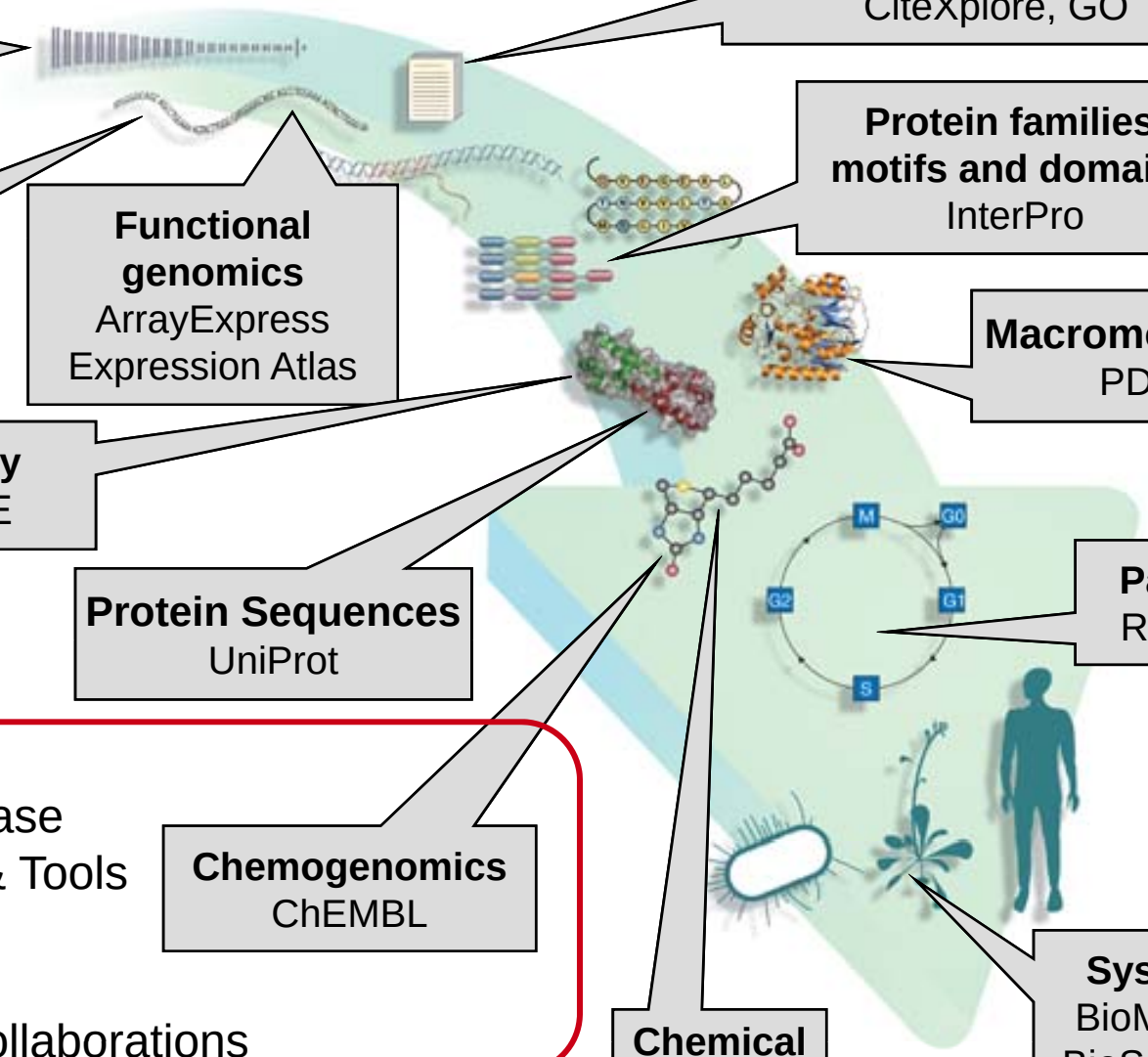
ChEMBL & Structural Alerts

Francis Atkinson

Chemogenomics Group

EMBL – EBI, Hinxton

European Bioinformatics Institute



Genomes
Ensembl
Ensembl Genomes
EGA

Literature and ontologies
CiteXplore, GO

Protein families, motifs and domains
InterPro

Functional genomics
ArrayExpress
Expression Atlas

Nucleotide sequence
ENA

Macromolecular
PDBe

Protein activity
IntAct , PRIDE

Protein Sequences
UniProt

Pathways
Reactome

- Service group
  - Core database
  - Interfaces & Tools
- Research group
  - IMI eTox
  - Industrial collaborations

