PROTEIN TARGET PREDICTION

Lazaros Mavridis

University
of
St Andrews

University of St Andrews

Overview

- Introduction
  - World Anti-Doping Agency (WADA)
  - Substances Prohibited in Sports
  - Protein Target Prediction
- Methodology
  - Circular Fingerprints (CFP)
  - Machine Learning (Parzen – Rosenblatt)
  - Databases (ChEMBL)
  - Refinement (Rule base – Clustering)
- Results
  - WADA explicitly prohibited compounds

WORLD ANTI-DOPING AGENCY
play true

ChEMBL

World Anti Doping Agency

The World Anti-Doping Agency's (WADA) mission is to lead a collaborative worldwide campaign for doping-free sport.

- WADA's funding is based on a unique hybrid private-public model: 50% Olympic Movement 50 % Governments of the world.
- WADA's governing bodies, namely Foundation Board and Executive Committee, are composed in equal parts by representatives from the sport movement and governments of the world.
- WADA is the funding body for this project.

Substances Prohibited in Sports

- WADA publishes and maintains a prohibited list world anti-doping code, which is updated every 6 months
- Substances are split into three main categories:

Substances prohibited at all times (in and out of competition)

S0. Non-Approved substances
S1. Anabolic Agents
S2. Peptide hormones, Growth Factors and Related Substances
S3. Beta-2 Agonists
S4. Hormone Antagonists and Modulators
S5. Diuretics and Other Masking Agents

Substances prohibited in competition

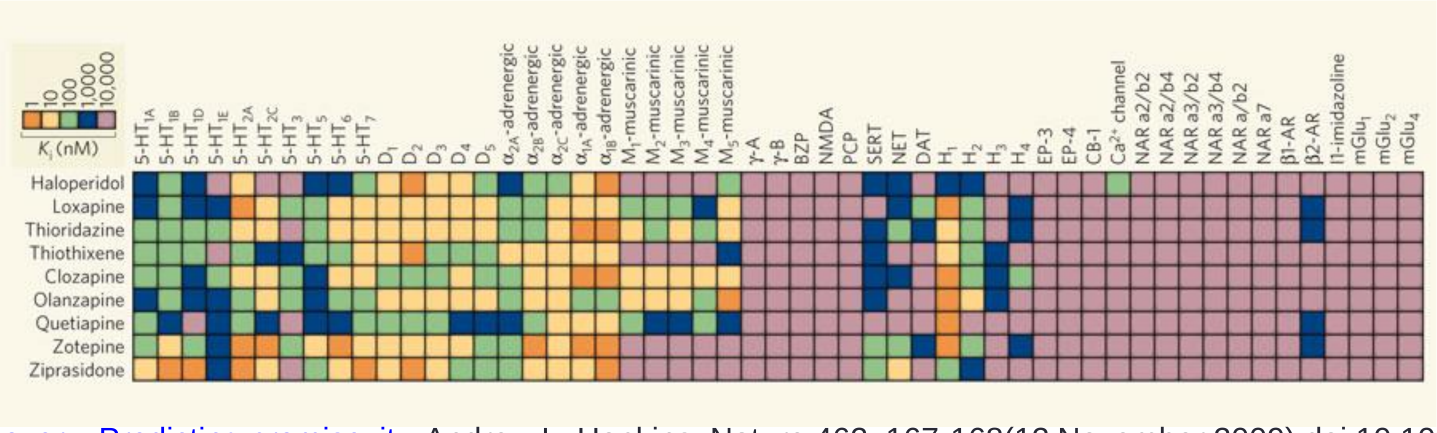
S6. Stimulants
S7. Narcotics
S8. Cannabinoids
S9. Glucocorticosteroids

Substances prohibited in particular sports

P1. Alcohol with a violation threshold of 0.10 g/L. (Archery, Karate etc)
P2. Beta-Blockers prohibited *In-Competition* only (Bridge, Curling, Darts, Wrestling, Archery etc.)

Protein Target Prediction

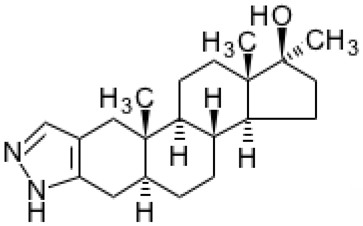
- Given a specific substance, is it possible to predict computationally **all** possible biological interactions of the substance?
- Very important for
  - *In silico* screening (time and money efficient)
  - off-target prediction (side effects)
- Can be used for identifying substances with performance-enhancing potential

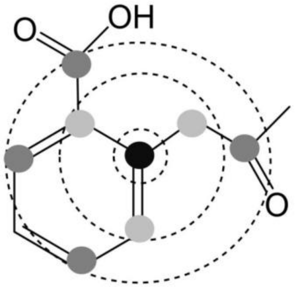


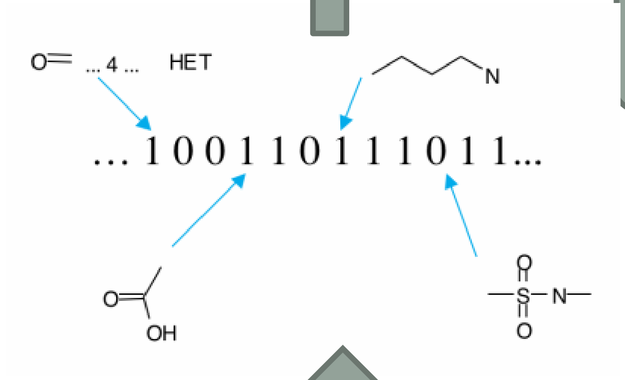
Drug discovery: Predicting promiscuity, Andrew L. Hopkins, Nature 462, 167-168(12 November 2009),doi:10.1038/462167a

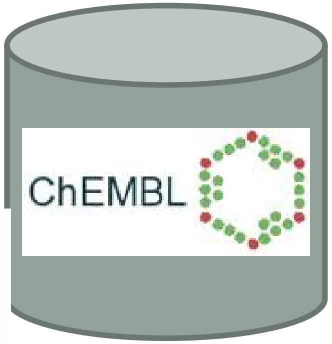
Methodology

Stanozolol





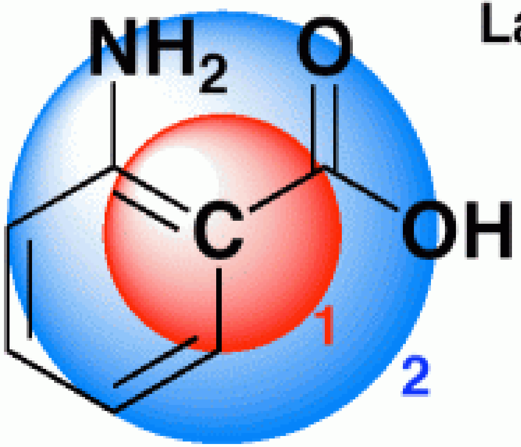




Rank	Class
1	Anabolic Agents
2	Vitamin D
3	Glucocorticoids

Circular Fingerprints (CFP)

- Atom-environment fingerprint of a compound
- 2D based descriptor
- Ideally suited for machine learning techniques
- Used for all pairwise comparisons of compounds

