

Strategies for assembling an annotated library for phenotypic screening

31/05/2018

Henriëtte Willems

LifeArc



**Alzheimer's
Research UK**
The Power to Defeat Dementia

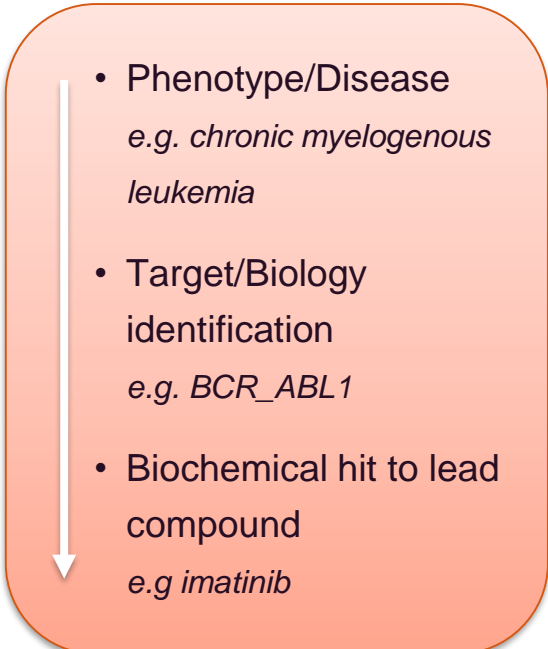


UNIVERSITY OF
CAMBRIDGE

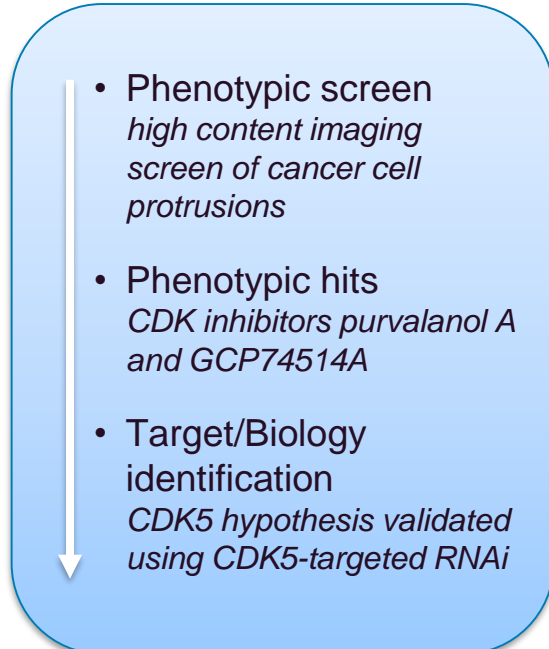
DRUG DISCOVERY INSTITUTE

What is phenotypic screening?

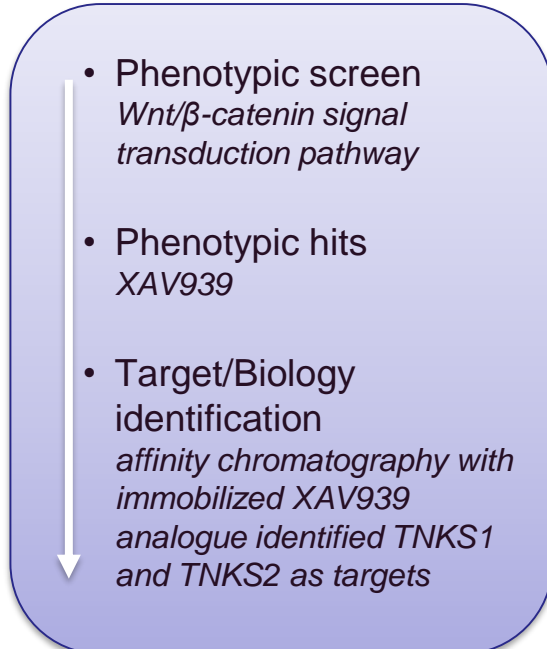
Target-based drug discovery

- 
- Phenotype/Disease
e.g. chronic myelogenous leukemia
 - Target/Biology identification
e.g. BCR_ABL1
 - Biochemical hit to lead compound
e.g. imatinib

Phenotypic screen: Cell-imaging

- 
- Phenotypic screen
high content imaging screen of cancer cell protrusions
 - Phenotypic hits
CDK inhibitors purvalanol A and GCP74514A
 - Target/Biology identification
CDK5 hypothesis validated using CDK5-targeted RNAi

Phenotypic screen: Pathway based

- 
- Phenotypic screen
Wnt/ β -catenin signal transduction pathway
 - Phenotypic hits
XAV939
 - Target/Biology identification
affinity chromatography with immobilized XAV939 analogue identified TNKS1 and TNKS2 as targets

Wang, Y. *et al.* (2016). *Cell Chemical Biology*, 23(7), 862–874.
Quintavalle, M. *et al.* *Sci. Signal.* 4, ra49 (2011).
Huang, S-M., *et al.* (2009) *Nature*, 461, 614-620.

Why phenotypic screening?