Chemogenomics
ChEMBL

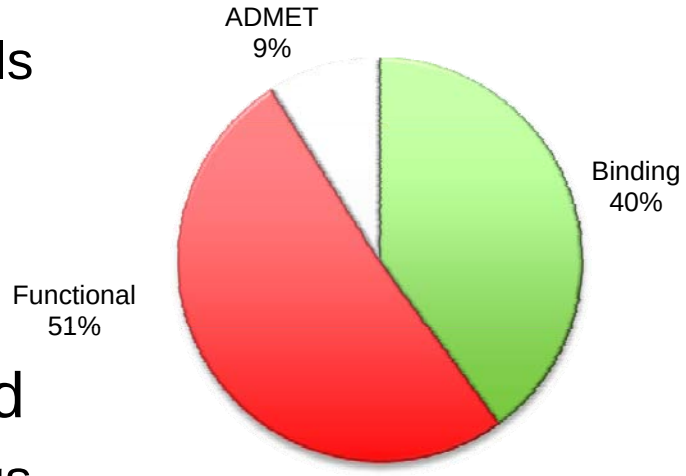
Chemical entities
ChEBI

Systems
BioModels
BioSamples

ChEMBL

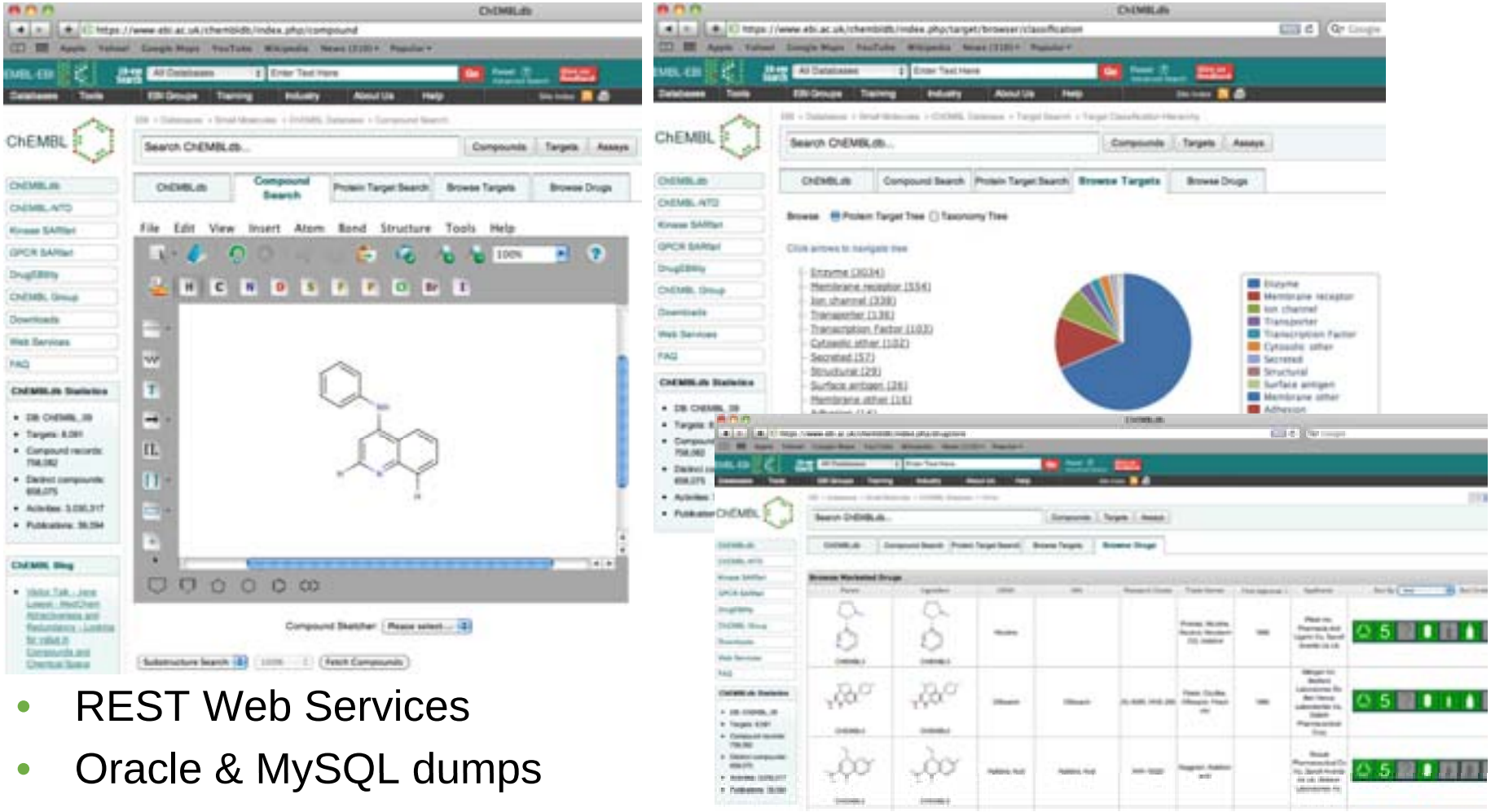
Current release: ChEMBL 09

- Open-access, curated database of bioactivity data
  - Focused on small, drug-like molecules
  - Mainly extracted from MedChem journals
  - Regularly updated
- > 3 M bioactivities
- ~ 660 K distinct compounds
- FDA Approved Drugs recently added
  - Development phase, Black Box warnings
  - Orange Book status (Rx, OTC, discontinued)
- Clinical Candidates coming soon
  - ~12,000 2-D structures/sequences
  - ~35-45,000 compounds
- ChEMBL data available through PubChem
  - Loading of PubChem data into ChEMBL underway



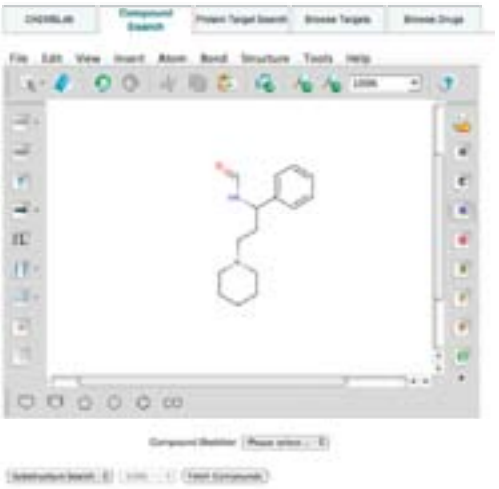
Accessing ChEMBL

- https://www.ebi.ac.uk/chembldb

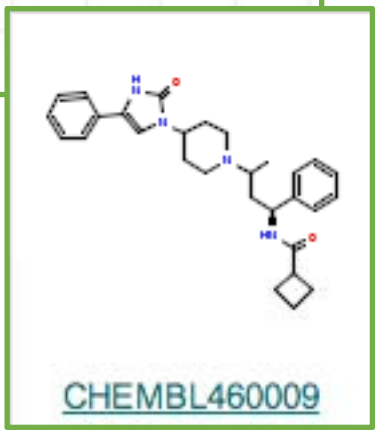


- REST Web Services
- Oracle & MySQL dumps

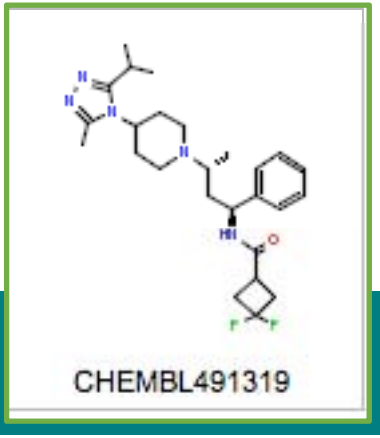
Using ChEMBL to design Safer Medicines



CHEMBL460009	CHEMBL460009	IC50	=	4.7 nM	HERG
CHEMBL460217	CHEMBL460217	IC50	=	117 nM	HERG
CHEMBL456019	CHEMBL456019	IC50	=	248 nM	HERG
CHEMBL469542	CHEMBL469542	IC50	=	849 nM	HERG
CHEMBL459600	CHEMBL459600	IC50	=	1650 nM	HERG



CHEMBL256907	CHEMBL256907	CL	=	74 ml kg-1 min-
CHEMBL491319	CHEMBL491319	CL	=	92 ml kg-1 min-
CHEMBL491319	CHEMBL491319	F	=	6 %



- See also: ‘Probing the links between in vitro potency, ADMET and physicochemical parameters’
- Nat. Rev. Drug Disc. 2011, **10**, 197-208.

