Predicting interactions of compounds and metabolites with toxicity-associated targets

Peter Hunt*, Matthew Segall, Francis Atkinson – ***EMBL-EBI***, **Ines Smit** – ***EMBL-EBI***

ACS National Meeting

Philadelphia – 24th August 2016

Introduction To Hecatos

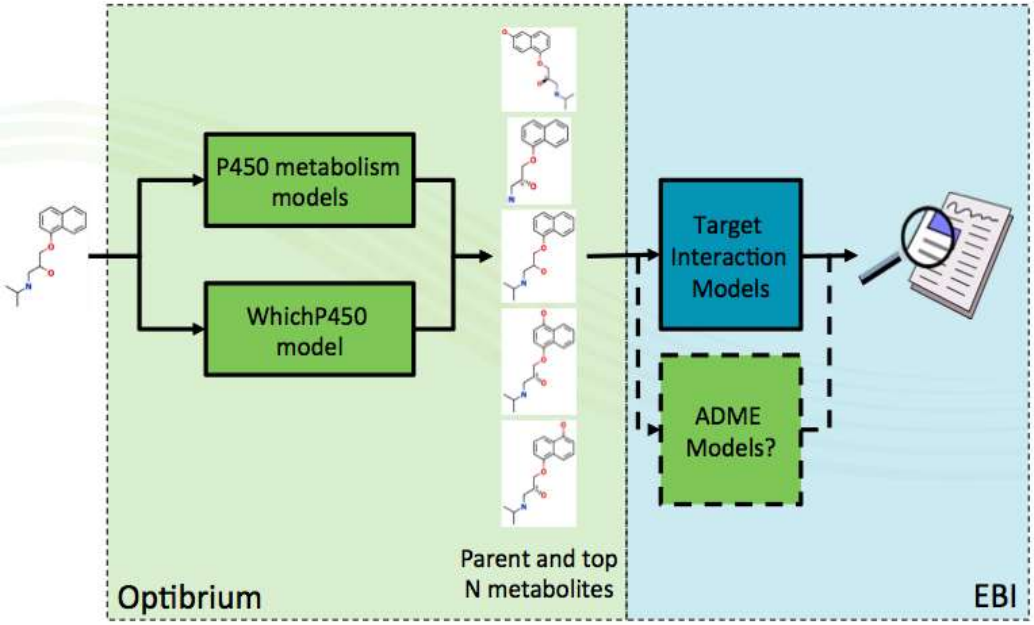
HeCaToS *aims at developing integrative 'in silico' tools for predicting human liver and heart toxicity. The objective is to develop an integrated modeling framework, by combining advances in computational chemistry and systems toxicology, for modelling toxic perturbations in liver and heart across multiple scales.*

http://www.hecatos.eu/

- European consortium organised as a private-public partnership
- Directed at **He**patic and **Ca**rdio **To**xicity **S**ystems modelling
- Linking *in silico*, 3D *in vitro*, 'omic and clinical data for better chemical safety testing

