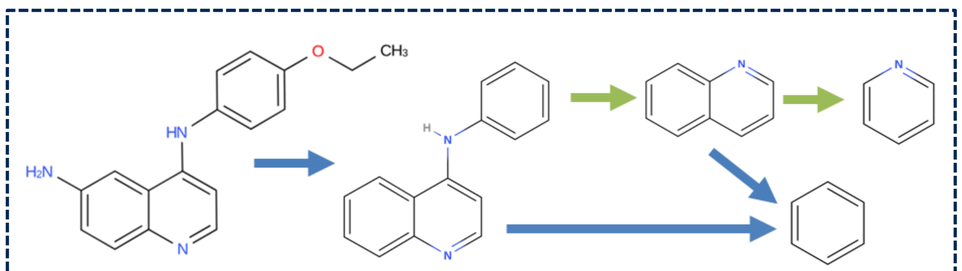
Classification and Analysis of Privileged Scaffolds in Protein Families

Oliver Scott , Xiaotong Zhang, AW Edith Chan

Wolfson Institute for Biomedical Research, University College London, Cruciform Building, Gower St, London WC1E 6BT, UK

- A '***privileged scaffold***' (PS), as defined by Evans in 1988, is a simple structural subunit common to compound structures that have high affinity against a diverse set of targets
- Since then this concept has been interpreted fairly liberally by chemists using different methods and definitions to derive collections of privileged scaffolds
- Another commonly used term, target-family privileged scaffolds (TFPS) defines a scaffold which is selective towards a specific protein family
- Results based on abundance of a scaffold within active compounds could be an oversimplification of biological interaction
- In this study, we analyze bioactivity from five protein-target (PT) super-families using a Compound Set Enrichment approach applied to sub-family scaffold networks
- PS is defined after scaffolds enrichment assessment based on their aggregated bioactivity