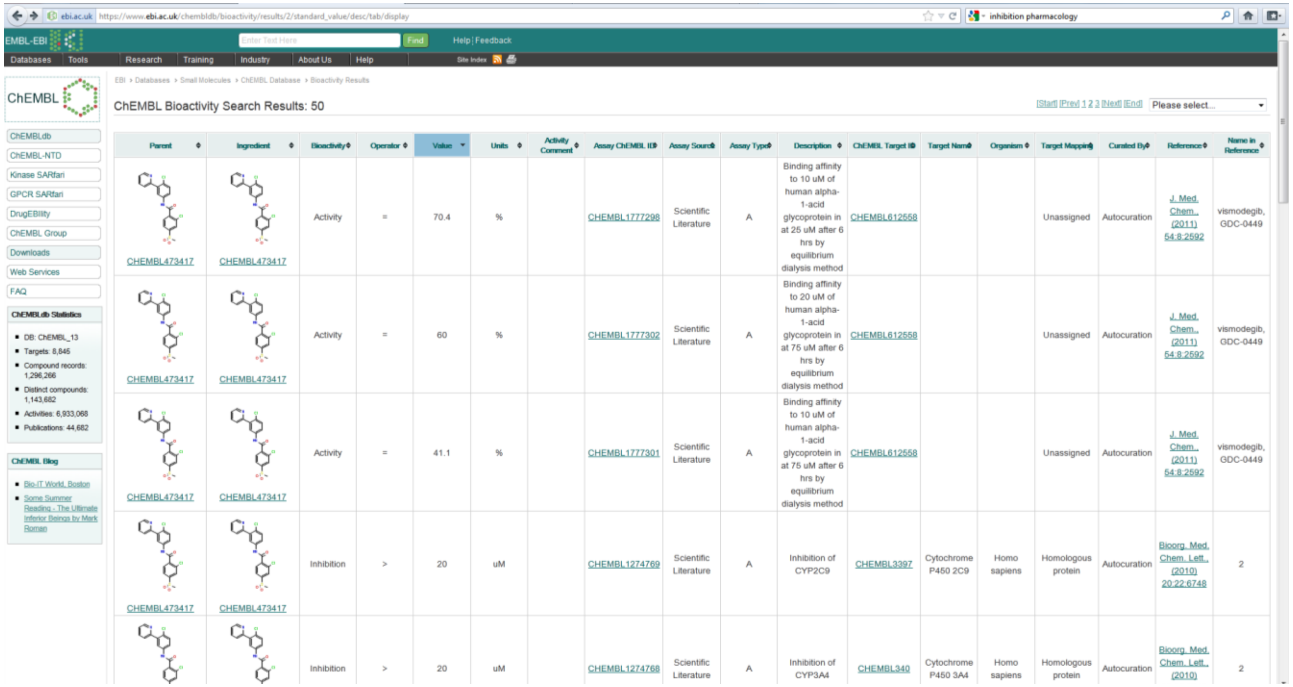


Layer:	0	1	2
	C.ar (sp^2)	C.ar (sp^2)	C.ar (sp^2)
		C.ar (sp^2)	C.ar (sp^2)
		C (sp^2)	N (sp^3)
			O (sp^2)
			O (sp^3)

[atom type];[layer]-[frequency]-[neighbour type];

ChEMBL

- DB: ChEMBL_13
- Targets: 8,845
- Compound records: 1,296,266
- Distinct compounds: 1,143,682
- Activities: 6,933,068
- Publications: 44,682



ChEMBLdb Statistics

- DB: ChEMBL_13
- Targets: 8,845
- Compound records: 1,296,266
- Distinct compounds: 1,143,682
- Activities: 6,933,068
- Publications: 44,682

ChEMBL Blog

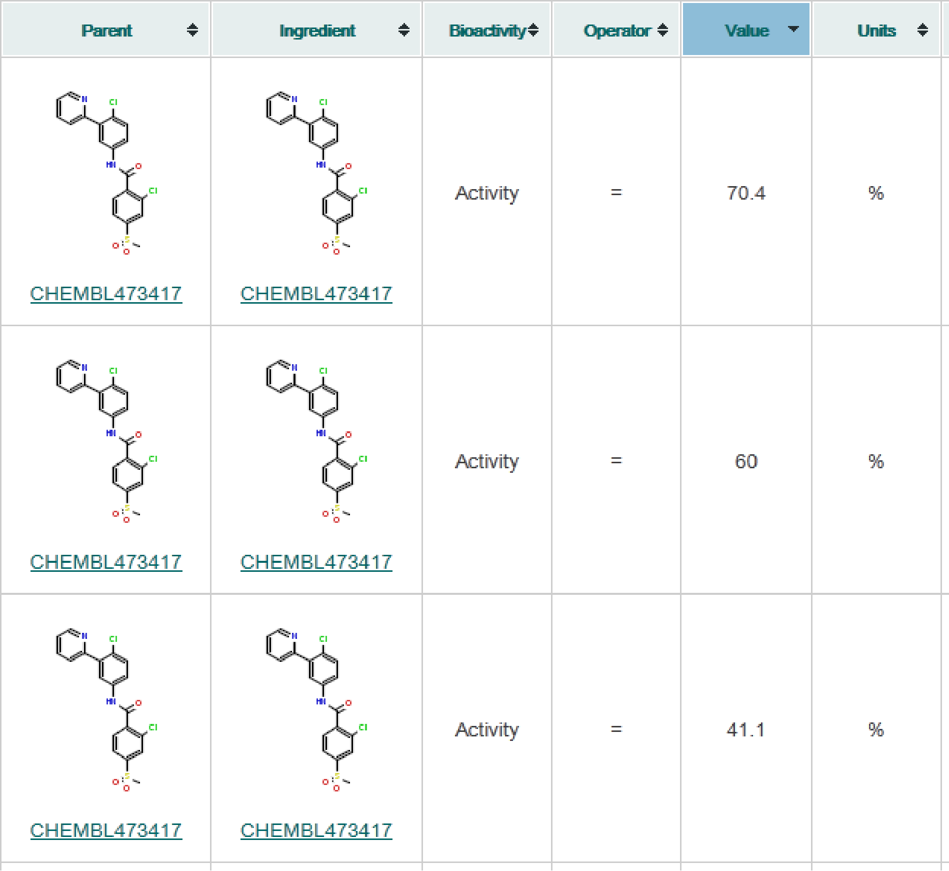
- Bio-IT World, Boston
- Some Summer Reading - The Ultimate Inferior Beings by Mark Roman

