University of Szeged
Doctoral School of Pharmaceutical Sciences

Educational Programme:	Pharmaceutical Chemistry and Drug Research
Programme director:	Prof. Dr. Ferenc Fülöp
Institute:	TargetEx Ltd.
Supervisors:	Dr. György Dormán Prof. Dr. Ferenc Fülöp

Krisztina Dobi

Selection and testing of protein target focused compound libraries using chemoinformatics methods integrated with MTS biological assay

Final examination committee:
Head: Dr. Zsolt Szakonyi, PhD
Members: Dr. Péter Krajcsi, DSc
Dr. Antal Péter, DSc

Reviewer committee:
Head: Dr. György Dombi, DSc
Reviewers: Dr. Gerda Szakonyi, PhD
Dr. Róbert Kiss, PhD
Members: Dr. Géza Tóth, DSc

1. Introduction and aims

One of the main driving forces of the modern drug discovery is to identify new chemical entities (structures) and to achieve a higher hit rate by integrating with contemporary chemoinformatics methods. To increase the hit rate is also economically important, because more hits are expected to be identified by creating focused libraries, while eliminating “junk” (non-active) molecules delivered by random screening, thereby the costs of purchasing and biological screening could be significantly reduced.

Two-dimensional (2D) similarity search is a widely used *in silico* technique in the pharma research selecting compounds from huge compound repositories based on the structural-similarity towards known, biologically active compounds. Even though this method is robust and fast it has several weak points.

The key concept of the ligand-based VS approaches is the Similarity Property Principle, which states that similar molecules should have similar biological properties.

The traditional fingerprint-based 2D similarity apporaches are able to find novel structures that in fact resemble to the parent, seed compounds, therefore, to achieve novelty and higher hit rates are critical questions. The similarity search uses 2D fingerprints, i.e., binary strings encoding the presence or absence of a substructure within the molecules. Applying simple 2D fingerprints is often the method of choice particularly when numerous reference compounds and multimillion compound databases are available.

Potentially active, target focused libraries can be obtained by determining the similarity to the biologically active reference compound for each molecule found in the database followed by ranking the molecules by similarity.

Focused library screening often results in a many fold increase in the hit rate compared to the random screening of commercial libraries. In many cases virtual screening methods and *in vitro* HTS were combined and found complementary to each other.

The 2D similarity search can be further refined with physico-chemical parameter filtering and diversity filtering. The empirical physico-chemical parameter ranges stand for drug-likeness. If most of the parameters of the drug candidate fall into the pre-defined range, the concerning molecule could be administered orally.

The major physico-chemical properties are Molecular weight, cLogP, H-bond donors, H-bond acceptors, rotatable bonds and Topological polar surface area. The first 4 parameters are included in the Lipinski’s Rule of 5, while the other 2 are noted as Veber rules.

During my work, two protein target focused compound libraries were generated and tested by *in vitro* biological screening:

1. The **phosphodiesterases** (PDEs) belong to a family of cyclic nucleotide degrading enzymes, which control the intracellular levels of cAMP and cGMP. There are 11 known PDE families, produced in the central nerveous system (PDE-1 - 11), with at least 21 subfamilies or subtypes which differs in structure, substrate specificity, body tissue dispersion, regulation by kinases, protein-protein interaction and inhibitor selectivity.

   PDE-4 isozymes play role in pathologies with inflammatory symptoms such as asthma, chronic obstructive pulmonary disease (COPD), inflammatory bowel disease, atopic dermatitis (AD), psoriasis and rheumatoid arthritis (RA). PDE-4 inhibitors have importance also in CNS, including antidepressant, and memory-enhancing effect mostly in long-term treatment.
2. **5-H$_{T6}$** is implicated in the pathogenesis of neurological and cognitive disorders including Alzheimer's disease, schizophrenia etc. as well as obesity˙ 5-Hydroxytryptamine (5-HT, serotonin) is a major neurotransmitter in the central (CNS) and in the peripheral nervous system, which mediates many effects through its interaction with a family of receptors called 5-HT receptors. The mediated processes are mood, cognition, perception, pain, feeding behavior, smooth muscle contractility and platelet aggregation. 5-H$_{T6}$ antagonists are able to increase acetylcholine levels, thereby these antagonists have therapeutic potential to improve normal and reduced memory.

