heavy atom count <150, rotatable bond count <15, peptides were eliminated, largest fragment was kept in case of multi-fragment structures.

Methods

Because of the size of the data set and the heavy computational load, simplifications were applied when running indexing. Therefore, only exocyclic cuts were included in further studies.
A 96GB RAM and 24 CPU machine took 4.5h to fragment, index and analyze the dataset with 80% average load and 80GB peak memory usage.

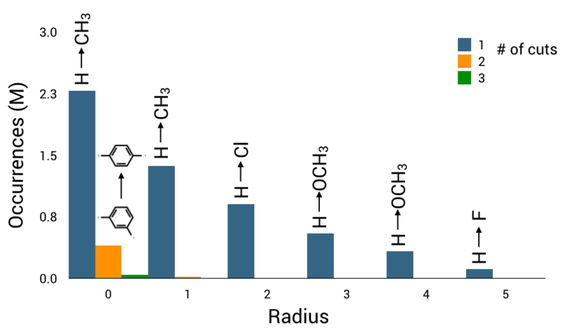


Fig 1. Types and occurrences of transformations identified in ChEMBL data set by MMP. Up to 5 atoms away from the cut atom(s) on the constant part has been considered.

ChEMBL molecules	1.6 M
MMPs	191 M
rules	26.5 M
frequent transforms	0.15 M

Table 1. Overview of the result of the fragmenting-indexing algorithm

Results

77% of the overall transforms is covered by 1500 distinct transformations. The most typical transformations are shown in *Fig 1* and an overview in Table 1.

Multiple cuts

The fragmentation phase cuts 1, 2, or 3 non-ring bonds to provide the constant and variable part in the chosen MMP rule. An overview of the most frequent transformation of a given number of cuts, grouped by occurrences are presented on *Fig 2*. A representative selection of rare transforms are presented on *Fig 3*.

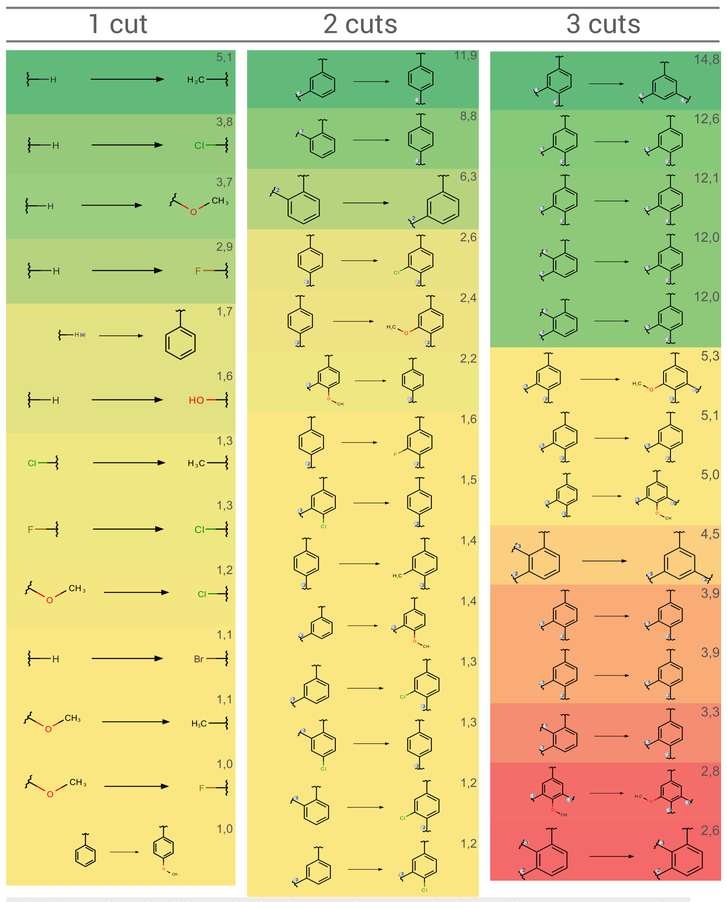


Fig 2. Examples of valid matched pairs at various cuts in descending occurrence order. In the corners the probability of occurrence is indicated as percentage.