ChEMBL Bioactivity Search Results: 50

Parent	Ingredient	Bioactivity	Operator	Value	Units	Activity Comment	Assay ChEMBL ID	Assay Source	Assay Type	Description	ChEMBL Target ID	Target Name	Organism	Target Mapping	Curated By	Reference	Name in Reference
CHEMBL473417	CHEMBL473417	Activity	=	70.4	%		CHEMBL1777298	Scientific Literature	A	Binding affinity to 10 uM of human alpha-1-acid glycoprotein in at 25 uM after 6 hrs by equilibrium dialysis method	CHEMBL612558			Unassigned	Autocuration	J. Med. Chem., (2011) 54:8:2592	vismodegib, GDC-0449
CHEMBL473417	CHEMBL473417	Activity	=	60	%		CHEMBL1777302	Scientific Literature	A	Binding affinity to 20 uM of human alpha-1-acid glycoprotein in at 75 uM after 6 hrs by equilibrium dialysis method	CHEMBL612558			Unassigned	Autocuration	J. Med. Chem., (2011) 54:8:2592	vismodegib, GDC-0449
CHEMBL473417	CHEMBL473417	Activity	=	41.1	%		CHEMBL1777301	Scientific Literature	A	Binding affinity to 10 uM of human alpha-1-acid glycoprotein in at 75 uM after 6 hrs by equilibrium dialysis method	CHEMBL612558			Unassigned	Autocuration	J. Med. Chem., (2011) 54:8:2592	vismodegib, GDC-0449
CHEMBL473417	CHEMBL473417	Inhibition	>	20	uM		CHEMBL1274769	Scientific Literature	A	Inhibition of CYP2C9	CHEMBL3397	Cytochrome P450 2C9	Homo sapiens	Homologous protein	Autocuration	Bioorg. Med. Chem. Lett., (2010) 20:22:6748	2
		Inhibition	>	20	uM		CHEMBL1274768	Scientific Literature	A	Inhibition of CYP3A4	CHEMBL340	Cytochrome P450 3A4	Homo sapiens	Homologous protein	Autocuration	Bioorg. Med. Chem. Lett., (2010)	2

ChEMBL - Activities

- Each compound has experimental data for a number of targets
- Activity data based on IC50, EC50, K_i, K_d *etc*.
- Some activities just labelled “inactive” or “active”
- Each compound can have more than one record for a given target



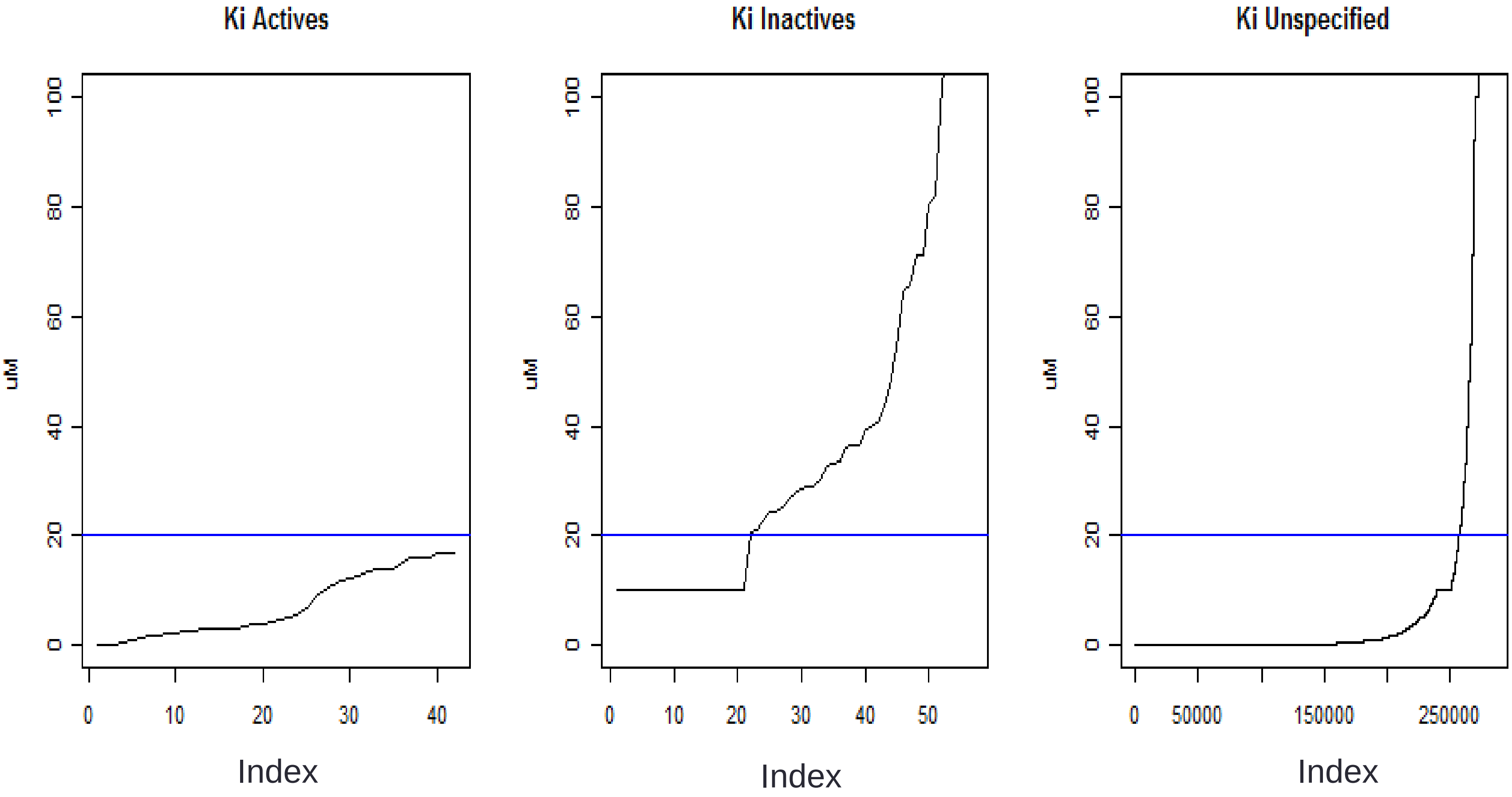
Parent	Ingredient	Bioactivity	Operator	Value	Units
CHEMBL473417	CHEMBL473417	Activity	=	70.4	%
CHEMBL473417	CHEMBL473417	Activity	=	60	%
CHEMBL473417	CHEMBL473417	Activity	=	41.1	%

Filtering the CheMBL Families

- Each of the 8,845 targets has a number of compounds assigned to them
- Not all compounds have actual data on the target or are active
- We performed a filtering for each of the families according to a number of rules
- The rules were decided after visual inspection of the most important bioactivity types