Objectives of the PhD work:

1. In general, testing and evaluating various integrated chemical biology model systems, which have an aim to indentify the interacting common portion of the chemical and biological space by combining *in silic*o selection of potential biological active compounds (generating target focused libraries) from huge databases with in-house developed *in vitro* biological assays for evaluating the biological activity.

2. Develop and improve 2D similarity search methods using available softwares and outside partners. Key drivers: effectiveness and the novelty. Integrating the simple and robust 2D similarity search with other approaches (3D alignment, pharmacophore modelling and search) in order to improve the hit rate and the novelty.

3. Develop rapid and reproducible biological assays for evaluating the focused libraries. Assay development and validation.

4. Analyzing the obtained biologically active compounds and the structural evolution of the *in silico* search pathway in order to draw conclusions how successful was the selection as well as what should be modified or changed in future *in silico* selection campaignes.

5. In order to investigate and achieve the above general objectives two cases studies were selected and investigated:

5.1. In the first study, we analyzed the relationship between 2D / 3D similarity based on the results of a screening project previously performed on PDE-5 with the aim of drawing conclusions and using experience in a new project (generating a PDE-4 inhibitor focused library). The focused library generated by 2D / 3D similarity search was validated by in vitro biological screening developed during the work.

5.2. In the second study, a 5-HT6 antagonist focused library was generated and screened by integrating the 2D similarity search and pharmacophore model and then the library was evaluated by performing in vitro biological screening developed during the work.

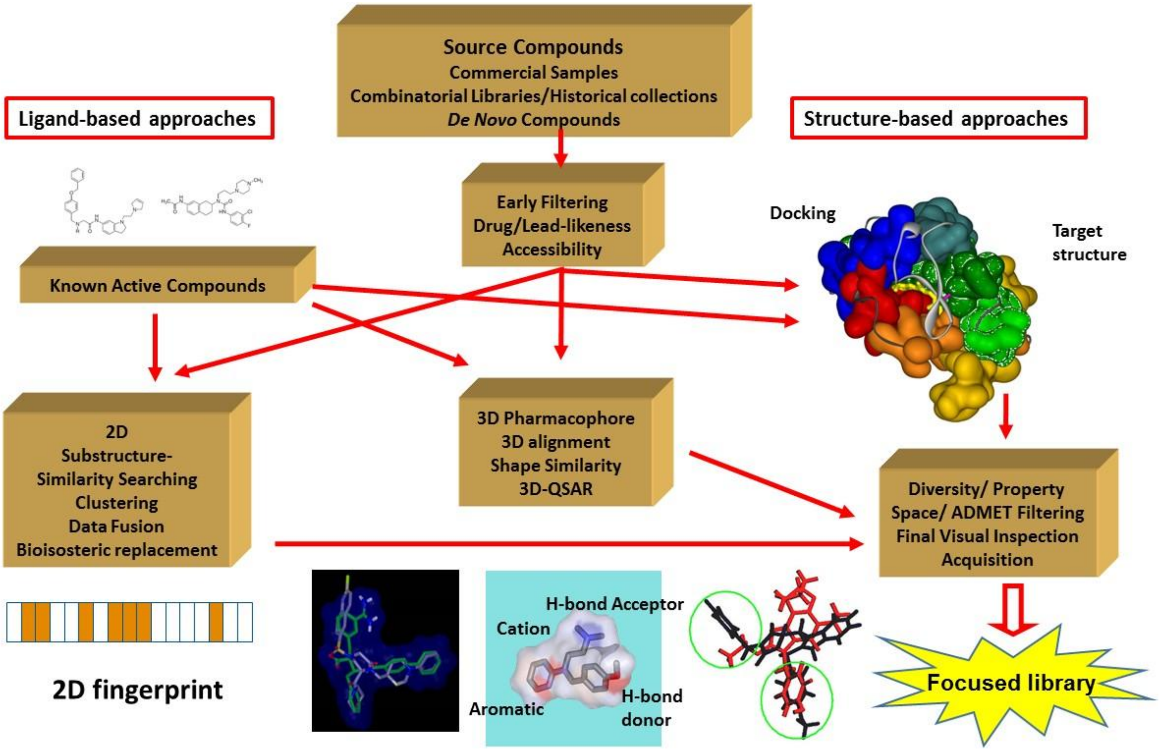

Figure 1. Overview of the Virtual Screening approaches

2. Applied methods

We applied InstJChem software (ChemAxon Ltd., Budapest, Hungary) for **2D similarity search**. InstJChem uses the Chemical Hashed Fingerprints for the 2D similarity search. Normally, in the initial similarity search phase a compound was defined as similar, if the Tanimoto coefficient was T2D ≥ 0.65 compared to any reference compounds, however, we applied higher Tanimoto similarity cut-off values (T2D ≥ 0.8) in second round (hit validation) studies.

