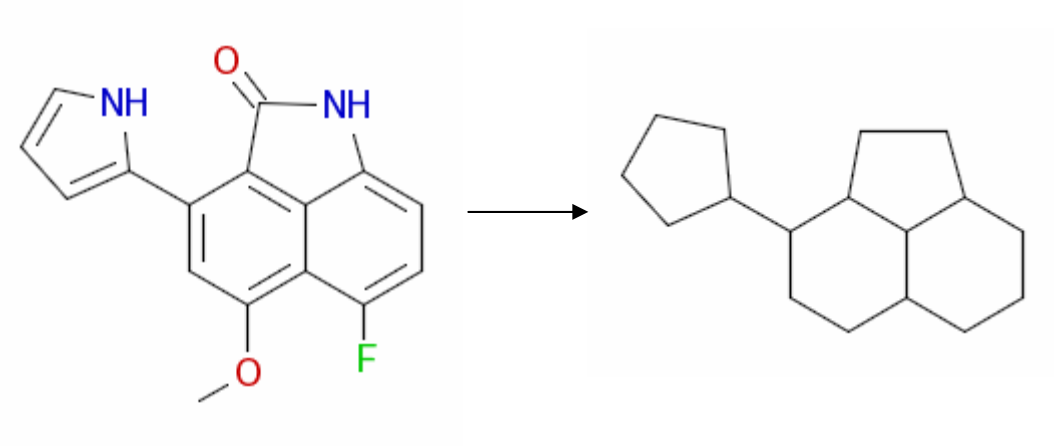
Clustering Procedures for DUD and WOMBAT data

All data was clustered and filtered according to the following procedure

The loose lead-like cutoff defined by Oprea et al. JCICS (2001), 41(5), 1308-1315 was used to filter each set. (AlogP <4.5 (5.5 for NHR targets to reflect the penchant for hydrophobic moieties) / MW <450). This was done both to create more realistic starting points and also reduce the complexity of the molecules to a level more in keeping with virtual screening sampling levels (for example with respect to conformational flexibility). DUD data was flitered using AlogP, while Wombat filtered using both ClogP and AlogP (both values had to meet the cutoff). Molecules were sorted by heavy atom count (primary sort) and then activity (secondary sort). Reduced graph assemblies were created for all molecules (example shown below - *J. Chem. Inf. Comput. Sci.*, **43** (2), 346 -356, 2003), and



.

The smallest molecule was then selected from each cluster of molecules with identical reduced graphs. For Wombat, where molecules have identical size, the molecule with lower activity has been chosen. A constraint on the activity range of the WOMBAT data was also toyed with but eventually abandoned, since the activity data is being provided allowing users to constrain their own searches accordingly. For DUD no secondary sorting was undertaken so selection is based on ordering in the original mol2 file (all sorting and clustering undertaken in Scitegix version 5.1)

For the WOMBAT data only molecules with specific activity data have been retained (all < xxx activities removed). This data has been retained in the sdf files to aid in analysis, as has the parent structural reference).

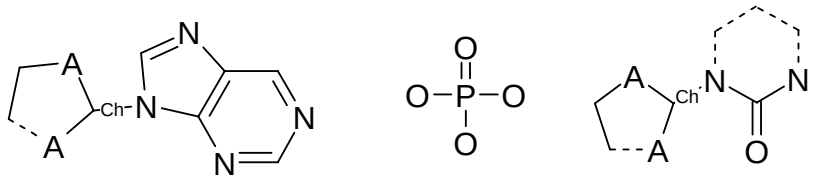
In all 13 target have been chosen for WOMBAT data extraction derived from 9 target classes. These include 3 kinases (tyrosine kinase EGFR, Cylcin dependent kinase CDK2 and map kinase P38), 3 nuclear hormone receptors (androgen receptor (AR) antagonists, PPAR gamma (PPARg) activators and estrogen receptor antagonists (ER), one GPCR for

homology model tests (D2 antagonists) and 6 assorted enzymes (FXa, COX2, PDE5, IMPDH, ALR2, HIV1RT).

Data limitations

No attempt has been made to separate human data from other species for any of the targets. Typically the sequence identity for the target chosen is high and often the alternative species has been chosen as a surrogate for human adta. Further, other than for Aldose reductase the no. of non-human data points is typically fairly limited. Nevertheless this may have some effect on the ability to interpret relative activities in some instances. Activity data also needs to be interpreted with the realization that the data has been extracted from multiple sources. This can have significant consequences. For example for CDK2 the cyclin variant chosen varies from assay to assay. Further the concentration of ATP is not known for each data point. Variations in ATP are known to to produce significant alterations in activity values returned. In addition, it is not known if each inhibitor class hits its given kinase in the same state (activated versus inactivated). Also the relative lop positions for the different chemotypes is unknown (e.g. P38 DFG loop in versus loop out). All this must be borne in mind during results interpretation, with users who find subtle issues based on these or other factors being encouraged to report them for future data annotation and refinement.