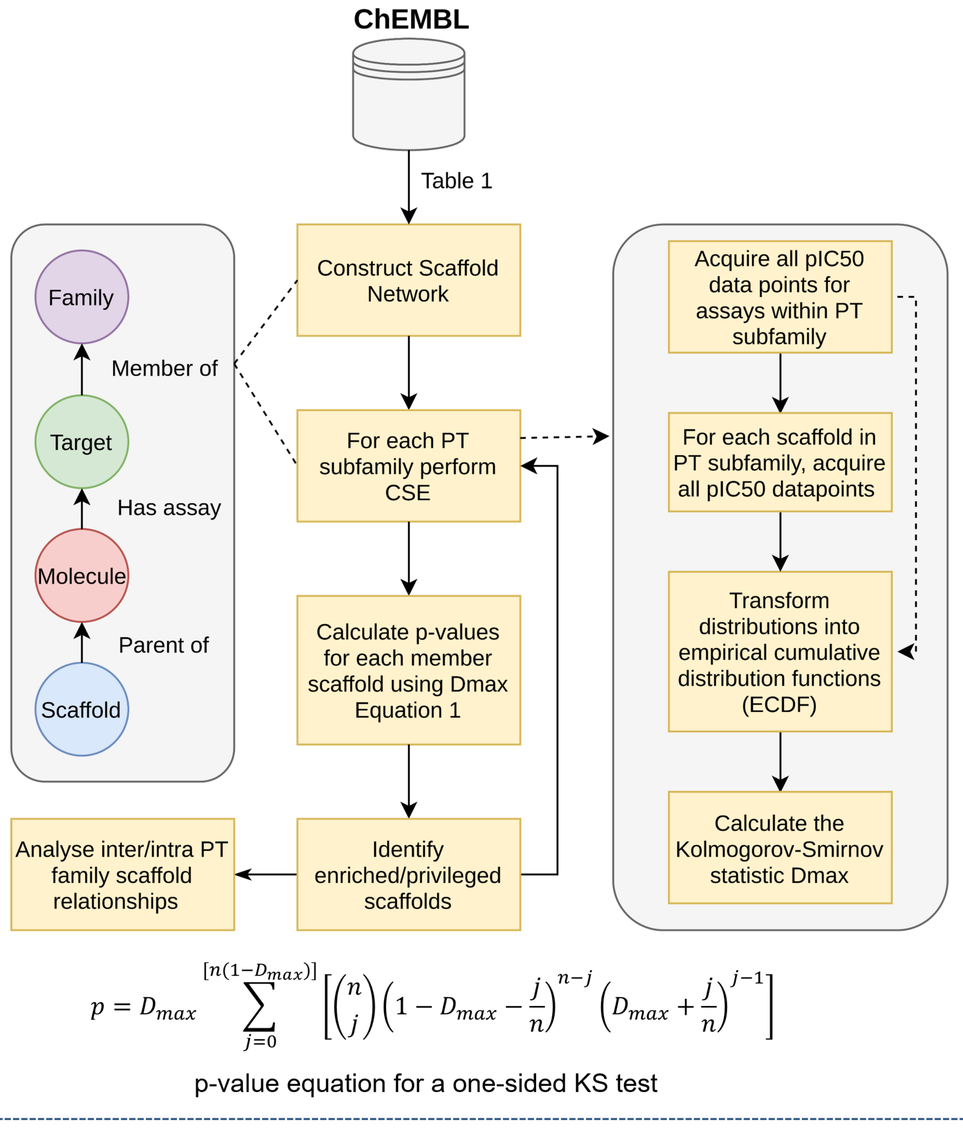
Definitions



Scaffold Network and Scaffold Tree representations

- **Scaffold Tree (ST)**
  - A hierarchical classification of scaffolds, obtained by the iterative removal of rings
  - A chemically intuitive set of prioritization rules determines the next ring to be removed from the molecule
- **Scaffold Network (SN)**
  - A scaffold classification consisting of all possible scaffolds within a molecule, constructed through recursive removal of rings
- **Compound Set Enrichment (CSE)**
  - An approach introduced by Varin et al. for identifying active series using SNs and STs coupled with non-parametric statistics.
  - The method identifies scaffolds statistically enriched for a biological measurement within primary screening data.

Methods & Workflow



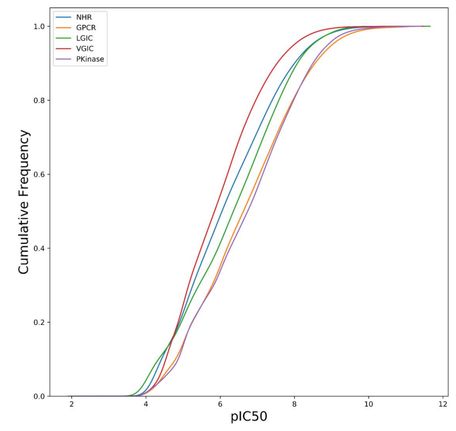
$$p = D_{max} \sum_{j=0}^{[n(1-D_{max})]} \left[\binom{n}{j} \left(1 - D_{max} - \frac{j}{n}\right)^{n-j} \left(D_{max} + \frac{j}{n}\right)^{j-1} \right]$$

p-value equation for a one-sided KS test

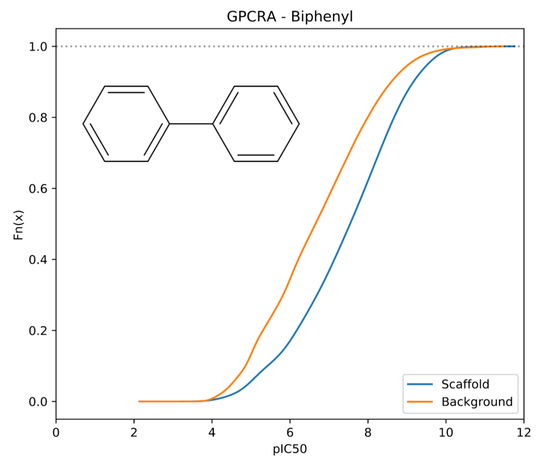
- All software was implemented in ***Python*** 3.6 using the open source cheminformatics software ***RDKit*** and graph analysis library ***NetworkX***
- Chemical structure data, pIC50 values and target family classifications were extracted from ***ChEMBL24***

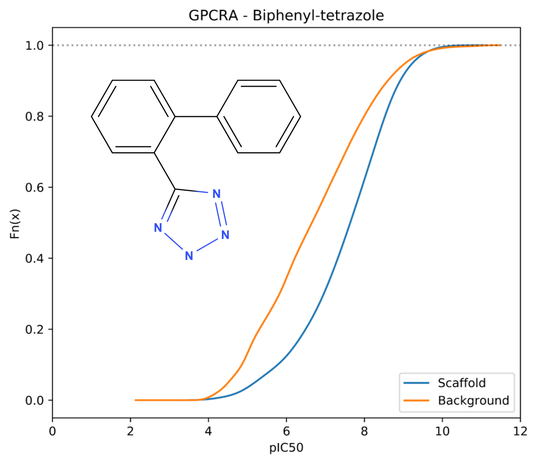
Results and Discussion

Family	# Subfamily (level 2)	# Molecules	# pIC50	# Scaffolds	# Privileged (<0.01)
P. Kinase	7	95,767	180,422	71,583	1859
GPCR	3	62,979	105,368	63,933	988
VGIC	5	23,208	31,685	22,576	418
NHR	6	12,184	20,417	11,200	184
LGIC	10	9,740	16,384	6,095	119