The searchable drug-like chemical space, which is the major target of the similarity search, was composed by purchasing compounds: non-exclusive commercial libraries were available from the actual edition of the top vendor databases (~5 million compounds).

Biologically active chemical space is composed of known, biologically active reference (seed) compounds. Known PDE inhibitors and 5-HT$_6$ antagonists were collected from the available literature, PubChem and various commercial databases.

The compounds obtained in the 2D similarity search were preferably **clustered** into groups (scaffolds/chemotypes) based on the chemical architecture (substructures) using ChemAxon’s JKlustor program.

The 2D similarity search can be refined by using the calculated physico-chemical parameter ranges (’parameter space’) of the known, active compounds. The **physico-chemical parameters** (Mwt, LogP, H-bond donors/acceptors, rotatable bonds and topological polar surface area) were determined by the Calculation Suit of InstJChem (ChemAxon Ltd.).

The property or parameter space of the known active compounds can be defined as the calculated min. and max. values, which were focused to the central 90 % of the range. Such focused property space was used for **property-based filtering** of the 2D similarity search results.

We applied for **3D ligand-based similarity search** flexible alignment through the Match algorithm of the Screen3D software (Screen3D 5.1, ChemAxon Ltd. Budapest). In practice Match algorithm carries out calculations of minimum and maximum possible intramolecular distances between every atom-atom pair during the high-throughput conformational scan of each compound. Intramolecular distances are collected for the formation of distance range histograms. As a result of the above calculation the software generates a 3D conformation that is suitable for alignment between two structures providing a measure of the 3D similarity as a 3D Tanimoto coefficient.

We could further reduce the filtered compound selection to a reasonable and affordable library size, by **diversity selection** using Similarity Manager (CompuDrug Int., Sedona, AZ, USA).

3. Results and discussion

3.1. PDE-5

1. In the first case study we analyzed the 2D/3D similarity values of a previous PDE-5 screening campaign to draw conclusions about the correlation. We involved a PDE-5 inhibitor focused library (41 compounds) that were generated by 2D similarity based on the structure of 3 known active (reference) compounds and generated their 3D similarity values. We found that while the 2D selection criteria was T2D=0.60, thus, T2D values were above this value, the 3D Tanimoto scores (T3D) were only above 0.2. Since the T3D values for the 5 hits were above 0.3, we indicated this value on the plot emphasizing that many „2D similar" compounds fell into the 0.2-0.3 category. We could conclude that even though the compounds selected by 2D methods had similar molecular architectures (atomic connectivity), they are rather different in terms of shape and conformational flexibility. The lower T3D region suggests that such 3D „dissimilarity" represents much different binding characteristics.

2. Comparing the 3D similarity values with the biological activity of the same cluster (41 compounds), we found that the 5 identified hits has T3D values higher than 0.3, but there are lots of 2D similarity selected compounds that has T3D similarity values above 0.3 but they are not hits. We could call them as 'virtual false positives' (45% of all compounds measured). (*Note:* Hit compounds are defined if the inhibitory activity is > 57% at 10 μM concentration.)