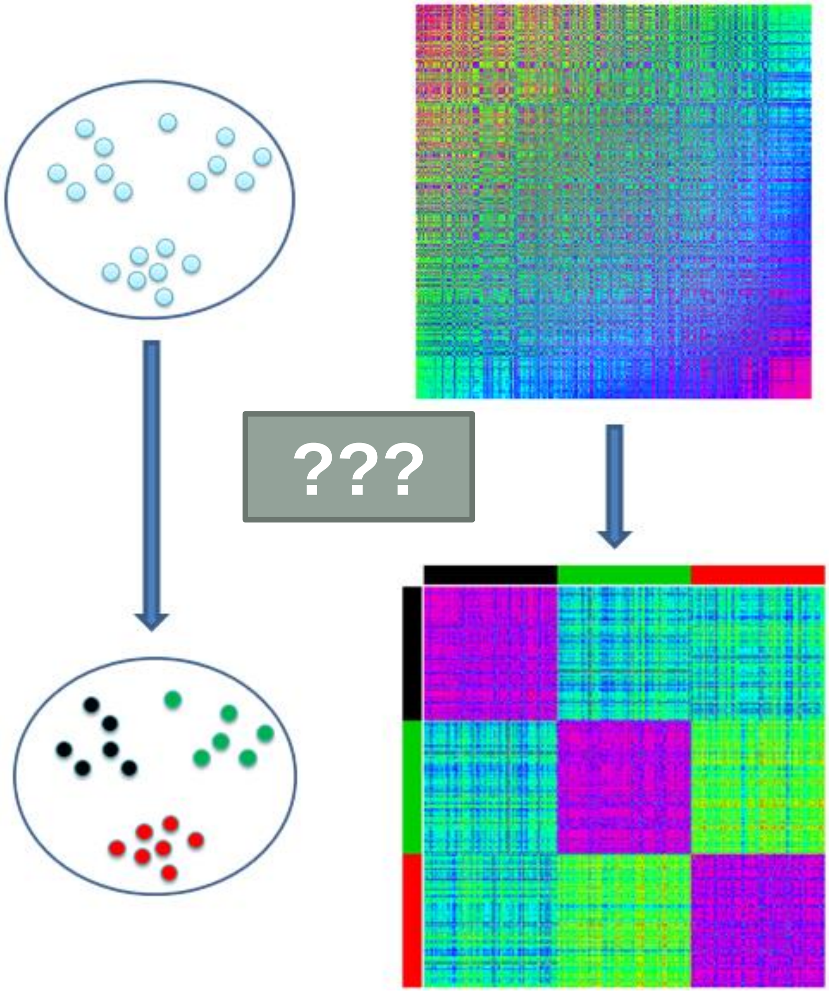
- **Rules**
  - **IC50**
    - **≤50000nM active & >50000nM inactive**
  - **K_i**
    - **<20000nM active & ≥20000nM inactive**
  - **K_d**
    - **≤ 10000nM active & >10000nM inactive**
  - **EC50**
    - **≤ 40000nM active & >40000nM inactive**
  - **ED50**
    - **≤ 10000nM active & >10000nM inactive**
  - **Potency**
    - **≤ 10000nM active & >10000nM inactive**
  - **Activity**
    - **≥40% active & <40% inactive**
  - **Inhibition**
    - **≥45% active & <45% inactive**

Example case of K_i



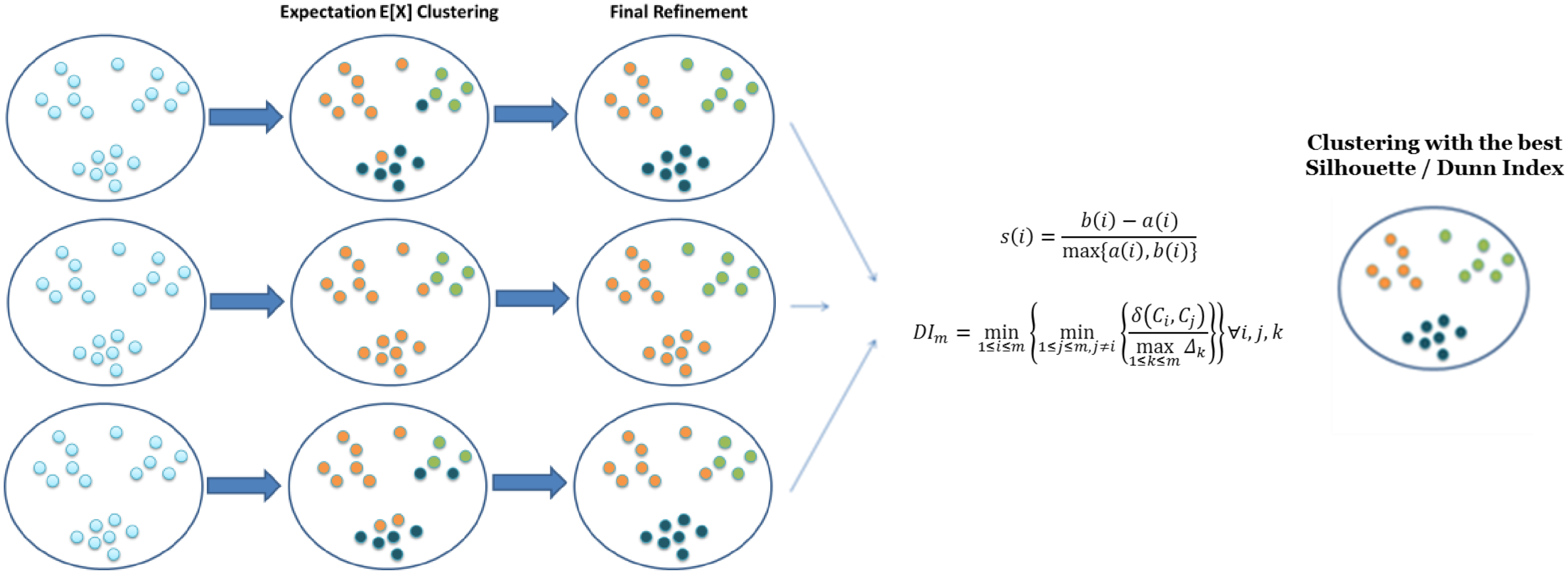
Refined Families

- Although the filtered families consist of compounds that have significant experimental activities against the relevant targets
- There are many targets that have distinct groups of ligands with different scaffolds.
- This may be because there is more than one binding site, or because different scaffolds can fit the same site.
- Splitting such a family into smaller groups based on ligand structure will allow us to identify the different sets of ligands



Refined Families - PFClust

We selected the PFClust algorithm because it is a parameter free clustering algorithm and does not require any kind of parameter tuning.

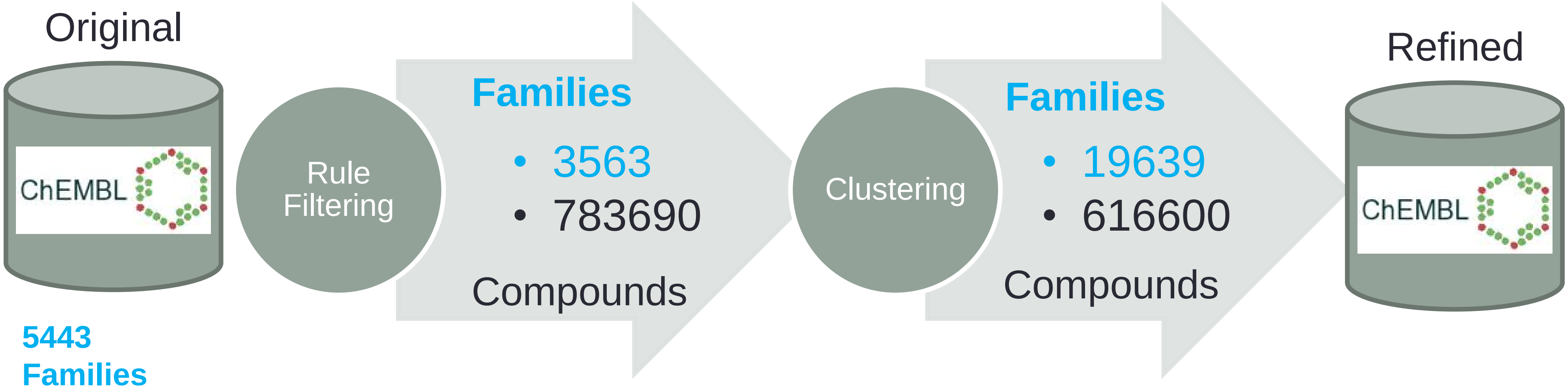

Expectation E[X] Clustering
Final Refinement
Clustering with the best Silhouette / Dunn Index

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$DI_m = \min_{1 \le i \le m} \left\{ \min_{1 \le j \le m, j \ne i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta_k} \right\} \right\} \forall i, j, k$$

PFClust : A novel parameter free clustering algorithm. Mavridis L, Nath N, Mitchell JBO. *BMC Bioinformatics 2013,* **14**:213.