- Phenotypic screening has been very successful in the past
 - 28 out of 50 first-in-class FDA-approved small molecules from phenotypic screens (Nat. Rev. Drug Discov. 2011, 507-519)
 - However, another study: 78 (69%) of 113 first-in-class drugs from target-based approaches, only 33 did not have a target hypothesis, but this study included biologics (Nat. Rev. Drug Discov. 2014, 577-587)
- More disease relevant, greater confidence that hits will deliver the desired therapeutic effect
 - Hits often inhibit a pathway, not just a protein that can be bypassed
 - Typically cell-based assay, so hits have already cleared permeability hurdle
- But, without a biological target it is difficult to derive SAR, or to avoid mechanism-based toxicity
- Therefore, identifying the biological target of a phenotypic screening hit accelerates drug development

What compounds to screen?

General consensus

- Lower throughput assays: 1000s of compounds
- Small molecules with well-annotated pharmacology

SGC criteria for tool compounds:

- in vitro potency <100 nM,
- >30-fold selectivity over related proteins
- profiled against off-targets and proteins relevant to drug discovery
- on-target effects in cells at <1 μ M.

'orthogonal' chemical probes

- different chemical structure with activity for same target
- reduces the probability of having common off-target

'dark chemical matter'

molecules that have been frequently screened in TDD projects but that have not shown any activity

Arrowsmith, C. H., et al. (2015). *Nature Chemical Biology*, 536–541.
Wassermann, A. M. et al. (2015). *Nature Chemical Biology*, 958–966

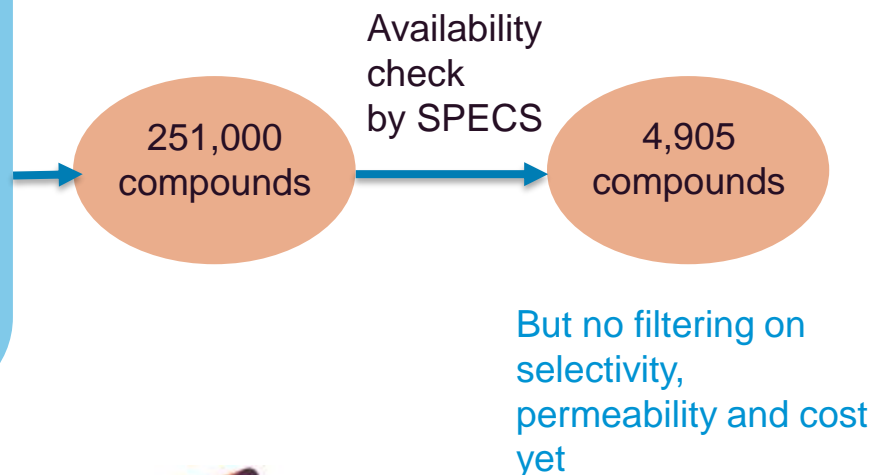
ARUK approach to phenotypic library

ARUK Cambridge DDI wishlist

- Small molecules that have well-annotated pharmacology
- Cell permeable
- Selective
- Wide target coverage, with focus on non-GPCR, non-kinase ligands
- 2000 to 5000 compounds in joint library with LifeArc

ChEMBL mining selection criteria

- Potency < 300 nM (pchembl value > 6.5)
- Human, rat or mouse assay data
- MW < 1000 and > 50
- Confidence score for target > 5
- No '>' values in potency
- Only assay type 'B' (binding)



What is selective?

How much data is needed before selectivity is meaningful?

2, 3, 10 different assays per compound?

Can a compound that hits multiple targets be selective?

What is an acceptable ratio between targets hit and targets not hit?

Absolute or relative potency values?

If a compound has potency of 0.1 nM at primary target is a potency of 10 nM at a secondary target acceptable?

How large should the potency gap be?
10-fold, 100-fold?

What is more important?
Selectivity against similar targets or against different targets?

What about data that doesn't specify the receptor subtype?

Selectivity metrics

- Windows score (Bosc, Meyer, Bonnet)
 - **WP-2 count**: how many targets are hit in 'window' that ranges from [Primary potency - 2] to [Primary potency] (i.e. 2 log units)
 - **WP-1 count**: [Primary potency - 1] count
 - **Score**: [targets counted in window] / [targets tested for in total]
 - Therefore smaller windows score means more selective
- Windows scores were implemented in R and calculated for all potent ChEMBL ligands

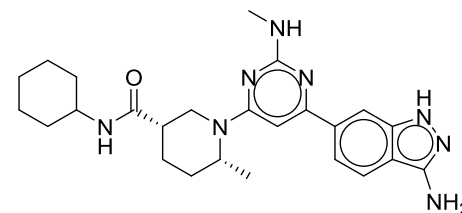
Reference:

The use of novel selectivity metrics in kinase research,

Bosc N, Meyer C, Bonnet P

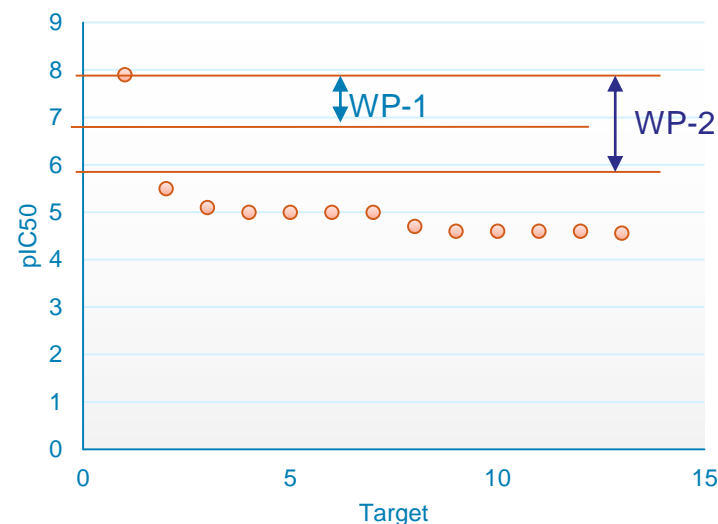
BMC Bioinformatics 2017 vol: 18 (1) pp: 17