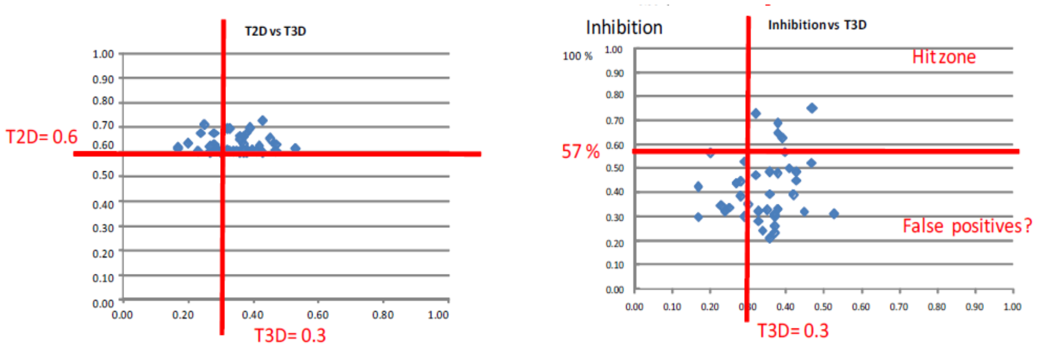

Figure 2 shows the scatter plot of correlating the T2D and T3D similarity scores of all the measuredcompounds (41, active and not active) derived from the three reference compounds.

3. After the first round selection and screening we had the conclusion that T3D = 0.3 might have been a useful cut-off value even if it might result in false positives as well.

In a further study we selected 104 compounds based on the structure of the 5 first round screening hits for the 2nd round screening by 2D methods. Now we set the selection criteria much higher, to T2D ≥ 0.8. We calculated again the 3D similarities for all the 2D selected compounds. While the T2D cut-off value was 0.8 and the corresponding T3D values were between 0.35 and 0.95. The 2D/3D correlation is shown in **Figure 3**. There is a certain trend that with increasing 2D values the 3D values are also increasing but in a lower extent. Interestingly, the majority of the compounds have T3D values above 0.6.

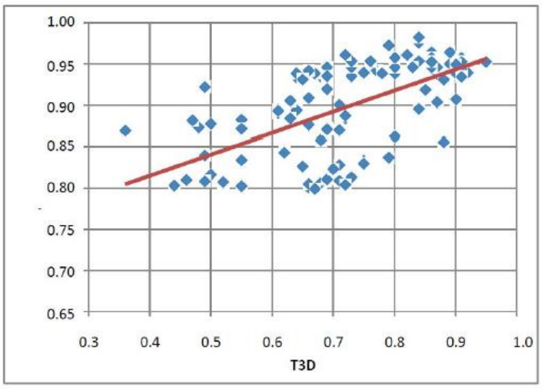

Figure 3. Correlation between T2D and T3D values of the compounds (104) selected for the second round screening

4. We analyzed separately two series: the second round hit compounds (13 out of 20) derived from two PDE-5 first round hits (#3-PDE-5; #2-PDE-5) and their parent seed (reference) compounds (#18-PDE-5; #44-PDE-5).

Significantly lower 3D similarity values were found in the case of compounds derived from Series #3-PDE-5 (between 0.47 and 0.69), than from Series #2-PDE-5 (between 0.7 and 0.91). We attempted to combine the T2D/T3D similarities into a single fusion score and the T2D + T3D values would range between 1.28 (0.8 + 0.47) and 1.71 (0.8 + 0.91) in the second round screening. We got the conclusion that T3D values are much more sensitive measures than their T2D counterparts, and they reflect important structural features (e.g., conformational flexibility) that strongly depend on the structure.

5. We concluded that combination of the 2D similarity search with 3D similarity measures might be useful and applicable approach in ligand-based virtual screening and we further refined this strategy in the selection a PDE-4 focused library.

3.2. PDE-4B

6. The typical 2D similarity search was carried out using 44 known PDE-4 inhibitors as seed compounds and we obtained 4833 compounds, when the Tanimoto cut-off value was set as 0.65. We calculated the property (parameter) space for the 44 known PDE-4 inhibitors and applied this property space as a filter, which allowed to reduce the number of the virtual hits to 1764. Finally, we selected 200 compounds using simple diversity selection. After visual checking, the i*n vitro* measurements were done with 105 compounds at 10 μM concentration, and we experienced inhibition (≥47%) in case of nine compounds.