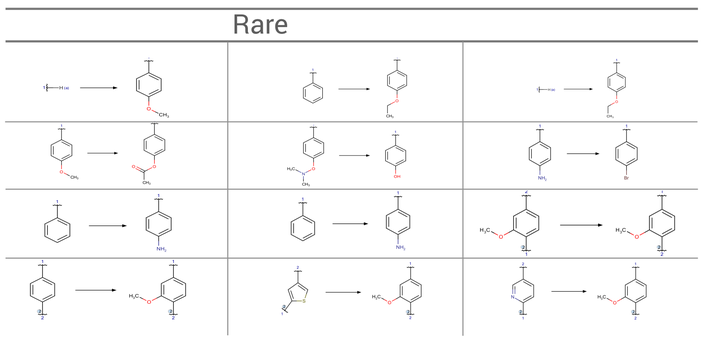


Fig 3. Examples of valid, rare matched pairs at various number of cuts

MMP on hERG

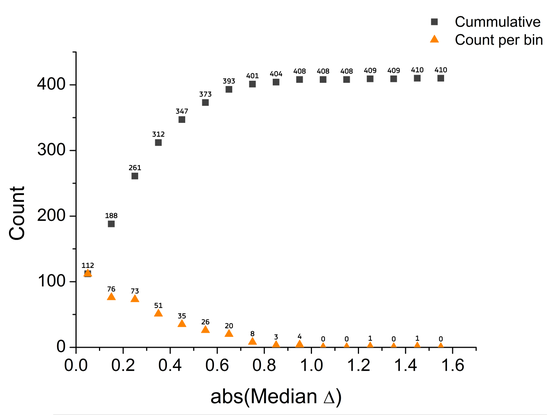
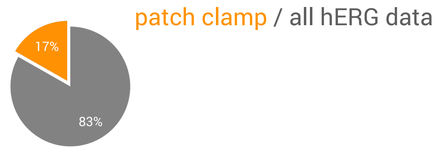


Fig 4. Distribution of median p activity deltas against the number of compounds, showing significant amount of data above 0.3 - a two-fold activity change.

Automated electrophysiological patch clamp allows assessment of hERG channel effects early in drug development to aid medicinal chemistry programs. **hERG** experimental data based on patch clamp method extracted from ChEMBL was investigated in MMP study. [6]



17% of the original dataset were relevant in this study. Methyl addition is the most favorable transformation (~25%)

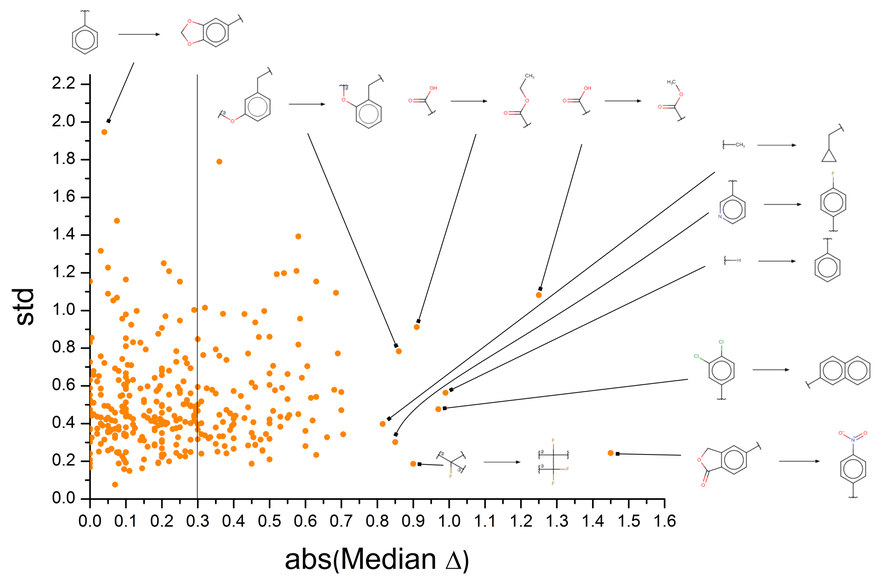


Fig 5. Standard deviation was plotted against median delta values to visualize impact and certainty.