Windows score implemented by Arushi Gandhi



Mean pIC50 = 7.9 at PDPK1

ChEMBL1765740



WP-1 count = 1

WP-2 count = 1

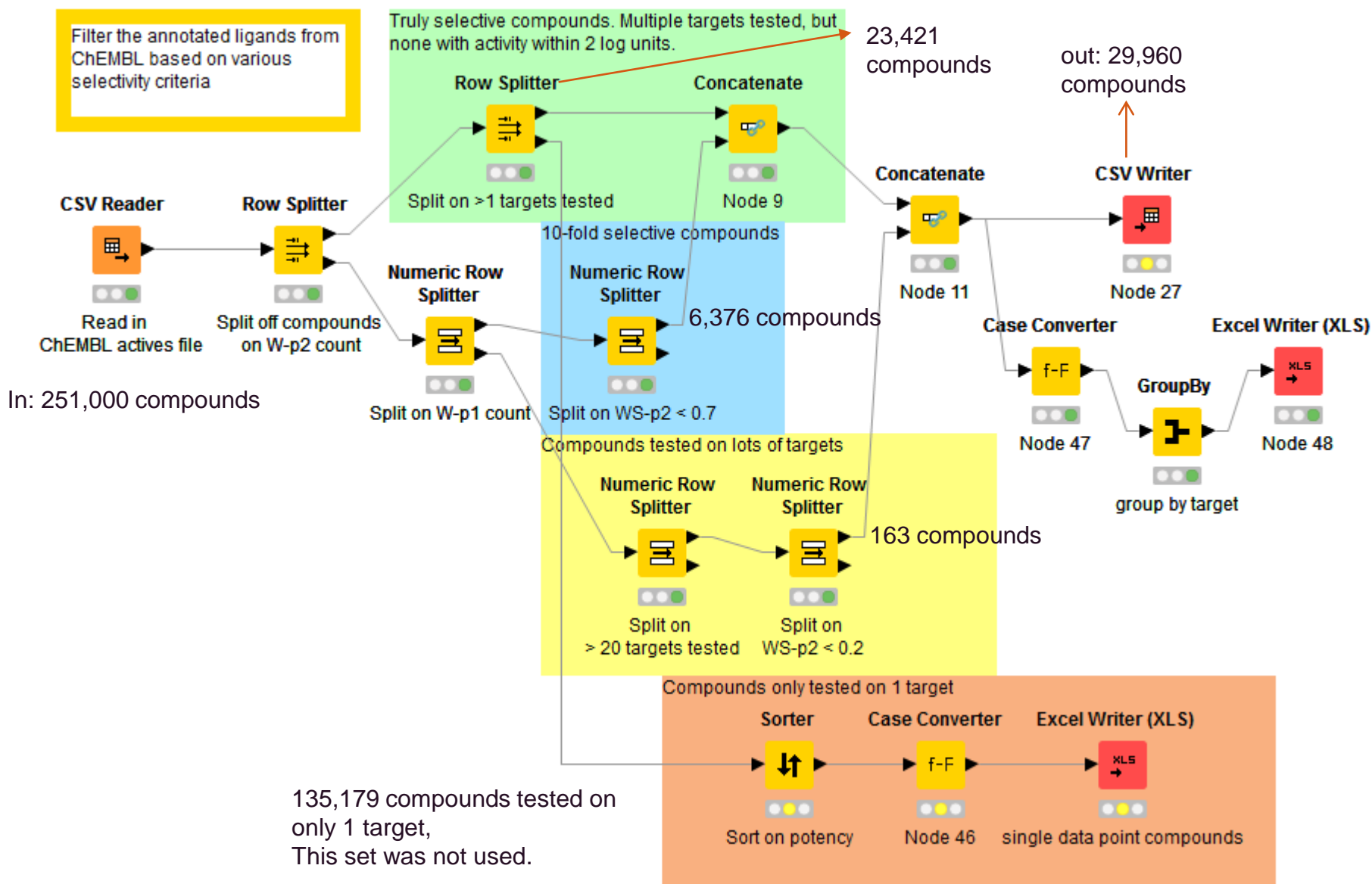
WP-1 score = WP-2 score = $1/13 = 0.08$

How we defined selective compounds

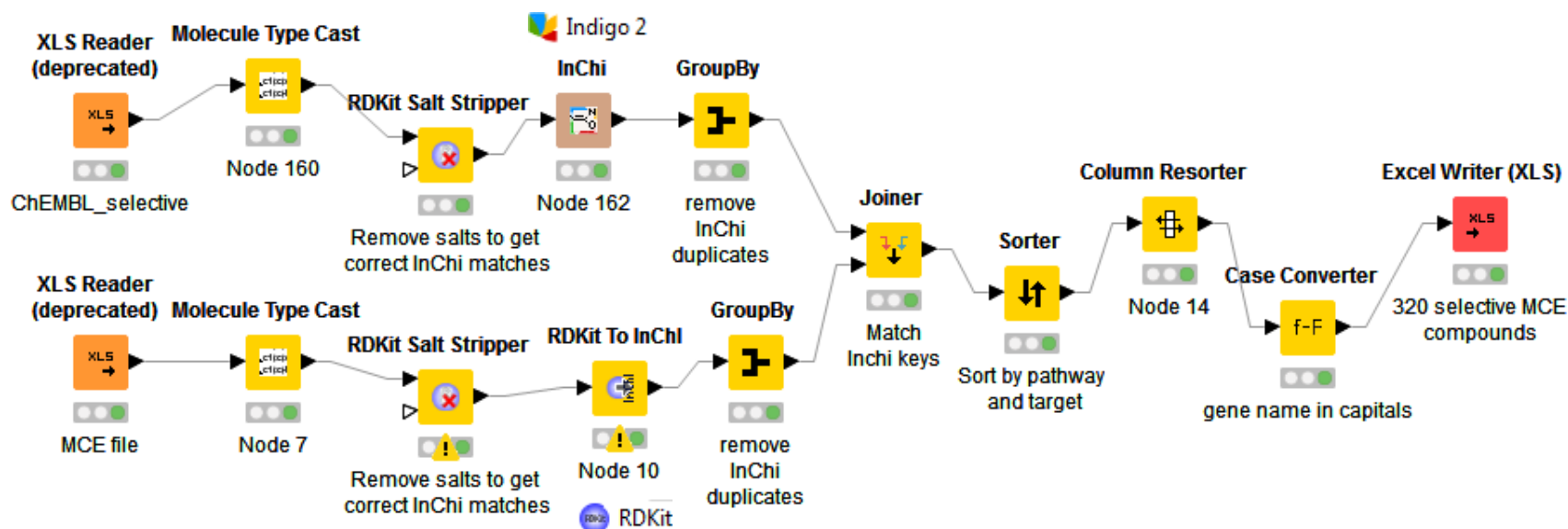
1. Green: WP-2 = 1 and number of targets > 1 (23,421 compounds)
 - This selects compounds that are active for 1 target and inactive/weakly active for at least one other
 2. Blue: WP-1 = 1 and number of targets > 1 and WP-2 score < 0.7 (6376 compounds)
 - Selects compounds that have only 10-fold selectivity over some other targets, but have more than 100-fold selectivity over some others.
 3. Yellow: WP-1 and WP-2 count > 1, tested on > 20 targets
AND have a WP-2 score < 0.2 (163 compounds)
 - This selects compounds that hit several targets with almost equal potency, but are inactive for 9 out of 10 targets they were tested on)
- Total: 29,960 'selective' compounds

- Issues:
- protein targets for different species have different identifiers
 - Transporters and CYPs included in selectivity scores
 - % inhibition and PubChem negative data not included
 - Duplicate compounds with different data due to multiple ChEMBL IDs

Filtering for selective compounds



Commercial availability



- MedChem Express and SPECS databases were searched with selective ChEMBL set and probes
- MCE was chosen because affordable, SPECS because it can source from many different vendors and has ongoing collaboration with LifeArc
- Both the ChEMBL set and the vendor sets contains >100 InChi duplicates
- Some targets were covered by many vendor compounds, so a further selection on price, properties (ChemAxon cxcalc logD, logP, PSA and pKa) and amount of annotation was carried out

Comparison with LifeArc and ARUK collections

- The phenotypic library project is a collaborative project with LifeArc
- LifeArc's existing phenotypic library was selected partly based on similarity, so did not have ChEMBL annotation matching current workflow
- Both the LifeArc and ARUK collections contained tool compounds that are not in ChEMBL, but have data in e.g. Chemical Probes or Selleck

Practical issues:

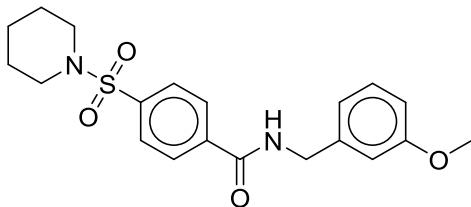
- InChiKeys calculated on desalted molecules work best for searching ChEMBL and matching data
- Compounds can have multiple ChEMBL ID's for different salts or isomers
- Chemical Probes website and various vendor websites with annotations are not set up to search with lists of molecules

More philosophical issues:

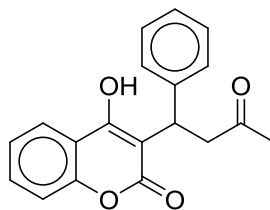
- How to annotate if not in ChEMBL?
 - Which other sources to check
 - Identifier to use?
- Annotating compounds that have a cell line or protein complex as target
- Should poorly annotated compounds be included?