Structural Alerts

- Substructures marking molecules as ‘of concern’
- Compound Filtering
  - Identify molecules that are somehow ‘undesirable’
  - Differ across industries, discovery phase, therapeutic area
- Category Formation
  - Identify chemicals with common mechanism of action
  - Form groups for QSAR or ‘read-across’ strategies
- Used in R&D and regulatory contexts
  - Pharmaceuticals
  - Food additives
  - Industrial chemicals
  - Environmental toxicity

How are they typically defined?

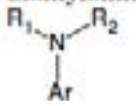
- Electrophilicity is a recurring theme…
- *In-vitro* mutagenicity or *in-vivo* carcinogenicity data
  - Often related to DNA damage by electrophiles
  - Mechanisms of carcinogenicity not always clear
- Skin sensitisation
  - Important endpoint, but also proxy for idiosyncratic ADRs
  - Electrophiles form adducts with proteins (haptenisation)
- Reactivity with thiols *in-vitro*
  - Cysteine residues in proteins or GSH
- Principles of organic chemistry
  - Reactivity prediction
- Dosed molecule and/or metabolites may be of concern
  - Need to consider metabolic activation

How are they typically defined?

- CYP450 inhibition
  - Heme-binding fragments
- CYP substrates
  - Particularly if metabolism generates reactive species
- Poor solubility or permeability
- Instability in solution, buffers or serum
- Compounds that interfere with assays
  - Reactives, fluorophores, aggregators
- Promiscuous chemotypes
  - Highly unselective kinase inhibitors
- Lack of novelty
- Experience & intuition
  - Chemotypes that have repeatedly failed in development

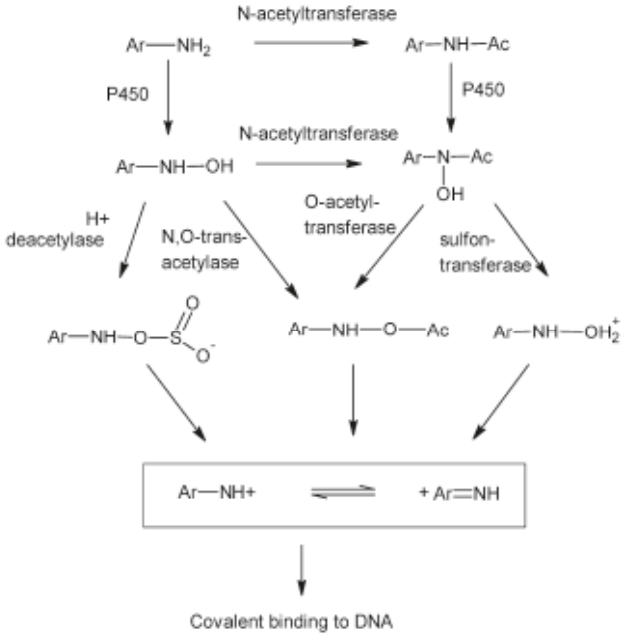
Example alert with ‘Mitigating Factor’

SA_28bis: Aromatic mono- and dialkylamine

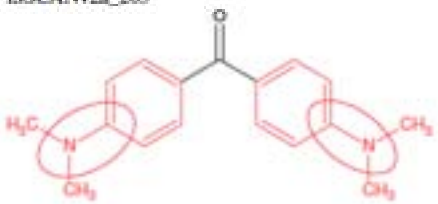


Ar = Any aromatic/heteroaromatic ring
R1 = Hydrogen, methyl, ethyl
R2 = Methyl, ethyl

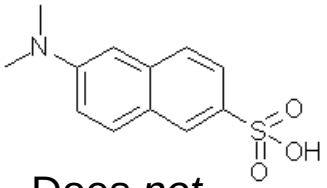
- Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded.
- Chemicals with a sulfonic acid group (-SO3H) on the same ring of the amino group are excluded .



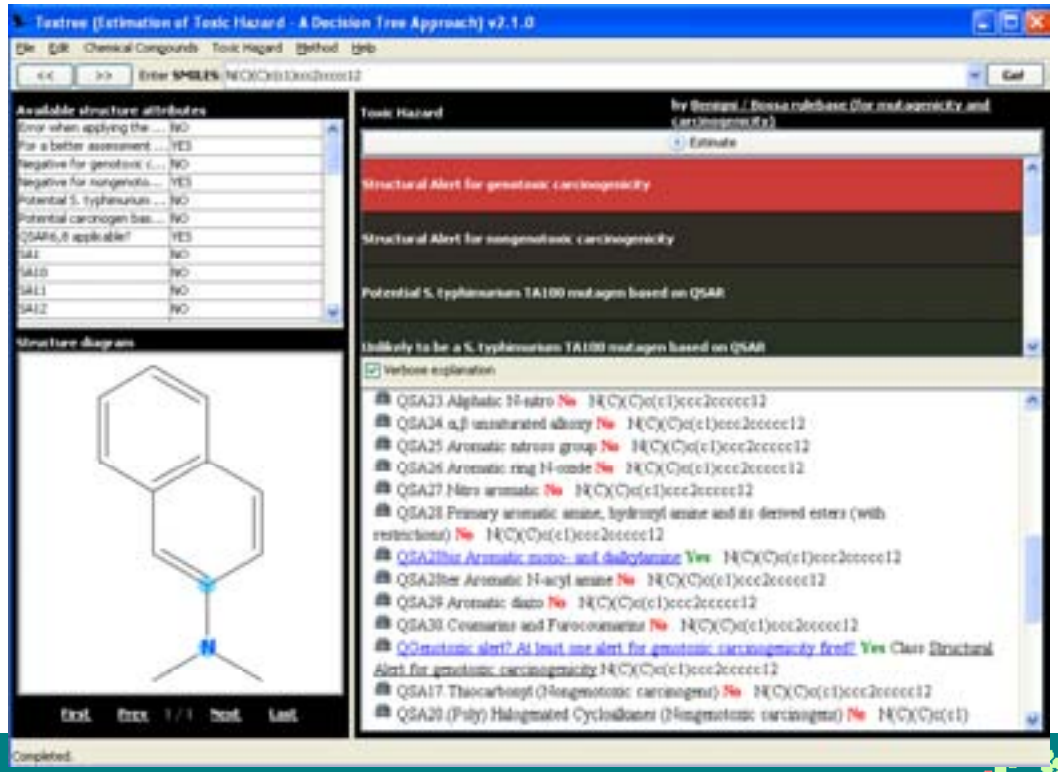
ISSCANv2a_203



ChemName: Michler's Ketone
CAS: 90-94-8
Mouse_Male : 3
Mouse_Female : 3
Rat_Male : 3
Rat_Female : 3