Database Refinement



Original

ChEMBL

Rule Filtering

Families

- 3563
- 783690

Compounds

Clustering

Families

- 19639
- 616600

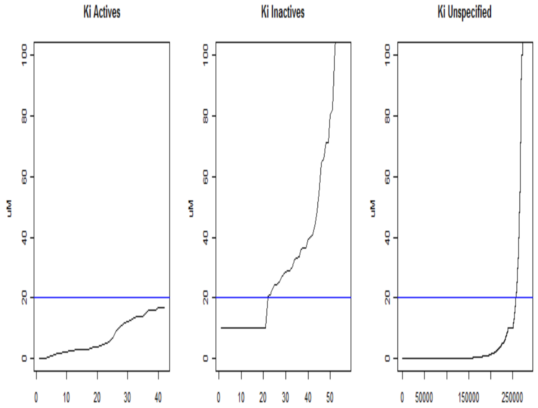
Compounds

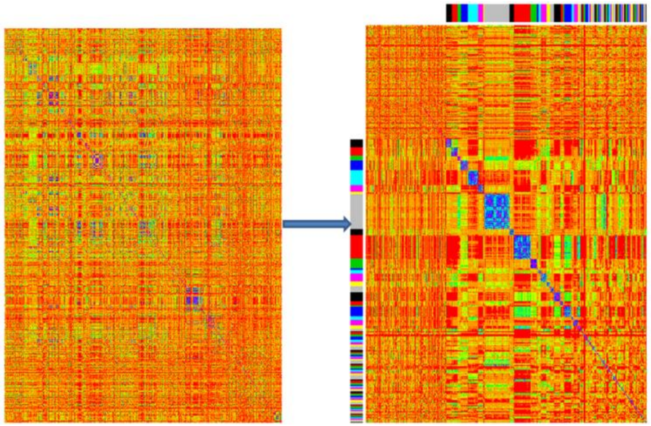
Refined

ChEMBL

5443 Families

1366460 Compounds





Predicting the protein targets for athletic performance-enhancing substances. Mavridis L, Mitchell JBO. *J Cheminformatics 2013,* **5**:31.

Database Searches

- Each database is split into groups according to annotated targets and activity data when available
- Each compound can be a member of more than one family
- For each query we would like to measure our confidence that query x_i is a member of a given family ω as $f(x_i, \omega)$
- What is the best estimate of this function

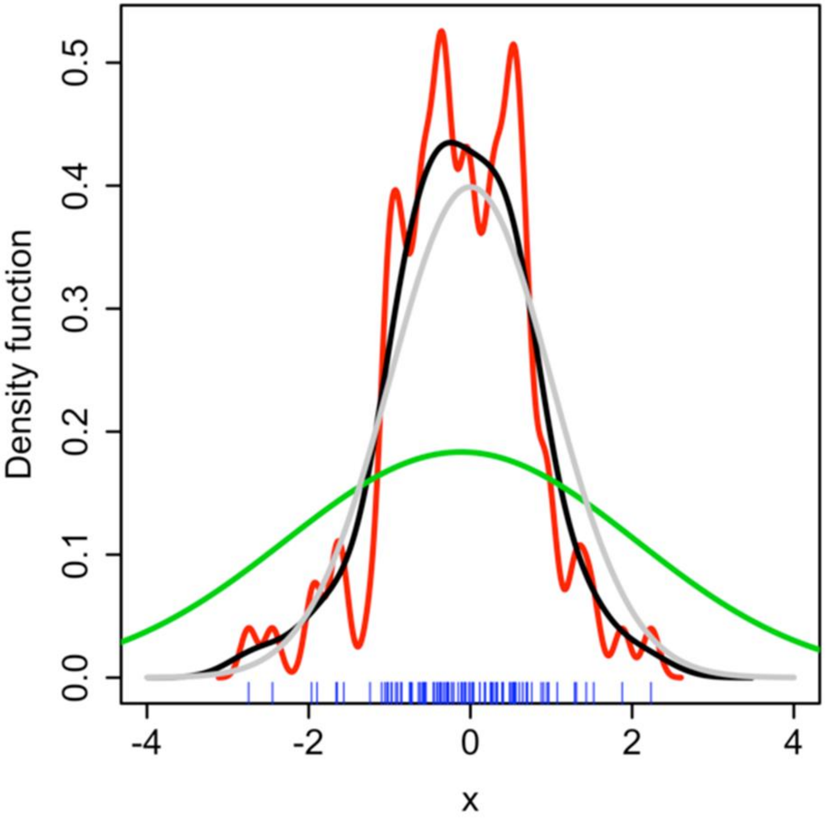
$$f(x_i, \omega) = ???$$

Machine Learning (Parzen–Rosenblatt)

- Kernel density estimation
- Appropriate for Multi-Labelling problems
- A non-parametric way of estimating the probability density function of a random variable {X}

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x - x_i)$$

where *n* is the number of samples and *k_h()* is the kernel.

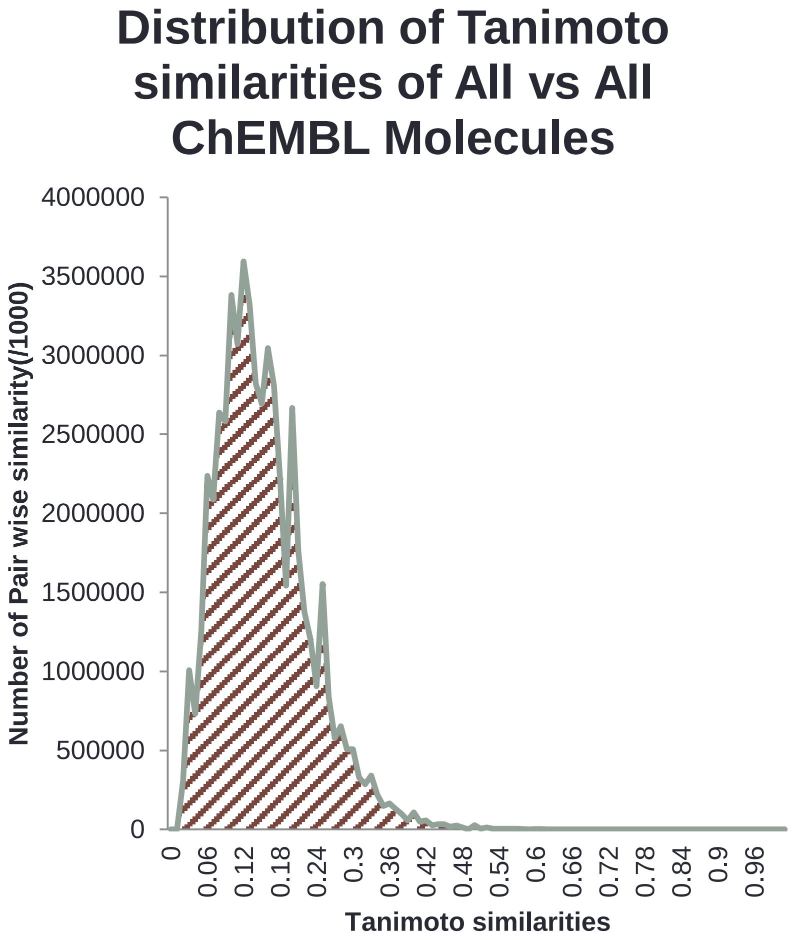


Kernel Density Estimation

- Comparison of molecules using a Tanimoto similarity score

$$f(A,B) = \frac{A.B}{|A|^2 + |B|^2 - A.B}$$

where A and B are the binary fingerprints of two molecules



Kernel Density Estimation

- We calculated the cumulative probability density (CDF) function of the Tanimoto scores
- We selected a Gaussian distribution as our kernel