Annotating existing ARUK and LifeArc collections

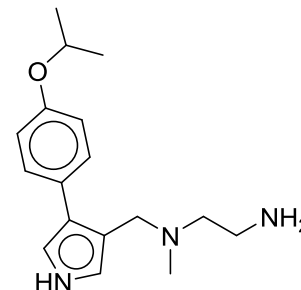
Some examples of troublesome compounds



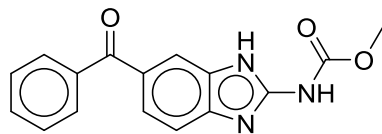
CHEMBL2130561,
No binding data, but 112 nM against
USP1 in PubChem primary screen,
activity comment: inconclusive



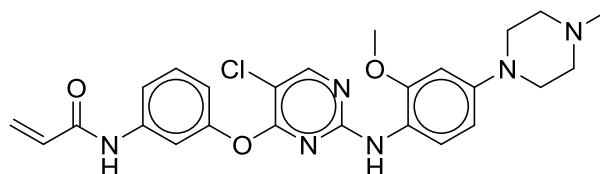
CHEMBL1464 (CHEMBL7252,
CHEMBL251073, CHEMBL251074,
CHEMBL1416568)
Warfarin, Approved drug,
No protein target with potency <10 uM in
ChEMBL



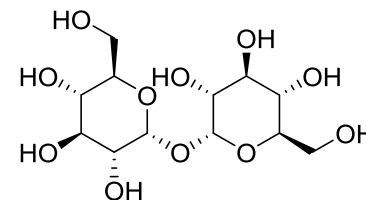
No ChEMBL ID,
MS023, Chemical Probe,
Inhibits PRMT1,3,4,6 and 8
4-119 nM



CHEMBL685,
Mebendazole, Approved drug,
Antiparasitic,
No mammalian protein target