Does *not* match



ChEMBL

Free Tools incorporating Structural Alerts

- ToxTree
  - Toxic Hazard Estimation by decision tree approach
  - Application for download or can be run online
- QSAR Toolbox
  - Tool for Category formation, but can also simply identify S.A.s
  - Application for download
- smartsfilter
  - Web application hosted at the University of New Mexico
- EPA Tools
  - ECOSAR & OncoLogic
  - Applications for download

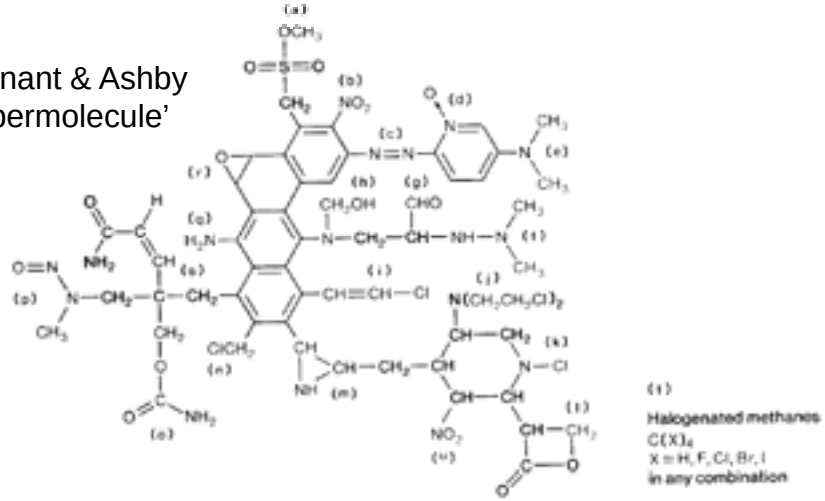
Motivation

- Provide tools to partners
  - Looking for ‘quick wins’
- Survey literature for sets of Structural Alerts
  - SMARTS (or equivalent) available to be used ‘as-is’
    - Several publications show alerts as depictions or text
      - coding these as SMARTS would be time-consuming
      - might not always capture the intent of the authors
  - Well-documented if possible
- Implement as Pipeline Pilot protocols
  - Straightforward SMARTS matching, not programmatic
  - More complex and/or property-based filters as a follow-on
- Investigate using ChEMBL data
  - Development phase, black box warnings, drug withdrawals
  - Also available are databases of side effects, ADRs *etc.*

Structural Alerts of historical interest…

- Tennant & Ashby's structural alerts for DNA reactivity based on analysis of *in-vitro* mutagenicity and *in-vivo* carcinogenicity data
  - Mut. Res. 1991, **257**, 209-227 (and others).
- Rishton's compilations of reactive moieties that can interfere with biochemical assays
  - Drug Disc. Today, 1997, **2**, 382.

Tennant & Ashby 'supermolecule'



Rishton's reactive groups

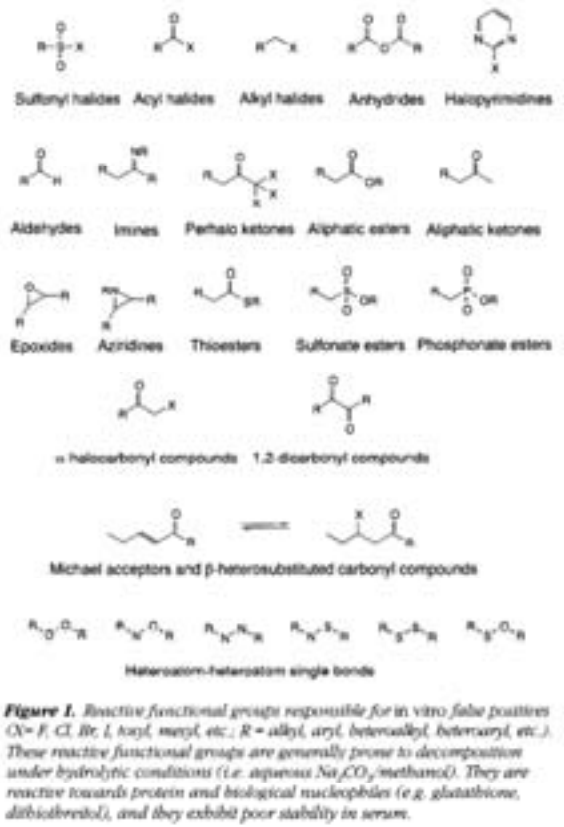


Figure 1. *Reactive functional groups responsible for in vitro false positives (X= F, Cl, Br, I, tosyl, mesyl, etc.; R = alkyl, aryl, heteroalkyl, heteroaryl, etc.). These reactive functional groups are generally prone to decomposition under hydrolytic conditions (i.e. aqueous Na_2CO_3/methanol). They are reactive towards protein and biological nucleophiles (e.g. glutathione, dithiothreitol), and they exhibit poor stability in serum.*

- These hugely influential sets of alerts have largely been superseded by others that incorporate and extend them

General Purpose Filter Sets

- Several ‘general purpose’ sets of alerts are available
- Intended for filtering out ‘undesirable’ compounds
  - HTS triage
  - Compounds for acquisition or synthesis
- Alerts are for a variety of the reasons discussed
  - Reactivity, PK, novelty *etc.*
  - Intuition / experience particularly important: “These filters, like many others currently being applied within the industry for triage, are based on chemical intuition and experience (bias).” (BMS)
- Level of annotation generally low
- Filters often dependent on historical context
  - As industry practice evolves, some alerts become redundant
- An important compilation of community knowledge

General Purpose Filter Sets

- Glaxo Wellcome: Hard Filters
  - J. Chem. Inf. Comput. Sci. 1999, **39**, 897-902.
- Pfizer: Lint procedure
  - Med. Chem. 2005, **1**, 649-655.
- BMS: HTS Deck Filters
  - J. Chem. Inf. Model. 2006, **46**, 1060**-**1068.
- NIH MLSMR: Excluded Functionality Filters
  - No literature reference, but see MLSMR homepage for details.
- University of Dundee: NTD Screening Library
  - ChemMedChem 2008, **3**, 435-444.
- Inpharmatica: Unwanted Fragments
  - 2006 (personal communication)