- The activity profile shows the cumulative distribution of pIC50 in the five protein super-families
- Activity profiles show that kinase and GPCR are the easiest targets to design high potency ligands

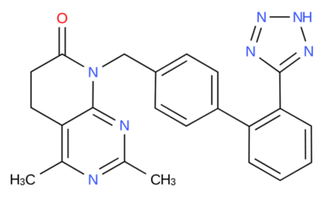




Family	Subfamily	# Targets	# pIC50	Log(p-val)
GPCR	GPCRA	172	5156	124
NHR	NHR1	11	691	20
VGIC	VGSC	9	958	42

Family	Subfamily	# Targets	# pIC50	Log(p-val)
GPCR	GPCRA	16	1304	51

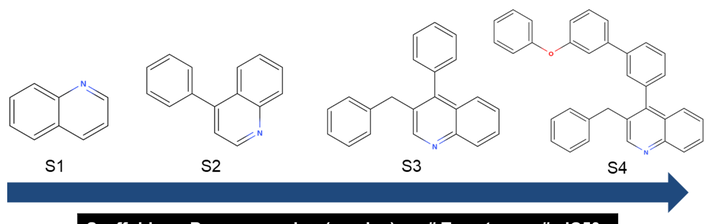
- Biphenyl is a commonly reported "privileged scaffold"
- Our analysis identifies biphenyl as privileged in three PT families
- Biphenyl-tetrazole is identified as privileged only in the GPCR PT family, suggesting that the addition of the tetrazole moiety generates selectivity, implying that it is a TFPS



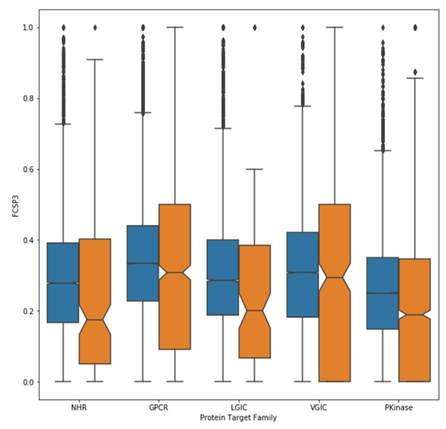
Tasosartan

- Biphenyl is present in 73 drugs spanning 10 ATC classes
- The biphenyl-tetrazole scaffold is present in 15 drug molecules targeting renin-angiotensin II

- In the NHR1 PT family quinoline is identified as privileged
- Compound optimisation can be observed through the increasing KS stat Dmax (distance between two ECDFs)
- The p-value represents a balance between the Dmax and the number of observations



Scaffold	Dmax	Log(p-value)	# Targets	# pIC50
S1	0.21	15.9	10	416
S2	0.35	27.3	2	256
S3	0.40	8.5	2	62
S4	0.54	6.6	2	26



Analysis of carbon hybridisation fractions
Privileged Scaffolds, *Molecules*

- As expected all scaffolds and molecules tend to have a higher Fsp2 contribution
- As expected, kinase's compounds have the lowest Fsp3 contribution; while GCPRs the highest
- More strikingly the PS generated from NHR and LGIC have low Fsp3 contribution not reflected in the molecules

Conclusions

- It is difficult to define privilege, without oversimplification of drug-target relationships
- Limited by the amount and quality of bioactivity data within public databases
- Scaffolds like Biphenyl-tetrazole may only be identified as a TFPS because it has only been tested within one TF
- Bias for particular scaffolds in library construction may still exist

References:

- Aric A. Hagberg, Daniel A. Schult, Pieter J. Swart. (2008). Exploring network structure, dynamics and function using NetworkX. *Proceedings of the 7th Python in Science Conference (SciPy2008)*
- Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., Chang, R. S. (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *Journal of Medicinal Chemistry*, *31*(12), 2235–2246.
- RDKit: Open-source cheminformatics; http://www.rdkit.org
- Varin, T., Gubler, H., Parker, C., Zhang, J.-H., Raman, P., Ertl, P., Schuffenhauer, A. Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data. *J. Chem. Inf. Model.* 2010, 50, 2067-2078
- Varin, T. Schuffenhauer, A., Ertl, P., Renner, S. Mining for Bioactive Scaffolds with Scaffold Networks: Improved Compound Set Enrichment from Primary Screening Data. *J. Chem. Inf. Model.* 2011, 51, 7, 1528-1538

oliver.scott.17@ucl.ac.uk