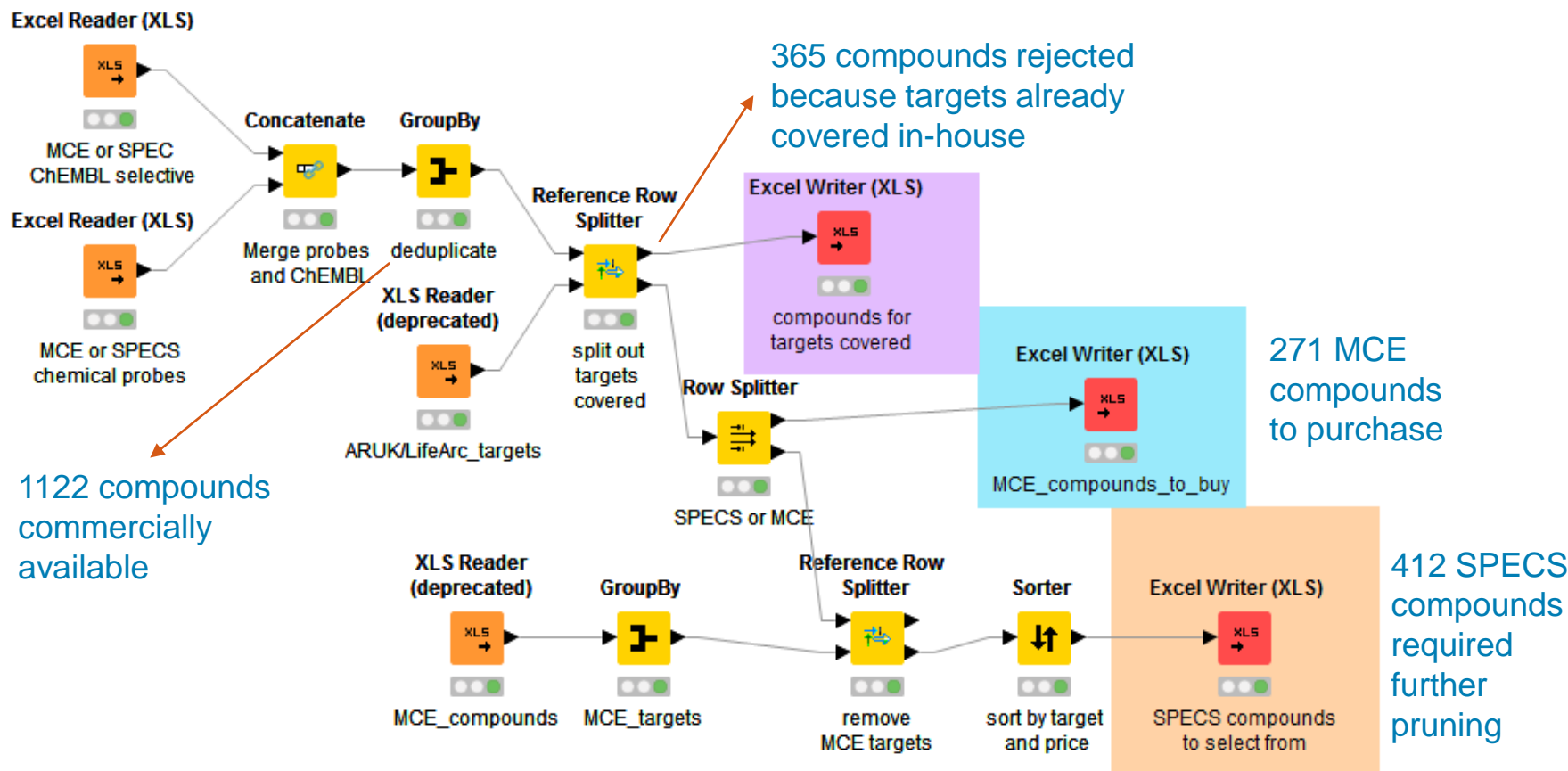


CHEMBL1229592,
Potent inhibitor of the epidermal
growth factor protein family,
Single ChEMBL target ID, but
multiple UniProt and gene name
identifiers



CHEMBL685,
Trehalose,
Active in huntingtin aggregation assay.
No known mammalian protein target

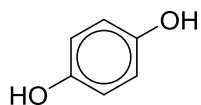
Compound selection to cover new targets



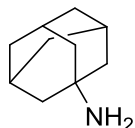
- 2 or more selective compounds in in-house collections: target filtered out
- Manual selection where >2 compounds per target were available
- Remove compounds with undesirable properties: ChemAxon logD>5, PSA>140, rot bonds>10

To buy or not to buy?

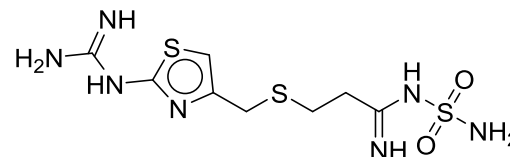
- Some examples of well annotated, selective compounds in ChEMBL