Specialised Alert Sets

- Focused on covalent modification of macromolecules
  - DNA: genotoxicity, mutagenicity and carcinogenicity
  - Proteins: skin sensitization, idiosyncratic ADRs
- Common mechanism
  - reaction of xenobiotic electrophiles with biological nucleophiles
- This mechanism is both important and ‘tractable’
  - Underlying organic chemistry well understood
  - Relatively abundant literature data for *e.g.* carcinogenicity
- Annotation tends to be much better
- Parent molecule or metabolites may be species of interest
- Some alerts can be used as filters
- Others are intended for defining mechanistic categories
  - As prelude to QSAR or read-across

Specialised Alert Sets

- Benigni/Bossa rulebase
  - Chem. Rev. ASAP 2011
  - Implemented in ToxTree
  - Includes some alerts for non-genotoxic carcinogens
- Cronin DNA-binding
  - Crit. Rev. Toxicol. 2010, 40. 728-748.
  - Implemented in QSAR Toolbox
  - Mechanistic: designed for use in category formation
- Cronin Skin Sensitisation
  - SAR QSAR Environ. Res. 2008, **19**, 555-578.
- ALARM NMR
  - J. Am. Chem. Soc. 2005, **127**, 217-224.
  - Derived from experimental binding to cysteine thiol in La protein

Others…

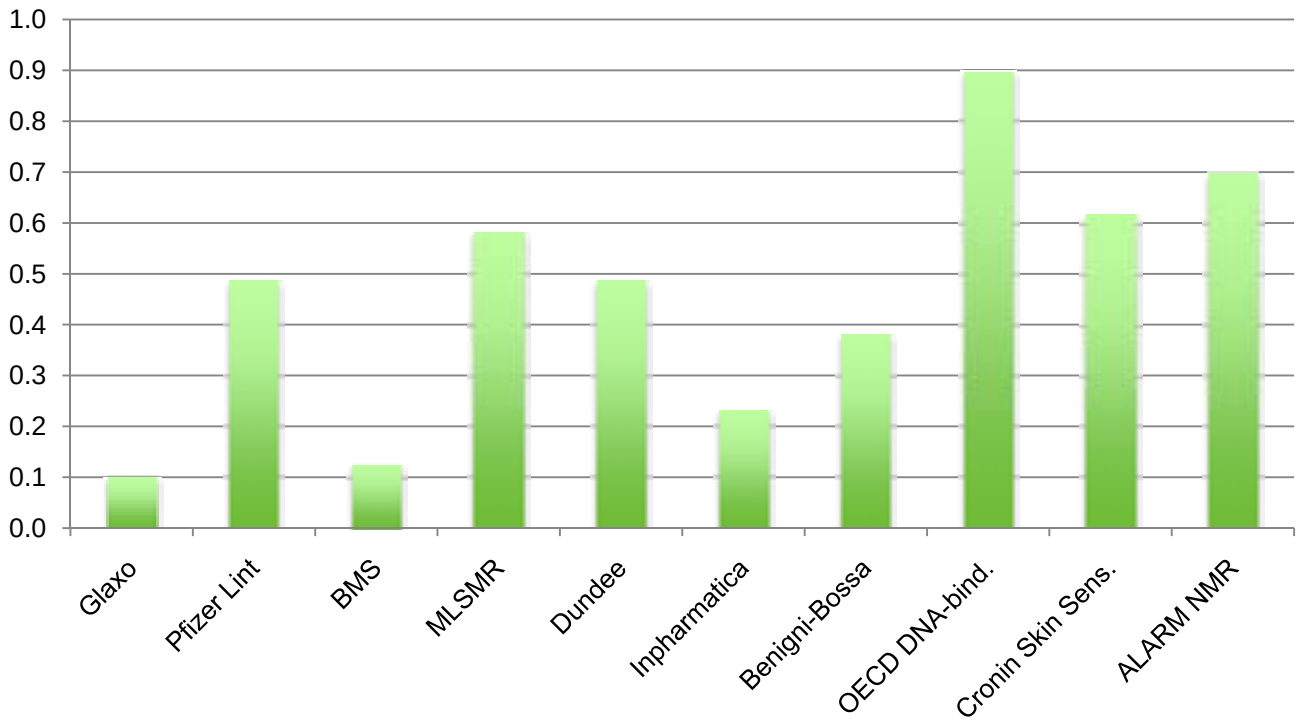
- Pan Assay Interference Compounds (PAINS)
  - J. Med. Chem. 2010, **53**, 2719-2740.
  - Alerts for nuisance compounds that elude usual reactivity filters
  - Available as SLN (not yet translated to SMARTS)
- SMARTCyp
  - ACS Med. Chem. Lett. 2010, **1**, 96-100.
  - Uses SMARTS in identification of sites of metabolic vulnerability
  - Available as a Pipeline Pilot protocol
- NIH CYP inhibition HTS
  - Nat. Biotechnol. 2009, **27**, 1050-1582.
  - Could the basis of a set of alerts for CYP inhibition
- TCSA New Chemicals Program Categories
  - Alerts of interest to EPA as possible environmental toxins

Issues encountered

- Alerts originally implemented using several systems
  - Daylight, ChemAxon, CDK, OpenEye, SLN…
  - SMARTS to must be ‘ported’ to Pipeline Pilot
- Subtle differences in SMARTS syntax
  - [!F&!Cl!&!Br&!I] matches Hydrogen in some dialects but not in PP
- Aromaticity definitions
  - phthalimide heterocycle is aromatic in some systems but not in PP
- Business rules for database normalisation
  - hypervalent *vs.* charge-separated nitro groups
- Some systems can use program logic for some alerts
  - Requires conversion to (sometimes complex) SMARTS
- Differences between SMARTS and documented alerts
- Errors in SMARTS

Hit rates for sets of alerts

- Fraction of ChEMBL matched by each set of alerts
  - Only compounds with AMW < 600 were used (~540 K)

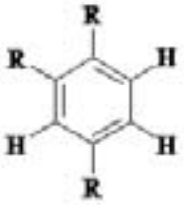


- Analysis ongoing: drugs *vs.* non-drugs, clean drugs *vs.* BBW *etc.*

Reasons for high ‘hit’ rates…

- Alert MA-10 (‘Arenes’) alone is hitting ~75% of the database...