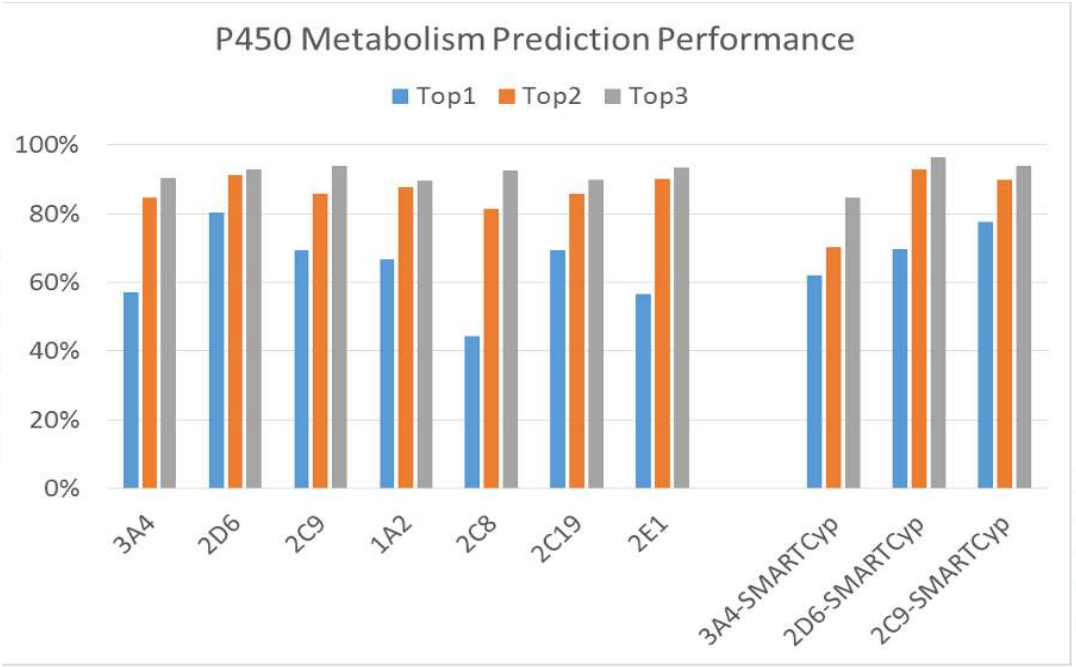
Integration Efforts For Toxicity Targets

- EMBL-EBI tasked with using the literature to identify a Cardio- and Hepato- related toxicity panel
- Integrate this panel into a workflow
  - Predict the metabolism hotspots
  - Suggest the likely Phase 1 metabolites
  - Predict the activity profile of parent ***plus*** likely metabolites
  - Indicate the toxicity potential based on that profile

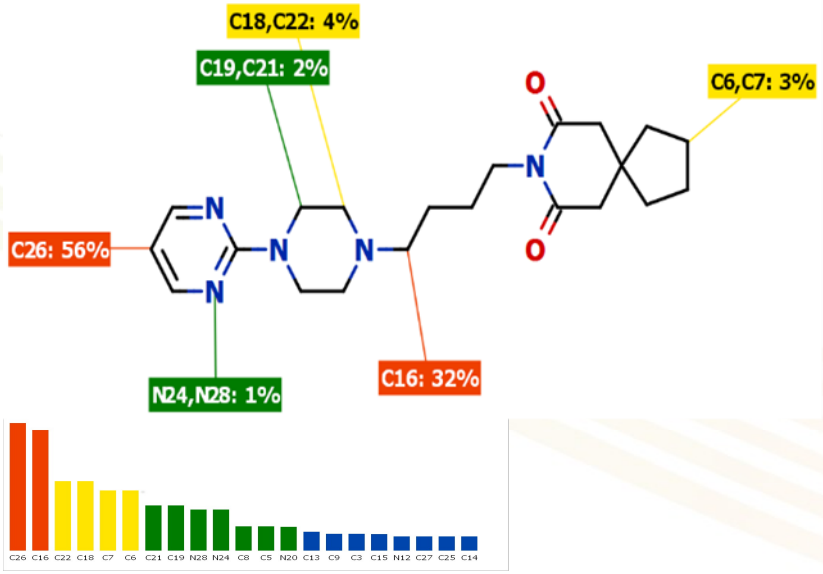


StarDrop P450 Regioselectivity Models

- A QM, AM1-based, energetic approach with correction factors
  - for systematic AM1 errors
  - for isoform specific steric/orientation effects
- Covers the leading 7 Human isoforms
- Good Top-2 predictive performance
  - "Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism using Quantum Mechanical Simulations" – preprint will be available from our community website http://www.optibrium.com/community/publications/in-silico-modelling/348-preprintp450metabolism
  - Assessment also made by using the Chemical Lift metric as proposed by Breneman et al for RS-Predictor - *J. Chem. Inf. Model.*, 2011, **51** (7), pp 1667–1689, **DOI:** 10.1021/ci2000488



StarDrop P450 Regioselectivity Models



	Site	atio %	Lability	Metabolites
1	C26	56	Labile	
2	C16	32	Labile	
3	C18, C22	4	Mod Labile	