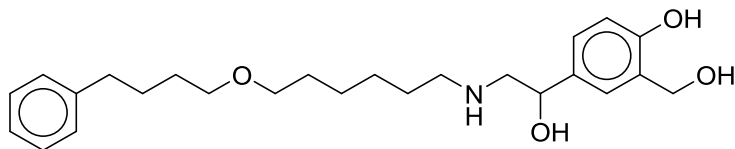
CHEMBL537, hydroquinone



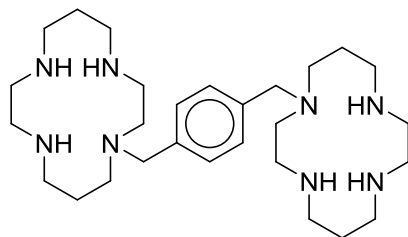
CHEMBL660, amantadine



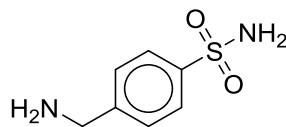
CHEMBL902, pepcid



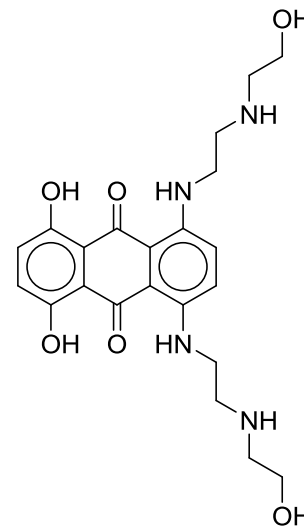
CHEMBL1263, Salmeterol



CHEMBL18442, Plerixafor

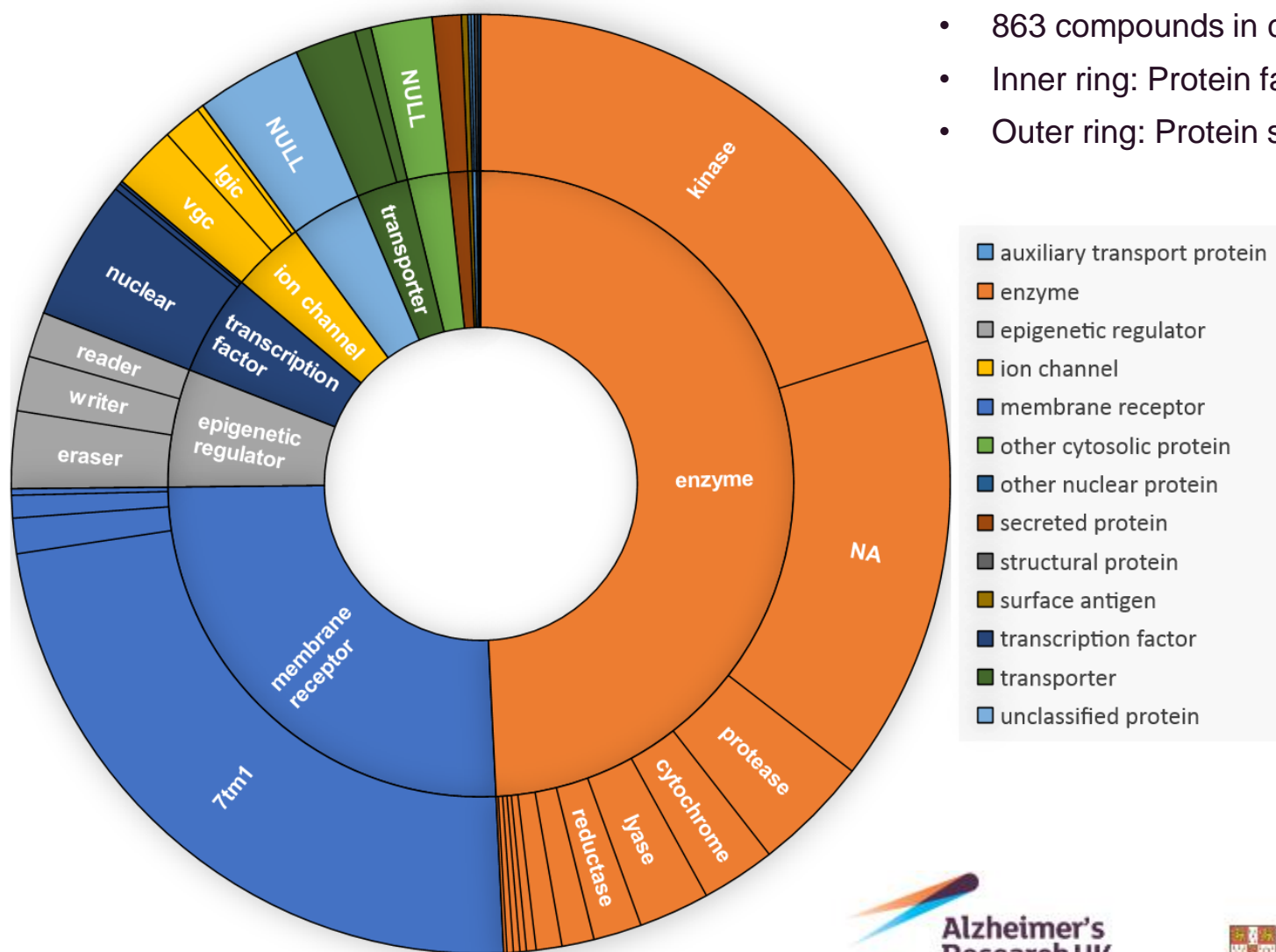


CHEMBL419, Mafenide



CHEMBL58, Mitoxantrone

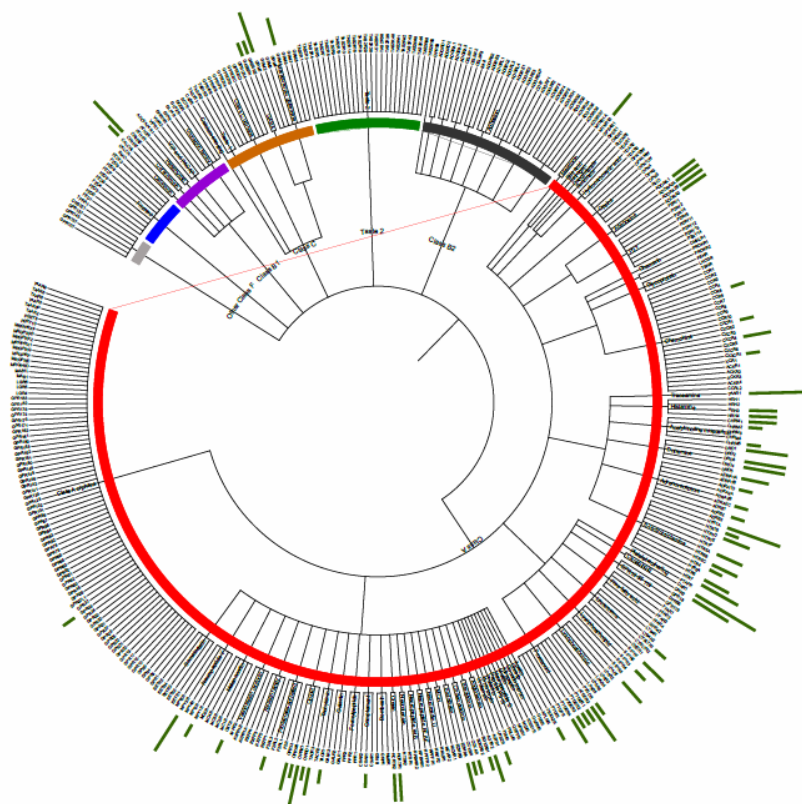
Target families and subfamilies in the phenotypic library