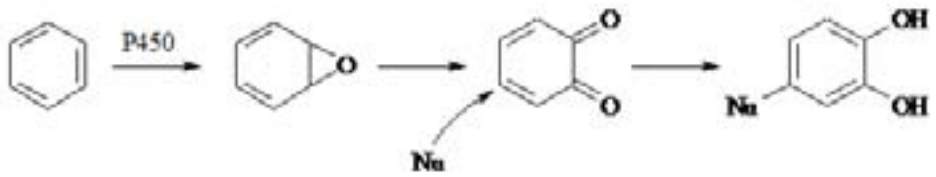
Category: Arenes



R = alkyl carbon, hydrogen

Mechanism

A P450 mediated epoxidation followed by conversion to a reactive quinone has been postulated as the primary cause of benzene derivatives ability to bind to biological nucleophiles (via a Michael addition mechanism) (Saghir et al 2009, Ishihama et al 2008).

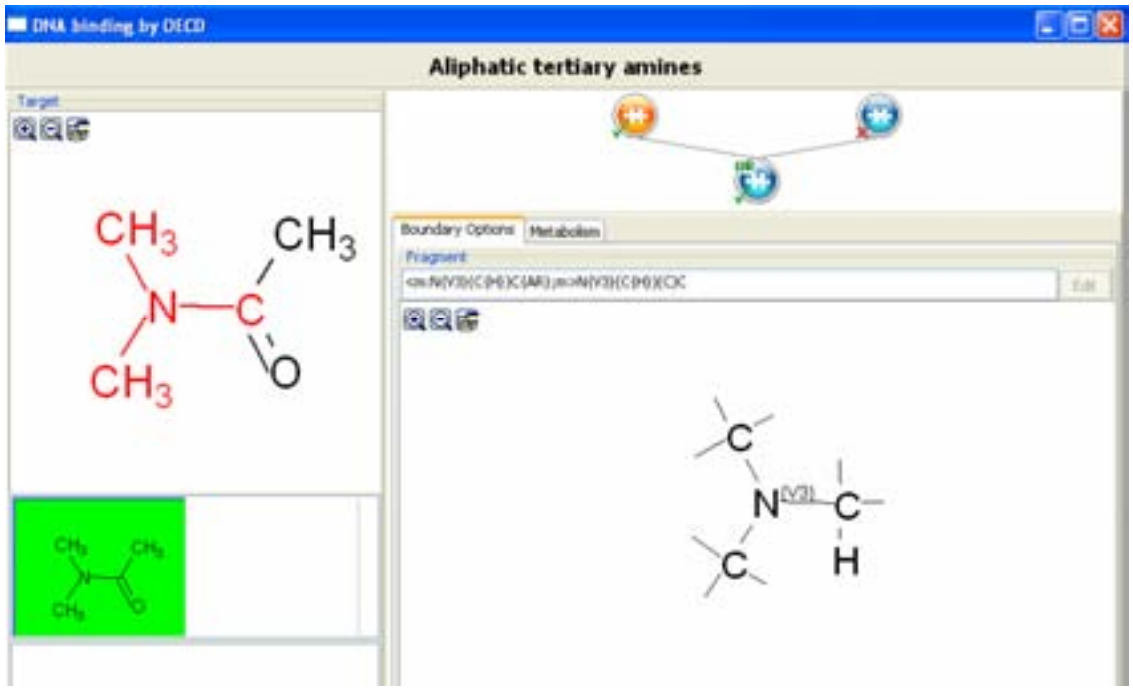


Nu = biological nucleophile

- This is indeed a *possible* route of metabolism, but…
  - The actual degree of vulnerability will depend on the molecular context
- Remember that these alerts are all not intended as filters!

Reasons for high ‘hit’ rates…

- SN1-15 (‘Aliphatic tertiary amines’) hits ~ 32% of the database...



Category: Aliphatic tertiary amines

R = aliphatic C

The cyclic aliphatic ring system can be any size above n = 3 (i.e. not aziridine). The ring system cannot be heterocyclic

Mechanism

P450 metabolism to a reactive iminium species has been suggested as a potential pathway to DNA adducts via an S_N1 mechanism (Kalgutkar et al 2005).

iminium ion (electrophile)

- Note that amides match the alert as well as amines
  - This is not mechanistically plausible
  - The category thus appears insufficiently restrictive

Similarity between sets of alerts

- Based on ChEMBL compound hit list
- Electrophile alert sets tend to resemble one another
- General filter sets also resemble one another, albeit slightly less so

	Glaxo	Lint	MLSMR	Dundee	Inpharmatica	BMS	Benigni-Bossa	OECD DNA-bind.	Cronin Skin Sens.	ALARM NMR
Glaxo		0.2	0.2	0.2	0.3	0.3	0.1	0.1	0.1	0.1
Lint	0.2		0.5	0.6	0.4	0.2	0.5	0.5	0.4	0.5
MLSMR	0.2	0.5		0.6	0.3	0.2	0.4	0.5	0.5	0.5
Dundee	0.2	0.6	0.6		0.3	0.2	0.3	0.5	0.4	0.5
Inpharmatica	0.3	0.4	0.3	0.3		0.3	0.3	0.2	0.3	0.3
BMS	0.3	0.2	0.2	0.2	0.3		0.1	0.1	0.1	0.1
Benigni-Bossa	0.1	0.5	0.4	0.3	0.3	0.1		0.4	0.3	0.4
OECD DNA-bind.	0.1	0.5	0.5	0.5	0.2	0.1	0.4		0.6	0.7
Cronin Skin Sens.	0.1	0.4	0.5	0.4	0.3	0.1	0.3	0.6		0.5
ALARM NMR	0.1	0.5	0.5	0.5	0.3	0.1	0.4	0.7	0.5	

- The issues of complementarity and redundancy between sets needs further investigation

Conclusions

- Several useful sets of Structural Alerts are available
  - Which are best probably depends on your compound collection
  - Some designed as filters, others to flag compounds for investigation
  - Need to choose sets of alerts carefully and check results
- Conversion to your platform many not be trivial
  - Free tools are available, so this step may not be necessary
- Lack of annotation an issue for the more general sets?
  - Lots of ‘traditional wisdom’ encoded, but rationale not always clear
- Newer mutagen/electrophile sets are the Gold Standard
  - Chemical mechanisms, example data & literature references
- New alerts will emerge as data becomes available
  - *e.g.* for toxicity endpoints other than carcinogenicity
    - However, MOAs will not be so straightforward as for genotoxicity
  - ‘Mitigating factors’ should be an interesting area for research

Additional Slides

Historical

- ‘Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. N.T.P.’, Mut. Res. 1991, **257**, 209-227 [and other refs in series].
  - Tennant & Ashby’s structural alerts for DNA reactivity based on analysis of *in-vitro* mutagenicity and *in-vivo* carcinogenicity data
- ‘Reactive compounds and *in vivo* false positives in HTS’, DDT 1997, **2**, 382-383.
  - Rishton’s compilation of reactive moieties that can interfere with biochemical assays. The chemical mechanisms by which this can occur overlap with those that which can cause ADRs or genotoxicity.
- These two sources were hugely influential and ancestral to several of the sets of S.A.s described below. The substructures described are fairly general, and would not be entirely straightforward to encode as filters for practical use. As they have been incorporated in other, more modern, alert sets their direct implementation was not attempted.
- ‘Filtering databases and chemical libraries’, JCAMD 2002, 16, 311-323.
  - Widely cited overview of Vertex’s ‘Rapid Elimination Of Swill’ (REOS) approach. Details of the actual S.A.s used are lacking, however.