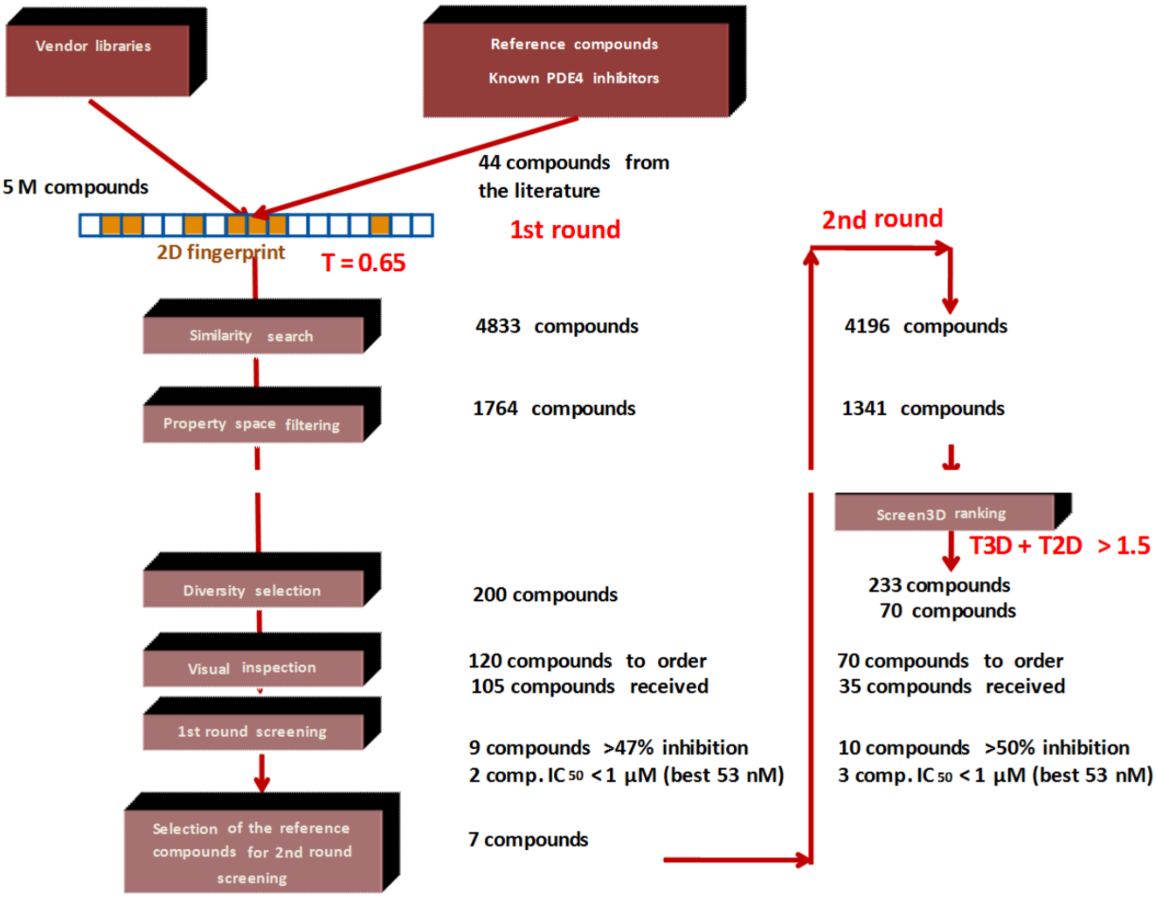

Figure 4. *In silico* selection scheme of the PDE-4B focused library

7. For the second round library selection (hit validation) the seed compounds were the seven first round hits. After a standard 2D similarity search and property space filtering 1341 compounds were obtained (T2D ≥ 0.65). In the next step we calculated the 3D similarity values for 1341 compounds towards their corresponding first round hits.

We applied a 2D/3D fusion score (T2D + T3D ≥ 1.5) for the 1341 compound library filtering, which resulted in 233 compounds. Finally, the library was reduced to 70 by diversity selection and this compound set was ordered and 35 compounds received. The contribution of T2D is gradually increasing but fluctuating between 0.65 and 1, while T3D is decreasing from 1 to 0.6 among the compounds that have a fusion score of 1.5. *In vitro* screening of the small focused library resulted in 10 hits (IC_{50}= 0.053–3.2 μM). The hit rate of the 2nd round screening was 28.5% (≥50% inhibition, at 10 μM), while 10% hit rate was calculated if only those compounds were considered that had an IC_{50} values ≤ 2 μM (seven compounds).

8. We also analyzed the T2D/T3D correlation of the hits. Such correlation and fusion score analysis applied on the 10 second round hits revealed that six out of the 10 hits have a fusion score around 1.5 which is close to the cut-off value (Figure 5). The T3D values were apparently lower in most of the cases. Apparently, the fusion score prefers this scenario and the selection leads to relatively close analogues (keeping the same chemotype) and helping to exclude false negatives, where the T2D is high but the T3D score is low.