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

As a result of our investigation and study on existing and available chemical and related information on hERG related data combined with ChemAxon's technology, a solution to create, filter, and suggest novel compounds for a given input structure within seconds is available.

Lead optimization support with hERG assistant - Marvin Live

Marvin Live

The extracted rules assists medicinal chemists to design out hERG blockade liabilities using MMP transformations in predictive mode. Results from superstructure search against the transformation dataset, sorted by the highest effect provide an array of modifications with expected change with regards to hERG. A proof of concept integration of MMP with Marvin Live compound design platform is shown on *Fig. 6*. MMP-based modification ideas relative to the compound on the canvas are visualized with the corresponding statistics, among other parameters characterizing the actual idea. [5]

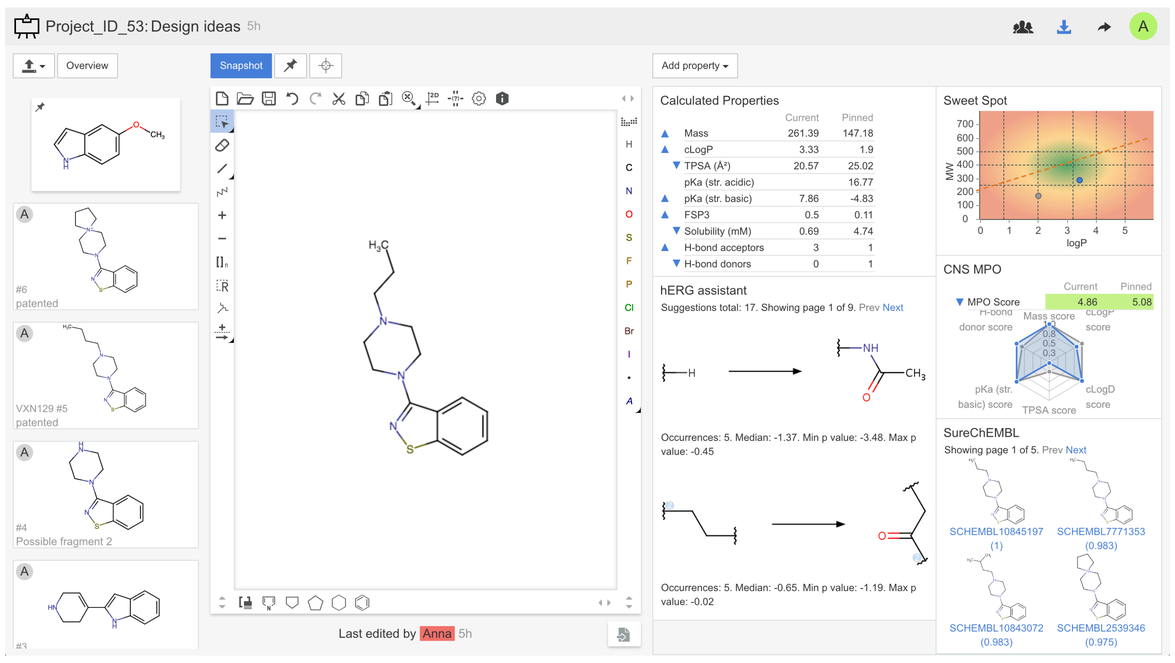


Fig 6. Characterization of benzothiazole compounds with med. chem predictions and freedom to operate search

Conclusion

Extracting information from available or in-house knowledge supported by experimental data is highly valuable. Incorporating such information into Computer-Aided Drug Design and optimization approaches using matched molecular pair analysis is a viable approach at medicinal chemists hand. MMP algorithms used in Marvin Live provides valuable input in designing novel compounds.

References

[1] C. Tyrchan, E. Evertsson, Comp. Struct. Biotech. J., 2017, 15, 86
[2] https://www.ebi.ac.uk/chembl/downloads
[3] J. Hussen, 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] www.chemaxon.com/products/marvin-live
[6] V.J Gillet et el, J. Chem. Inf. Model. 2010, 50, 1872

atomin@chemaxon.com

ChemAxon