Glaxo Wellcome ‘Hard Filters’ (1999)

- ‘Strategic Pooling of Compounds for High-Throughput Screening’ J. Chem. Inf. Comput. Sci. 1999, **39**, 897-902.
  - “A set of substructure search filters was used to remove compounds containing inappropriate functional groups. These comprise filters for reactive functional groups, unsuitable leads (*i.e.* compounds which would not be initially followed up), and unsuitable natural products (i.e., derivatives of natural product compounds known to interfere with common assay procedures).”
- SMARTS available as supplementary info (Daylight)
  - *N.B.* Includes SMARTS defining chemical functionalities that could react when pooled, which are not implemented here.
- *N.B.* Superceded within GSK, but still of interest.
- Also available *via* smartsfilter web app @ UNM.

Pfizer ‘Lint’ (1995-)

- ‘Identification and Evaluation of Molecular Properties Related to Preclinical Optimization and Clincal Fate’, Med. Chem. 2005, **1**, 649-655.
  - This paper cites ‘Lint: a Computational Procedure for Removing Problematic Functionality. Tripos Inc Sybyl Users Meeting, Princeton, NJ; May 1995’
  - Also see this presentation by JF Blake (Ex-Pfizer, now of Array BioPharma)
- Lint procedure originally implemented at Pfizer as an SPL script
- SLN translated to SMARTS (ChemAxon) by J. Yang of UNM and made available *via* smartsfilter web app.

BMS (2006)

- ‘An Empirical Process for the Design of High-Throughput Screening Deck Filters’ J. Chem. Inf. Model. 2006, **46**, 1060**-**1068.
  - “The FG filters that comprise this study are a combination of exclusion and informational filters. Exclusion FG filters are those intended for compound removal from screening decks. Informational filters are useful for compound annotation.”
    - Which of the SMARTS are ‘exclusion’ filters and which are ‘informational’ isn’t obvious from the publication
  - “These filters, like many others currently being applied within the industry for triage, are based on chemical intuition and experience (bias).”
- SMARTS available as supplementary info (Pipeline Pilot)

NIH MLSMR (2006)

- NIH Molecular Libraries Small Molecule Repository ‘Excluded Functionality Filters’
- No literature reference: see MLSMR Home Page for details.
- SMARTS extracted from PDF document (Daylight?)
  - The original filter set was relaxed somewhat (to facilitate compound acquisition) so some SMARTS are flagged as ‘Excluded’ and some ‘Allowed’.

University of Dundee (2008)

- ‘Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases’, ChemMedChem 2008, **3**, 435-444.
  - “Compounds containing unwanted functionalities were removed as it is not desirable to waste resources removing such functionalities in the hit optimization phase. These included potentially mutagenic groups such as nitro groups, groups likely to have unfavourable pharmacokinetic properties such as sulfates and phosphates; and reactive groups such as 2-halopyridines or thiols. Furthermore, compounds which are likely to interfere with typical HTS assays were also excluded.”
- SMARTS available (OpenEye)

Inpharmatica (2005)

- Filters used at Inpharmatica to identify ‘unwanted fragments’ (personal communication).
- SLN converted to SMARTS by hand.
- SLN file available on request.

Benigni/Bossa rulebase (2008)

- 'Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree' JRC Scientific and Technical Reports EUR 23241 EN – 2008 (*N.B.* also included with Toxtree)
- 'Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology' Chem. Rev. 2011, 10.1021/cr100222q
- Together, these two sources provide thorough documentation for the alerts, with mechanisms (for DNA-reactives) and example chemicals for each.
- Mainly SAs for DNA-reactivity, but includes a small number of alerts for non-genotoxic carcinogens.
- The SA are given graphically in the JRC document. However, the Java source contains the SMARTS (CDK).
- Some alerts are defined programatically, so SMARTS were coded to give as close a concordance with Toxtree as possible.
- Based on various historical sets of genotoxicity/carcinogenicity/mutagenicity SAs, which were thus not investigated further in any great detail.
- Some alerts take account of 'mitigating factors' that act to reduce DNA reactivity (steric hindrance, electronic deactivation, increased detoxification).
- Toxtree also includes QSARs for some classes.
- An implementation is also included in the QSAR Toolbox.

ChEMBL

OECD DNA-binding (2010)

- "A review of the electrophilic reaction chemistry involved in covalent DNA binding" Crit. Rev. Toxicol. 2010, 40. 728-748.
- "Report of the expert consultation on scientific and regulatory evaluation of organic chemistry mechanism-based structural alerts for the identification of DNA binding chemicals" OECD ENV/JM/MONO(2010)8/PART1 & PART2
- Also see HTML documentation supplied with QSAR Toolbox for most up-to-date & accurate info on the S.A.s
- SMARTS available from authors: however, some minor differences mean substructures defined in Toolbox configuration files were taken as definitive.
- Alerts focus on mechanistic chemistry of DNA modification, with substructures designed to group chemicals into seven 'mechanistic domains' (Michael addition, S_N1, S_N2 *etc*.) as a prelude to QSAR / read-across.
- 'Alerts' thus perhaps more general than BB, as they are not designed to be used as 'hard' filters; some are very general (*e.g.* 'arenes').
- High-throughput use-case would be to identify cases that might require more detailed investigation.

Cronin Skin Sensitisation (2008)

- ‘Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach’, SAR QSAR Environ. Res. 2008, **19**, 555-578.
- SMARTS patterns define electrophilic ‘reactivity domains’ of relevence to binding to protein nucleophiles. Predictivity tested with mouse LLNA data.
- SMARTS included in paper [extacted from PDF] (PerlMol).