These finding suggests that the fusion score approach is valuable in the hit validation stage particularly within the preferred chemotype. Of course, this finding raises the question if the fusion score would have been decreased to 1.3, which would allow the involvement of more compounds with lower T2D and T3D values. This would extend the selection to approx. 300 compounds from 233.

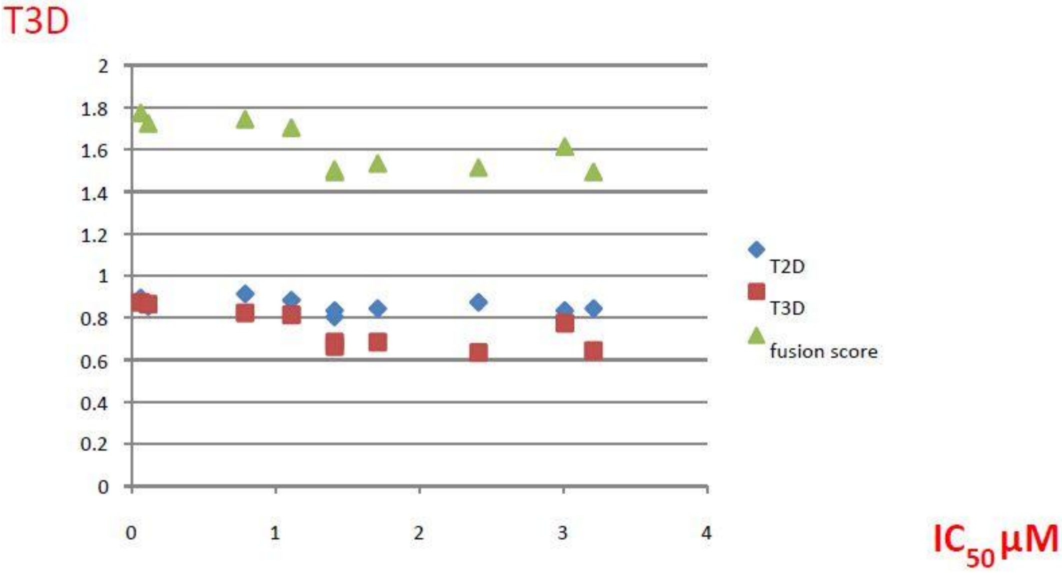

Figure 5. T2D/T3D and fusion score analysis of the 10 second round hits

9. We also found that most of the hit compounds showed some selectivity towards PDE-4B and in some cases the selectivity was surprisingly high (50–80 fold) over PDE-4D. Some of the hits inhibited PDE-5 at low concentration which could be useful in the therapy, while PDE-10 inhibition could be avoided, since that target is more implicated in CNS disorders.

3.3. 5-HT$_6$

10. For focused library generation first we collected 49 known 5-HT$_6$ antagonists („seeds") from the available literature. At the initial 2D similarity search we set the similarity threshold (T2D $\geq$ 0.65 Tanimoto coefficient) following the „reference property space" (Table 1.) and applying diversity selection.

	MolWeight	LogP	TPSA	H bond donor	H bond acceptor	Rotatable bond
P95	476.0	4.72	96.60	3	7	5
P5	330.4	1.10	40.10	0	3	2

Table 1. The property space ranges for the known 5-HT$_6$ antagonists applied as seed compounds

After property space filtering we obtained 872 compounds, followed by diversity selection and visual inspection. Finally we acquired 91 compounds for *in vitro* screening (Figure 6).
The hit rate of the first screening round was 13%, 12 compounds showed inhibition.

11. We used for the second round focused library selection a pharmacophore search algorithm. For pharmacophore model building 49 structurally different seed compounds and 11 first round hit molecules were involved. We divided the molecules into three distinct clusters, which allowed to create a pharmacophore hypothesis with five sites (two hydrogen bond acceptors (A), one hydrophobic group (H) or hydrogen bond donor group (D), and two aromatic rings (R) for these three datasets (cluster 1: AAHRR type; cluster 2: AADRR type; and cluster 3: AAHRR type).