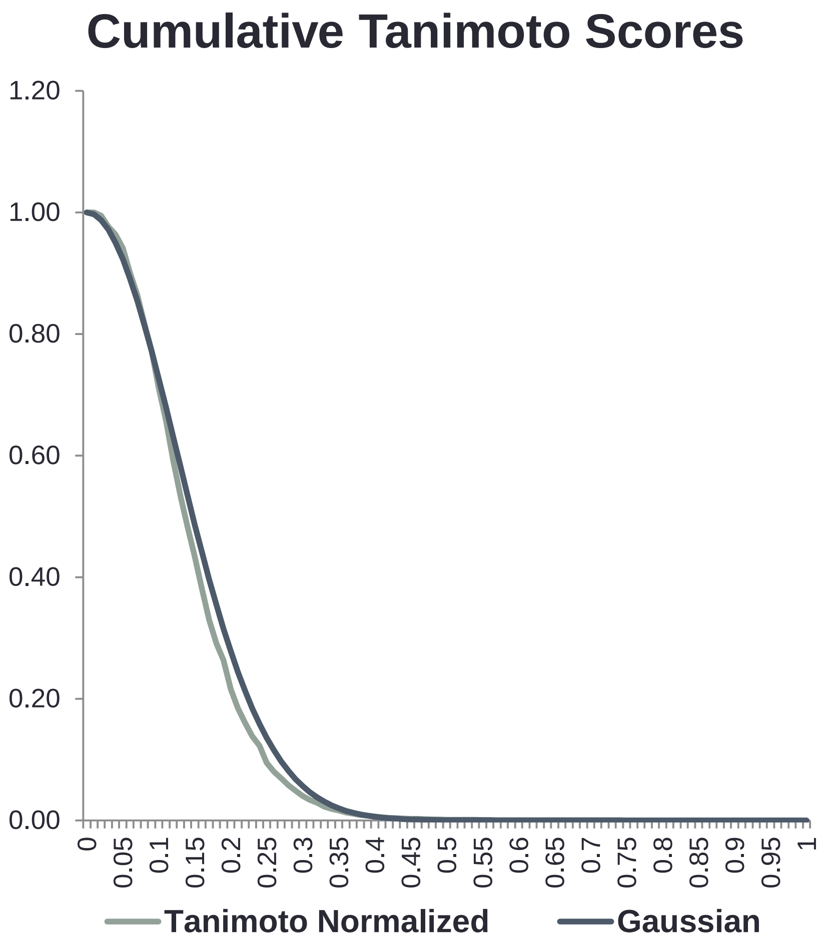
$$p(X > x) = p(X > t(x_i, x_j)) = e^{-\frac{t(x_i,x_j)^2}{2h^2}}$$

where *h=0.125* is a smoothing factor

- Hence we can calculate $f(x_i, \omega)$ as:

$$f(x_i, \omega) = \frac{1}{N_\omega} \sum_{j=1}^{N_\omega} p(X > t(x_i, \omega_{x_j}))$$

where N_ω is the number of molecules in family and t_i is the Tanimoto score of *x* with the *i*-th member of family *ω*



Database Refinement - Validation

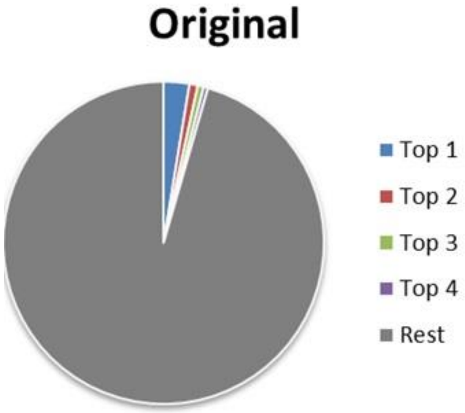
- Monte Carlo Cross-Validation

- The three versions of the database were examined (Original, Filtered and Refined)

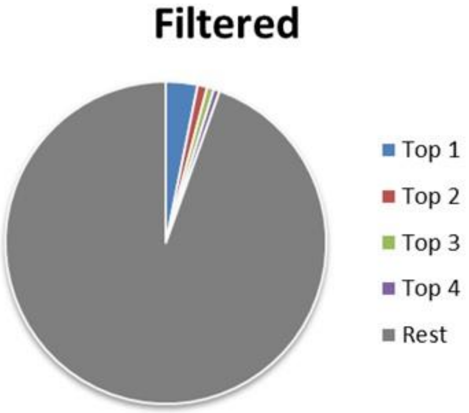
- 10% of each family were randomly removed and used as queries

- If the top prediction was the family that the query was a member of, a TP would be counted; if not, a FP

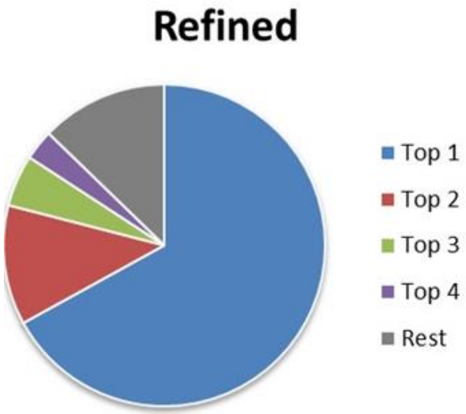
- Average Matthews Correlation Coefficient (MCC)
  - Original : 0.02
  - Filtered : 0.03
  - **Refined : 0.66**



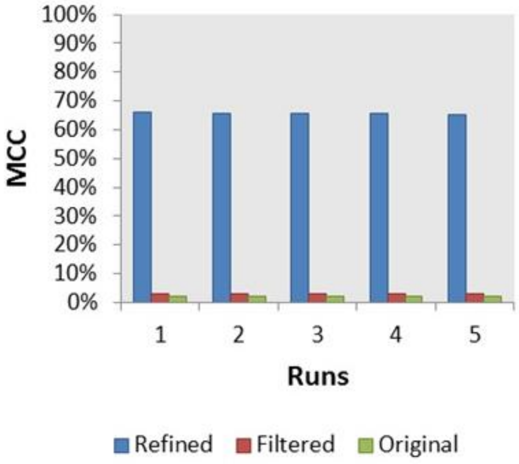
2.58% (6.61%)



3.18%(7.21%)



66.98% (87.25%)



Top Hit (Top four)