ALARM NMR (2004)

- 'ALARM NMR: A Rapid and Robust Experimental Method To Detect Reactive False Positives in Biochemical Screens', J. Am. Chem. Soc. 2005, **127**, 217-224.
- Describes experimental technique to identify compounds that react with 'an exceptionally reactive' cysteine in the La protein.
- "On the basis of the compound reactivity profiles that have been observed, we have identified chemical substructures that are prone to being thiol-reactive. These substructures can be included in filtering protocols to identify potential thiol-reactive compounds in silico."
- SMARTS available as supplementary info (dialect not specified).
- Also available *via* smartsfilter web app @ UNM.
- Validation reported in: 'Toxicological Evaluation of Thiol-Reactive Compounds Identified Using a La Assay To Detect Reactive Molecules by Nuclear Magnetic Resonance', Chem. Res. Toxicol. 2007, **20**, 1752–1759.

PAINS (2010)

- 'New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays' J. Med. Chem. 2010, **53**, 2719-2740.
- "This report describes a number of substructural features which can help to identify compounds that appear as frequent hitters (promiscuous compounds) in many biochemical high throughput screens. The compounds identified by such substructural features are not recognized by filters commonly used to identify reactive compounds"
- SLN available as supplementary info.
  - Not yet translated to SMARTS

SMARTCyp (2010)

- ‘SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism’, ACS Med. Chem. Lett. 2010, **1**, 96-100.
- “SMARTCyp is an in silico method that predicts the sites of cytochrome P450-mediated metabolism of druglike molecules. The method is foremost a reactivity model, and as such, it shows a preference for predicting sites that are metabolized by the cytochrome P450 3A4 isoform. SMARTCyp predicts the site of metabolism directly from the 2D structure of a molecule, without requiring calculation of electronic properties or generation of 3D structures.”
- Combines SMARTS (OpenEye) patterns for identification of metabolically vulnerable positions, a reactivity descriptor for each and an ‘accessibiliy descriptor’ (calculated programmatically).
- A Pipeline Pilot protocol implementing the algorithm (N. Malcolm) has been made available *via* the Accelrys Community website.

NIH CYP inhibition HTS (2009)

- ‘Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries’, Nat. Biotechnol. 2009, **27**, 1050-1582.
- “We determined potency values for 17,143 compounds against five recombinant CYP isozymes (1A2, 2C9, 2C19, 2D6 and 3A4) using an in vitro bioluminescent assay. The compounds included libraries of US Food and Drug Administration (FDA)-approved drugs and screening libraries. We observed cross-library isozyme inhibition (30–78%) with important differences between libraries.”
- Supporting info contains an analysis of the data showing the propensity of ~300 to inhibit the various isoforms.
- These could form the basis of a set of alerts for CYP inhibition, though as they are provided as SMILES not SMARTS some further work might be necessary to achieve an acceptable degree of specicifity.

TCSA NCP Categories (2010)

- 'TCSA New Chemicals Program (NCP) Chemical Categories' published by the Office of Pollution Prevention and Toxics of the EPA.
- "The categories included in this compilation represent chemicals for which sufficient assessment experience has been accumulated so that hazard concerns and testing recommendations vary little from chemical to chemical within the category. Thus, these categories do not necessarily represent the chemicals of greatest concern to the Agency. By the same token, the categories are also not intended to be a comprehensive list of all substances that may be subject to further action in the New Chemicals Program."
- Substructures are defined pictorially, and the translations to SMARTS sometimes not obvious; concordance with ECOSAR attempted where possible.
- *N.B.* The QSAR Toolbox contains an implementation which could also be used for verification.
- Alerts without substructural definitions were not included.
- Alerts generally include a physicochemical component (often as a mitigating factor) which have not been addressed.
- See 'Ranking and prioritization of environmental risks of pharmaceuticals in surface waters' Regul. Toxicol. Pharmacol. 2004, **39**, 158-183 for an application of ECOSAR to pharmaceuticals.

Some references referring to SAs but without including SMARTS (or equivalent)

- 'Components of Successful Lead Generation', Curr. Top. Med. Chem. 2005, **5**, 421-439. (AstraZeneca)
- 'Assessment of chemical libraries for their druggability', Comp. Biol. Chem. 2005, **29**, 55-67.
- 'Managing, profiling and analysing a library of 2.6 million compounds gathered from 32 chemical providers', Mol. Div. 2006, **10**, 389-403.
- 'Leadlikeness and structural diversity of synthetic screening libraries', Mol. Div. 2006, **10**, 377-388.
- 'Analysis and hit filtering of a very large library of compounds screened against Mycobacterium tuberculosis', Mol. Biosyst. 10.1039/C0MB00104J.
- 'Structure-Activity relationships for *In vitro* and *In vivo* Toxicity', Ann. Rep. Med. Chem. 2006, **41**, 353-368.
- 'The Identification of Toxicophores for the Prediction of Mutagenicity, Hepatotoxicity and Cardiotoxicity', J. Iran. Chem. Soc. 2005, **2**, 244-267.
- 'The use of structure-activity relationship analysis in the food contact notification program', Reg. Toxicol. Pharmacol. 2005, **42**, 225-235.
- 'Discriminating toxicant classes by mode of action: 3. Substructure indicators' 2007, **18**, 155-168.

Not SAs as such, but useful for reference…

- ‘Cytochrome P450 Enzymes Mechanism Based Inhibitors: Common Sub-Structures and Reactivity’
  - Curr. Drug Metabol. 2005, 6, 413.
  - Compilation of compounds known to irreversibly inactivate CYPs, with reacting moiety indicated where known. Structures and references included.
- ‘Cytochromes P450: A Structure-Based Summary of Biotranformations using Representative Substrates’
  - Drug Metabol. Rev. 2008, **40**, 1-100.
  - Compilation of compounds known to be substrates of P450s with sites of metabolism indicated. Structures and references included.
- ‘Minimising the potential for metabolic activation in drug discovery’
  - Expert Opin. Drug Metab. Toxicol. 2005, **1**, 91-142.
  - Contains table of functional groups vulnerable to metabolic activation, and of examples of susceptible drugs.
  - Companion to ‘A Comprehensive Listing of Bioactivation pathways of Organic Functional Groups’, Curr. Drug Metabol. 2005, **6**, 161-225.

Other Relevant Review Articles

- ‘False positives in the early Stages of Drug Discovery’
  - Curr. Med. Chem. 2010, **17**, 4231-4255.
- ‘Towards a Comprehensive Molecular Design Framework for reduced Hazard’
  - Chem. Rev. 2010, 110, 5845-5882.