For HIV reverse trascriptase data, Wombat does not differentiate NNRTIs from NRTIS. An AlogP constraint > 1 and LogS constraint < -2 have been used to differentiate the two classes, and substructure searches on the primary NNRTI chemotypes reveal no hits, suggesting good differentiation. It is still possible that an NRTI or two still lurks in the data, however.



NRTI substructure check

On the same lines for acetylcholinesterase no differentiation is made between catalytic / peripheral / dual binding inhibitor classes, so both active sites should be considered for searches involving this target.

For two NHR target selections and the D2 data the biological effect field has been used to only keep molecules designated as antagonists. This definition is to some extent defined by the nature of the assay applied, however, so additional filters removing steroid and dopamine specific substructures from the lists have been used to further refine the data. There is still, however, some ambiguity regarding the antagonist definition. For example the biphenyl chemotype, a well known estrogen mimic toxin is defined as an antagonist within WOMBAT. This particular ambiguity is something that will need further consideration before data release. One possibility would be to refer to the sets as binders with a bias toward antagonists.

The clustering method chosen is, like all others, not a perfect technique for chemotype partition. Overall within a given target data set it does a good job of providing perhaps the most conservative method of chemotype assignment, with the following limitations. First as one might imagine it is less discriminating with simple systems, particularly

containing only 1 ring. As such the method will tend to produce less fragment-like chemotypes than one might want and struggles with target that contain uniquely small ligands (e.g. COMT). Blending this with an alternative similarity measure might be useful for these smaller systems. Secondly, the technique will still differentiate small changes in ring structure (e.g. phenyl for thiophene) as rendering a new chemotype when really said change creates a simple analogue. Similarly the addition of a small carbocycle e.g. cyclopropyl in place of a small alkyl will also produce a new chemotype. As such the method is still prone to throw the odd analogue into the list. These analogues are not particularly common, however, and could be removed manually (in the end the only way to guarantee removal without potential wholesale slaughter of the data set by more draconian automated measures), or left in to spice up the data. In the end people who use the data should be asked to provide all resulting hit lists with their article and include a 2D search test comparison as a control. This should allow any remaining analogue enrichment to be flagged. The method can also not differentiate from changes to the key binding elements versus changes to structure IP (e.g. amidine domination of the fXa data set) or pk optimization (thiazolidinedione chemotype modifications in PPARgamma). As such the intrinsic chemotype exploration between classes is not equivalent and should be borne in mind during results interpretation. Finally all compounds that do not contain a ring will not be assigned to a specific cluster and will thus bin into cluster 0 (the no. of molecules without no rings is very small).
Data users should feel free to contextualize their results from the perspective of these limitations. It is important to point out, however, that the method still produces clusters that in the vast majority of cases make sense. It is particularly adept at clustering the analogues that abound for chemotypes popular with medical chemists. PDF files associated with each clustered sd file have been provided to highlight this.

Results

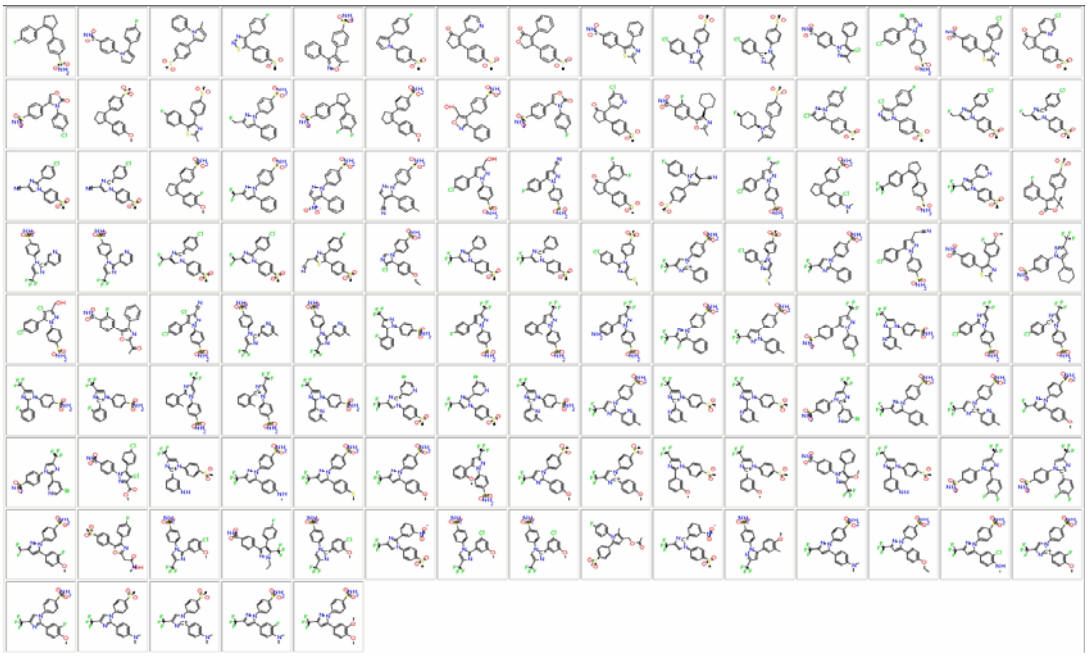
The results of the DUD data set filtering and clustering are shown below