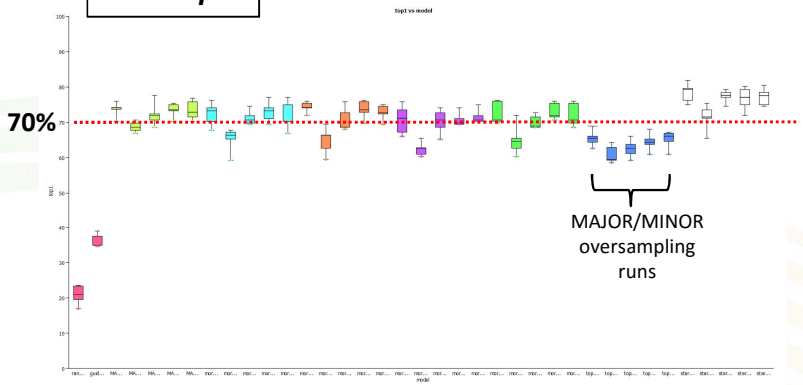
- The predictions show the percentage of metabolism to occur with that isoform at each site
  - Should the compound be a substrate for that isoform
- Suggestions for metabolites based on each centre are made using SMIRKs
  - NB: only Phase 1 metabolites currently
  - Conceptually one could suggest likely Phase 2 metabolites structures in a similar way

WhichP450 Models

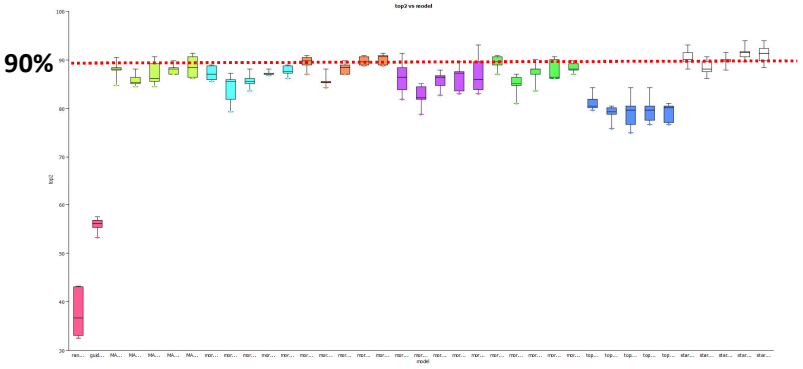
- Knowledge of the regiospecificity is only half the story
- Utilised the literature data collected for our Regioselectivity models
  - Annotated those molecules with not only where but by which CYP and a *personal* judgement of MAJOR vs MINOR metaboliser
- Total number of molecules used in this work is 484 with a 196 molecule/isoform test set
  - A molecule, and site, can be a substrate (major or minor) for more than one isoform,
- A Random Forest based multi-class model produces an ordered list of probabilities for all 7 isoforms
- Measure how successful a prediction is by seeing if a MAJOR isoform is one of the top-k predictions
  - 140 unique compounds in test set
  - Reporting the % of the 196 molecule/isoform test set is predicted correctly
  - Repeated for 5 different training/test splits of the data set
- Also calculated a “random” prediction and a “guided random” biased by the known proportion of compounds that each isoform metabolises
  - **3A4** = 0.345, **2D6** = 0.230, **2C9** = 0.149, **1A2** = 0.103, **2C19** = 0.08, **2C8** = 0.057, **2E1** = 0.036

WhichP450 Models

RF – Top 1



RF – Top 2



- RF methodology preferred due to
  - less variability,
  - less influence of the MINOR isoform information,
  - less influence on the descriptors used (esp. in a 256 bit fingerprint),
  - less influence due to dataset column or row ordering
- *Previous results were presented at the spring 2016 ACS meeting in San Diego*