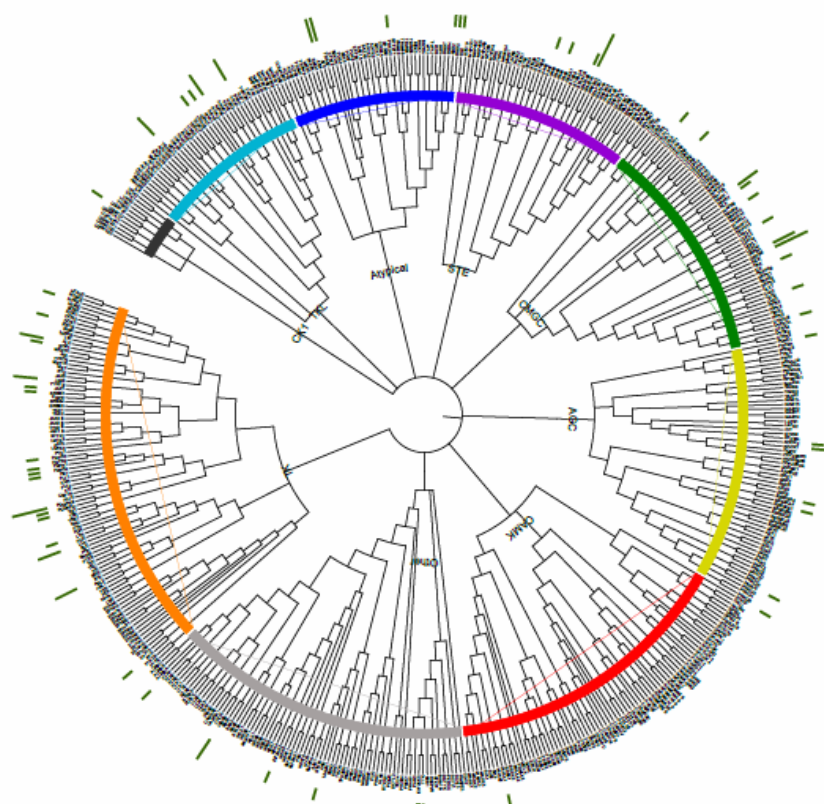
- 863 compounds in current library
- Inner ring: Protein families
- Outer ring: Protein subfamilies

Which targets have we not covered?

Plots generated by Peter Sterk



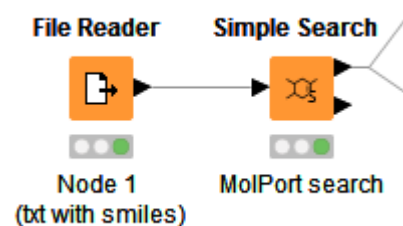
GPCR targets covered by our current set, mapped onto all GPCR targets in UniProt



Kinase targets covered by our current set, mapped onto all kinase targets in UniProt

What selective compounds have we missed?

- Target was for 2000-5000 compound library, so far <1000
- After mining ChEMBL and chemical probes/SGC websites where else can we look?
 - PubChem? Best data in ChEMBL, but much more data available in PubChem. Is it reliable?
 - SureChEMBL? Difficult to mine for non-specialists
 - Biology literature? Not comprehensively covered by ChEMBL, but time-consuming to mine
 - PDB?
- Other vendors:
 - MolPort and MCule have excellent portals for searching for compound availability with large lists of SMILES
- Custom synthesis?
 - Expensive and time-consuming, but large list of compounds of interest available



Should we include unselective or inactive compounds?

- 23,600 somewhat selective compounds
- Mostly for targets already covered.

- 135,000 ChEMBL compounds are <300 nM but only tested on 1 target
- App. 44,000 of those active at targets not covered in current library
- But % available probably quite low

- 62,500 ChEMBL unselective compounds
- Some of these contain a lot of 'inactive' data, could be useful as control

- Inactive compounds of same chemotype as actives
- These could be useful as controls for involvement of target

- Trade off between quality of annotation and range of targets covered
- No selective compounds are available for underexplored targets

Future expansion of our annotated library

- Investigate other vendors
 - MolPort showed availability for some compounds not available from SPECS.
 - So even vendors that aggregate data from other vendors have different compound collections.
 - Mcule and eMolecules can also be searched with large lists of compounds.
- Find compounds to purchase that only have activity data against 1 target
 - Pool of 135,000 compounds, of which app 44,000 were active at a target that is not yet represented
 - App. 500 of appear to be available (but not necessarily affordable)
- Find compounds that have been tested against a range of targets but are inactive: 'dark matter'
- Find similar, but inactive compounds as controls

Acknowledgements



ARUK DDI

Steve Andrews
John Skidmore
Peter Sterk

Chemistry Department

Stephanie Ashenden
Andreas Bender
Arushi Gandhi



LifeArc

Chido Mpamhanga
Andy Merritt



EBI

Francis Atkinson
Patricia Bento
Anne Hersey
Andrew Leach