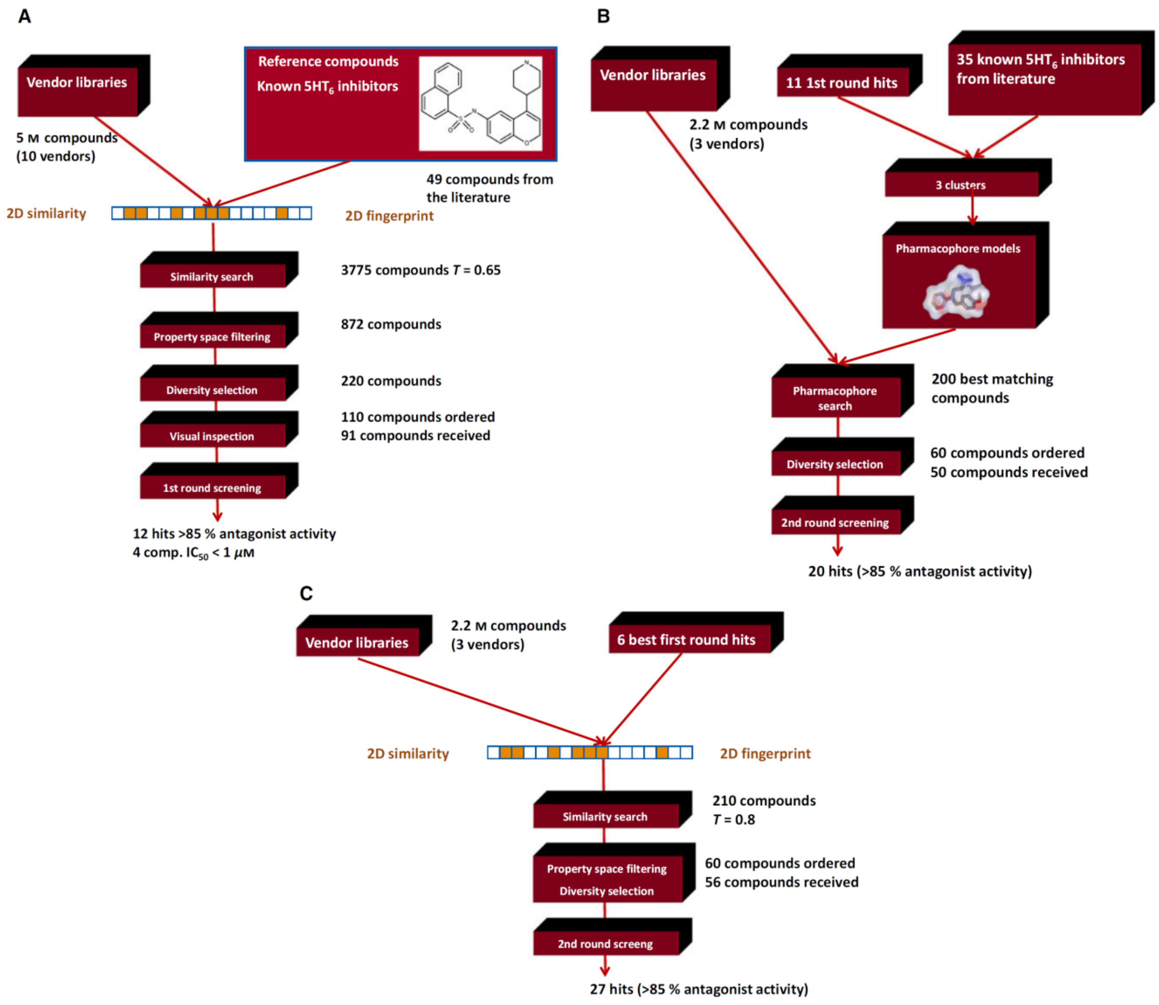

Figure 6. Multistage *in silico/in vitro* screening cascade to select and validate 5-HT$_6$ target-focused libraries in two rounds including similarity and pharmacophore search.

12. In our experience pharmacophore search was found as a particularly powerful tool for selecting compounds from large size compound libraries: 20 compounds out of 50 were found active (>85% antagonis activity, hit rate = 40%). The average 2D similarity to the closest similar first round hit compounds was 0.586, which has already anticipated novelty, and really, 3 novel chemotypes were found compared with the first round hits.

13. In terms of the hit rate, the purely 2D similarity search (T2D=0.8) around the 6 first round hits (IC$_{50}$ ≤ 1 μM) gave somewhat better results (27 compounds out of 56 were found active, hit rate= 51 %). The average 2D similarity values to the original seeds were 0.704 (at that level there is less chance for novelty).