P2 – Beta Blockers

20 explicitly prohibited compounds

Every compound, except timolol and levobunolol, gave a strong prediction (PR-Score) for at least one family

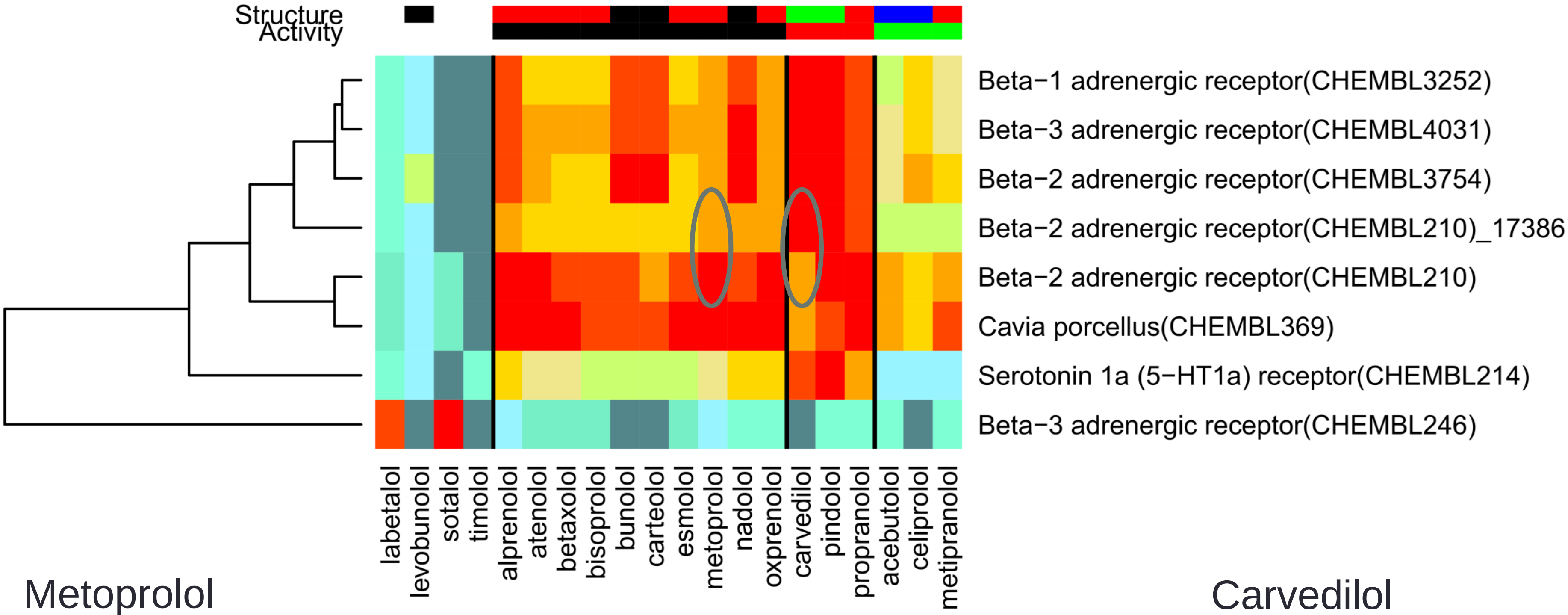
Good experimental validation

We see that the majority of the families are Beta-1,2 & 3 adrenergic receptor ligands, as expected.

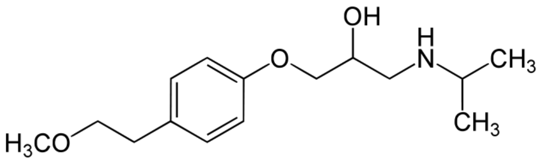
Other families also generate some interesting results, such as the serotonin 1a receptor, indicated to make off-target interactions with pindolol

Compound	Target	PR-Score	E-Value
	P2-Beta Blockers		
Alprenolol *(266195)*	*Cavia Porceullus (369)*	*0.039*	*LogB/F = −0.158*
Carvedilol *(723)*	*β-1 adrenergic receptor (3252)*	*0.032*	*Ki = 0.81 nM*
	β-2 adrenergic receptor (210)	*0.044*	*Ki = 0.166 nM*
	β-2 adrenergic receptor (3754)	*0.047*	*Prediction*
	β-3 adrenergic receptor (4031)	*0.036*	*Prediction*
Pindolol *(500)*	*β-1 adrenergic receptor (3252)*	*0.017*	*Ki = 1 nM*
	β-2 adrenergic receptor (210)	*0.015*	*Ki = 0.4 nM*
	β-2 adrenergic receptor (3754)	*0.026*	*Inhibition = 84%*
	β-3 adrenergic receptor (4031)	*0.018*	*Ki = 1 nM*
	Serotonin 1a (5-HT1a (214)	*0.026*	*Ki = 24 nM*
Propranolol *(27)*	*β-2 adrenergic receptor (210)*	*0.003*	*IC50 = 12 nM*
Sotalol *(471)*	*β-3 adrenergic receptor (246)*	*0.009*	*IC50 = 7200 nM*