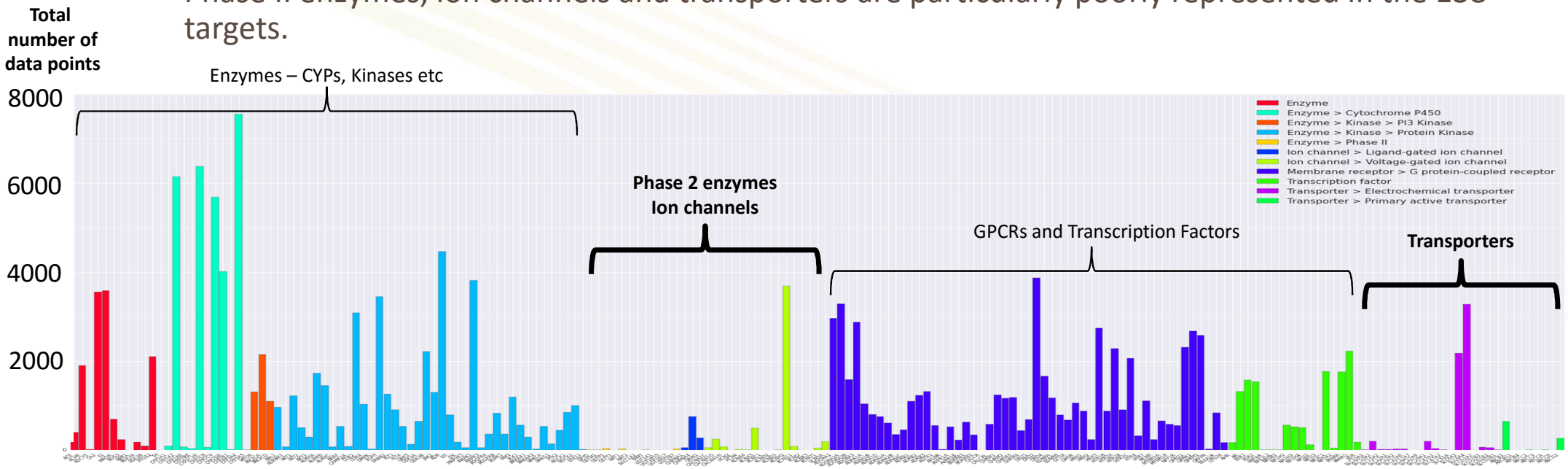
Random & Guided Random
Morgan Radius 3
MACCS
Morgan features Radius 3
Morgan Radius 2
RDKit topological
Morgan features Radius 2
StarDrop fingerprint

Toxicity Target Panel
Creation

- Toxicity target list compiled from a trawl of literature & screening lists at CRO's
- List of 215 number of targets –
  - Targets mapped to ChEMBL data via HUGO gene symbols
  - Protein family data where isoform is not specified or protein complexes where the binding target is unknown were not incorporated
  - 20 targets were not found in ChEMBL (Na/K pump, cardiac ion channels & transporters)
  - 18 of remaining 195 had multiple ChEMBL entries eg CDK2, CDK2/Cyclin A, CDK2/Cyclin E etc – highlighted the duplication of data in the lit by mislabelling
- Diversification of modelling approaches taken by EBI and Optibrium
  - Optibrium take the quantitative and EBI take the qualitative approach
  - Needed to be easy to update