14. We searched the best hit compounds came from the 2nd round screening in the PubChem DB for novelty and biological activity. Those compounds were chosen mostly, where T2D ≤ 0.6 to the original seed (coming from the pharmacophore search) because those compounds had a better chance for novelty. We identified compound #1-5-HT_6. by pharmacophore screening (T= 0.755, IC_{50}= 45 nM), although this compound is not completely novel, and we can find it in the DB but no biological activity is given. Therefore, this compound represents a novel 5-HT_6 antagonist.

The calculated developability scores (Ligand efficiency indices: LE, LLE, LLE_{AT}) in all cases of the molecules met the the required values (LE is preferred if >0.3; LLE if >5, and LLE_{AT} if >0.3).

Conclusion

Finally, we can say that in the two parts of the dissertation we have discovered various new biologically active compounds and achieved a higher hit ratio with our integrated chemoinformatics methods (in case of PDE-4B 28.5%, in 5-HT_6 study 40% and 51%).

Publication list

Papers related to the thesis

1. **Dobi K**, Hajdú I,Flachner B, Fabó G, Szaszkó M, Bognár M, Magyar Cs, Simon I, Szisz D, Lőrincz Zs, Cseh S, Dormán Gy. **(2014)**
Combination of 2D/3D ligand-based similarity search in rapid virtual screening from multimillion compound repositories. Selection and biological evaluation of potential PDE-4 and PDE-5 inhibitors.
Molecules, 19(6):7008-39 ; **IF: 2.095**

2. **Dobi K**, Flachner B, Pukáncsik M, Máthé E, Bognár M, Szaszkó M, Magyar Cs, Hajdú I, Lőrincz Zs, Simon I, Fülöp F, Cseh S, Dormán Gy. **(2015)**
Combination of pharmacophore matching, 2D similarity search and *in vitro* biological assays in the selection of potential 5-HT$_6$ antagonists from large commercial repositories
Chem Biol Drug Des;86(4):864-80 **IF:2.396**

Other publications

3. Flachner B, Hajdú I, **Dobi K**, Lőrincz Zs, Cseh S, Dormán Gy, **(2013)**
Melanin koncentráló hormon receptor-1 (MCHR1) antagonista fókuszált könyvtár kiválasztása és in vitro biológiai szűrése AequosCreen esszével
Acta Pharmaceutica Hungarica 83:(3) pp. 71-87.

4. Hajdú I, Flachner B, Bognár M, Végh B, **Dobi K**, Lőrincz Zs, Lázár J, Cseh S, Takács L, Kurucz I, **(2014)**
Monoclonal antibody proteomics: Use of antibody mimotope displaying phages and the relevantsynthetic peptides for mAb scouting
Immunology Letters. 160(2):172-7; **IF:2,37**

5. Flachner B, Tömöri T, Hajdú I, **Dobi K**, Lőrincz Zs, Cseh S, Dormán Gy. **(2014)**
Rapid in silico selection of an MCHR1 antagonists' focused library from multi-million compounds' repositories. Biological evaluation
Medicinal Chemistry Research, Vol. 23, Issue 3, 1234-1247; **IF:1,61**

6. Szaszkó M, Hajdú I, Flachner B, **Dobi K**, Magyar C, Simon I, Lőrincz Z, Kapui Z, Pázmány T, Cseh S, Dormán G

Identification of potential glutaminyl cyclase inhibitors from lead-like libraries by in silico and in vitro fragment-based screening

Molecular Diversity 21:(1) pp. 175-186. **(2017)** **IF:1,752**

Posters:

Beáta Flachner, Krisztina Dobi, János Varga, Zsolt Lőrincz, Sándor Cseh:

Epitope mapping of mAbs recognizing protein markers of obesity: a phage display study

Cecon II, Budapest 2009.

Krisztina Dobi, Mária Pukáncsik, Beáta Flachner, István Hajdú, , Zsolt Lőrincz, Sándor Cseh and György Dormán:

DEVELOPMENT OF MT ASSAYS TO DISCOVER POTENTIAL 5-HT6 ANTAGONISTS FROM FOCUSED LIBRARIES

Hungarian Molecular Life Sciences, Siófok 2013.