Target	Total ligands	"Lead-like" Filter pass	filtered reduced graph clusters
ACE	**49**	**46**	**18**
ACHE	**106**	**101**	**18**
ADA	**37**	**37**	**8**
ALR2	**26**	**26**	**14**
AMPC	**21**	**21**	**6**
AR	**74**	**63**	**10**
CDK2	**58**	**55**	**32**
COMT	**11**	**11**	**2**
COX1	**24**	**23**	**11**
COX-2	**412**	**250**	**44**
DHFR	**407**	**387**	**14**
EGFR	**458**	**379**	**40**
ER Agonist	**67**	**63**	**10**
ER Antagonist	**39**	**18**	**8**
FGFR1	**170**	**73**	**12**
FXA	**146**		**19**

GART	**31**	**13**	**5**
GPB	**52**	**52**	**10**
GR	**78**	**9**	**2**
HIVPR	**62**	**6**	**3**
HIVRT	**41**	**35**	**17**
HMGA	**35**	**25**	**4**
HSP90	**25**	**24**	**4**
INHA	**86**	**58**	**23**
MR	**15**	**13**	**2**
NA	**49**	**49**	**7**
P38	**353**	**219**	**20**
PARP	**35**	**33**	**7**
PDE5	**76**	**34**	**22**
PDGFRB	**169**	**136**	**22**
PNP	**30**	**30**	**4**
PPAR gamma	**82**	**7**	**6**
PR	**27**	**22**	**4**
RXR alpha	**20**	**18**	**3**
SAHH	**33**	**33**	**2**
SRC	**159**	**102**	**21**
THROMBIN	**68**	**26**	**14**
TK	**22**	**22**	**7**
TRYPSIN	**46**	**10**	**7**
VEGFR2	**78**	**49**	**31**
Average	**94**	**66**	**13**

The large scale reduction in data set size on clustering highlights the analogue bias intrinsic to the data.

A sample cluster from Cox-2 representing > 25% of the total data set sharing the reduced graph of celebrex and vioxx is shown below



The WOMBAT clustered set sizes are shown below, together with the equivalent DUD sets. The additional data present in these sets is pretty clear, with on average >3 times the number of compounds in the WOMBAT data sets

Target	Wombat	DUD
CDK-2	152	32
EGFR	74	40
P38	59	20
AR	36	10
PPARG	27	6
ER alpha	64	8
COX-2	76	44
ALR2	42	14
PDE-5	88	22
HIV-RT	99	17
fXa	107	19
IMPDH	49	
D2 antagonists	323	
Average	72	21

DATA output and future directions

The dud_clustered.zip file contains lead filtered DUD data organized by cluster (*_cluster.sdf), with the smallest example from each list being placed in a secondary list (*_parents.sdf) (with associated pdf for your reference to browse the clusters). It would seem reasonable to recommend to users that any enrichment calculations include

enrichment graphs corrected to include no. of unique clusters (and thus by extension chemotypes) located (rather than hits alone) by the techniques analyzed.

Note: Scitegix cocks up the odd structure when converting from mol2 to 2D to sdf. Most look fine but I have noticed oddities in the kinase structures where the whole system has been saturated e.g. ZINC03815528 for FGFR1. Name and cluster info is fine, however, and this is all that is required to allow the enrichment adjustment to occur (given the reduced graph approach it has no effect on cluster assignment, which is lucky!).

Wombat.zip contains the cluster parents only. This is always the smallest compound in the cluster so should always be dockable. This was done so as not to place too much of the database in the public domain, though we might want to consider including multiple examples from each cluster as another option.

This should likely be an ongoing project, and as such perhaps the Wombat data set collation can be transferred to UCSF for further optimization/augmentation. Something to consider…. (as part of this the Scitegix scripts are available on request).