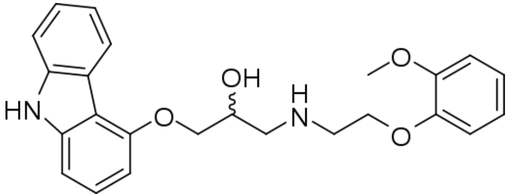
WADA – P2 Beta Blockers



Metoprolol



Carvedilol



S8 - Cannabinoids

10 explicitly prohibited compounds

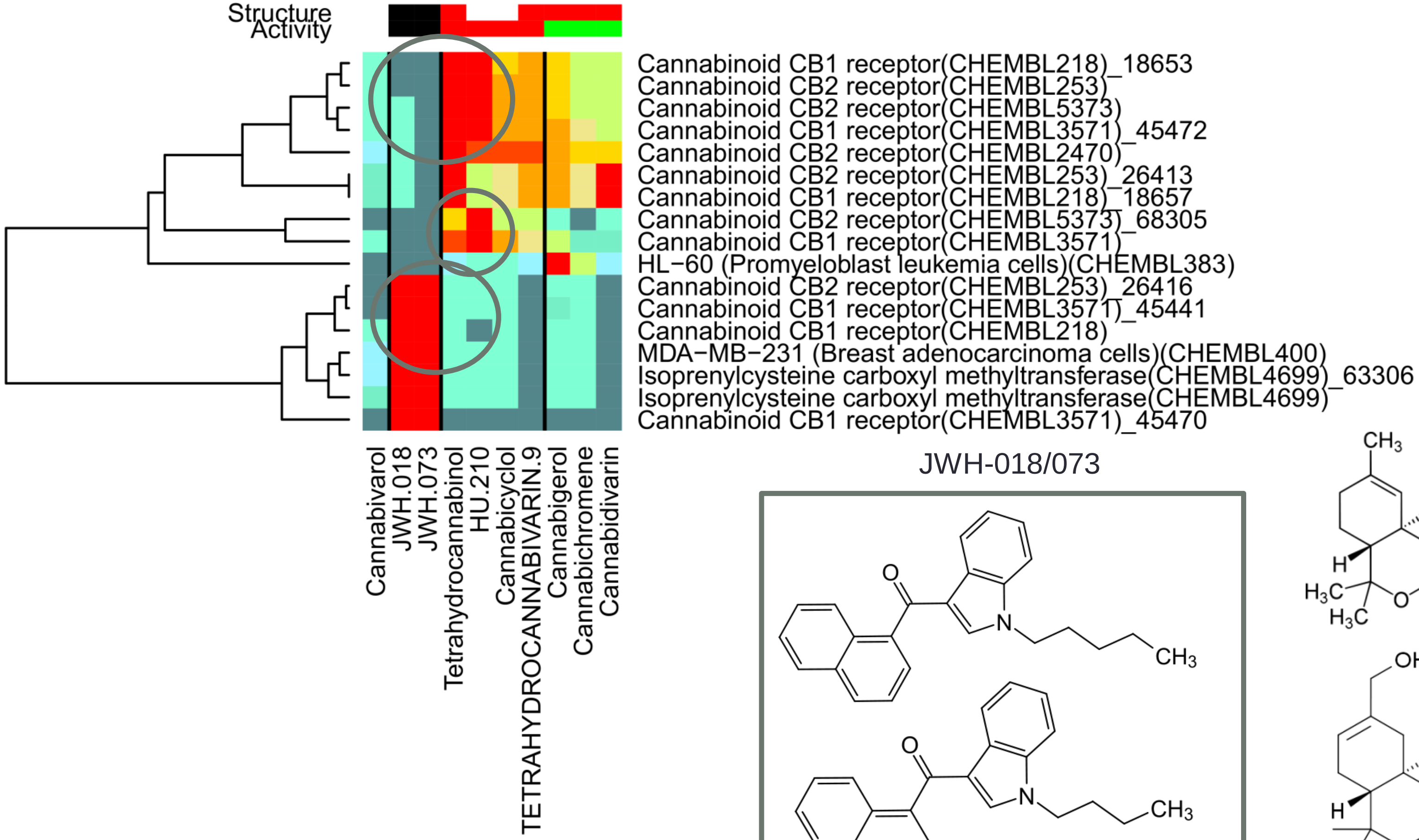
17 refined families of which 13 are cannabinoid CB1/2 receptors

All compounds show strong predicted affinity to at least one cannabinoid receptor, except cannabivarol

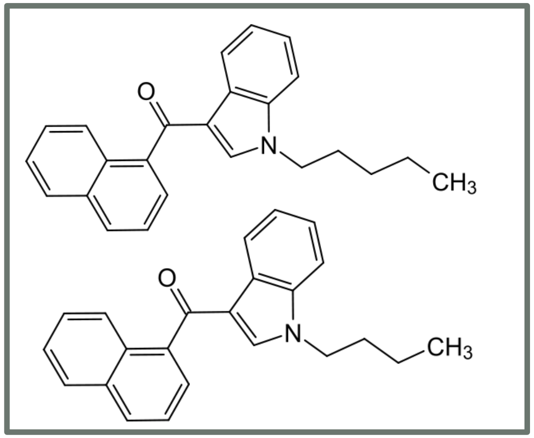
Excellent agreement between PR-scores and experimental results

Compound	Target	PR-Score	E-Value
	S8-Cannabinoids		
Cannabidivarin (−)	*Cannabinoid CB1 receptor (218)*	*0.037*	*Prediction*
	Cannabinoid CB2 receptor (253)	*0.037*	*Prediction*
Cannabigerol (497318)	*HL-60 (383)*	*0.047*	*Prediction*
HU-210 (70625)	*Cannabinoid CB1 receptor (3571)*	*0.035*	*Ki = 0.82 nM*[a]
	Cannabinoid CB2 receptor (5373)	*0.029*	*Prediction*
JWH-018 (561013)	*Cannabinoid CB1 receptor (218)*	*0.002*	*pKi = 8.7*
	Cannabinoid CB1 receptor (3571)	*0.015*	*pKi = 8.045*
	Cannabinoid CB2 receptor (253)	*0.009*	*pKi = 8.2*
	Isoprenylcysteine carboxyl methyltransferase (4699)	*0.031*	*Prediction*
	MDA-MB-231 (400)	*0.030*	*Prediction*
JWH-073 (−)	*Cannabinoid CB1 receptor (218)*	*0.002*	*Prediction*
	Cannabinoid CB1 receptor (3571)	*0.025*	*Prediction*
Tetrahydrocannabinol (465)	*Cannabinoid CB1 receptor (218)*	*0.037*	*Ki = 2.9 nM*
	Cannabinoid CB1 receptor (3571)	*0.037*	*Ki = 37 nM*
	Cannabinoid CB2 receptor (2470)	*0.034*	*Ki = 20 nM*
	Cannabinoid CB2 receptor (253)	*0.033*	*Ki = 3.3 nM*
	Cannabinoid CB2 receptor (5373)	*0.049*	*Ki = 9.2 nM*

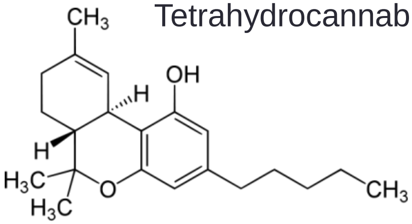
WADA – S8 Cannabinoids



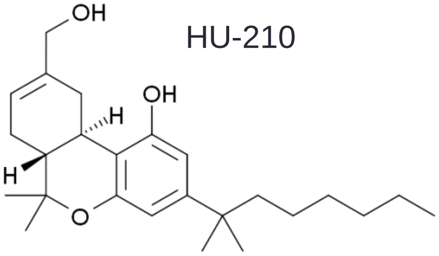
JWH-018/073



Tetrahydrocannabinol



HU-210



Discussion

- As for any method, the success of our approach depends on the quality of the underlying data that are available.

- Our methodology tries to address the problem that, for each molecule that could be synthesised and tested, only a small fraction of its activities against different targets have been assayed.

- For ChEMBL families that are not well populated, or for protein targets which too few compounds are assayed against, we cannot make predictions since we do not have the required data. Hence we cannot produce any predictions for a number of the compounds that are already in the WADA prohibited list.

Discussion (cont.)

- Our current methodology has proved that it **enhances** the predictive power of the CFP representations, and that the filtering and refinement of ChEMBL families **enriches** our results.

- However, the **portability** of our target prediction approach is as important as the quality of the results for the WADA prohibited compounds.

- This workflow can easily be used with different molecular representation techniques, new sets of rules, and with a different clustering algorithm (with due consideration of the stopping criterion); **hence it represents a truly portable methodology.**

Conclusions

- Automated data-curation of the ChEMBL families greatly increases the precision of our protein target prediction technique.
- Our validations show an encouraging correspondence with independent experimental results, with 87.25% having the parent refined family among the top four hits.
- Across the seven WADA classes considered, we find a combination of expected and unexpected protein targets for their constituent molecules.
- Analysis of the literature, however, demonstrates that many of the non-obvious targets have biochemically or clinically validated connections with the expected bioactivities.

Acknowledgments

John Mitchell research group - http://chemistry.st-andrews.ac.uk/staff/jbom/group/

Dr John Mitchell

Dr Luna
De Ferrari

James
McDonagh

Rosanna
Alderson

Neetika
Nath

Ava Sih-Yu
Chen

This work has been funded by the WORLD ANTI-DOPING AGENCY play true