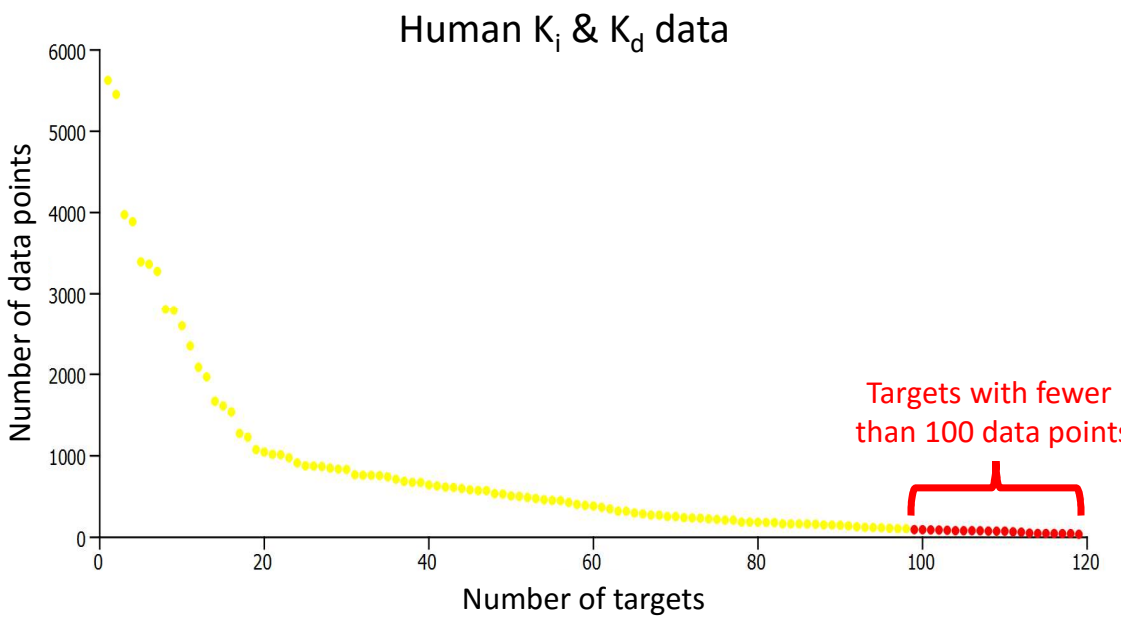
ChEMBL Data Summary For The Targets

- Graph of targets grouped by type vs *total* number of data points
  - Size of compound applied (<50 heavy atoms), actives are 10uM or better
  - If a minimum data set size of 40 is applied then only 138 targets remain
  - Phase II enzymes, ion channels and transporters are particularly poorly represented in the 138 targets.



ChEMBL Data Summary For Quantitative Modelling

- How does that data look when you start triaging for only IC_{50} or K_i..?
  - Greater constraints placed on the quality of the data



- NB the graph shows the *total* number of data points in ChEMBL 20,
  - that are non qualified,
  - Human binding data only

  compared to
  - the MCNB method using all Human data <10uM from all assays

- These do not represent number of *unique* compounds.
  - IC_{50} data has a similar distribution
  - 109 targets are in common and only 88 have >=100 data points in both

Current Quantitative Model Panel

GPCR Target name	
Dopamine D2 receptor	*Alpha-1a adrenergic receptor*
Adenosine A3 receptor	Serotonin 7 (5-HT7) receptor
Adenosine A2a receptor	Alpha-1b adrenergic receptor
Adenosine A1 receptor	Muscarinic acetylcholine receptor M2
Histamine H3 receptor	**Alpha-1d adrenergic receptor**
Delta opioid receptor	**Histamine H1 receptor**
Mu opioid receptor	Dopamine D1 receptor
Kappa opioid receptor	Muscarinic acetylcholine receptor M1
Serotonin 2a (5-HT2a) receptor	*Muscarinic acetylcholine receptor M3*
Dopamine D4 receptor	**Adenosine A2b receptor**
Serotonin 2c (5-HT2c) receptor	

The **bold** targets have R^2 values above 0.6 and the *italicised* targets have R^2 values below 0.5

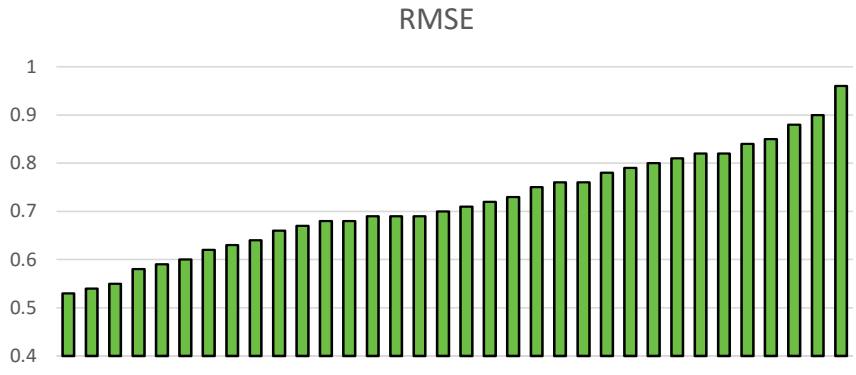
- 34 targets in the panel
  - linear regression,
  - 2-category (split at pK_i of 7.5)
  - 3-category (split at pK_i of 6.5 & 8.5)
- Note the over representation of GPCRs

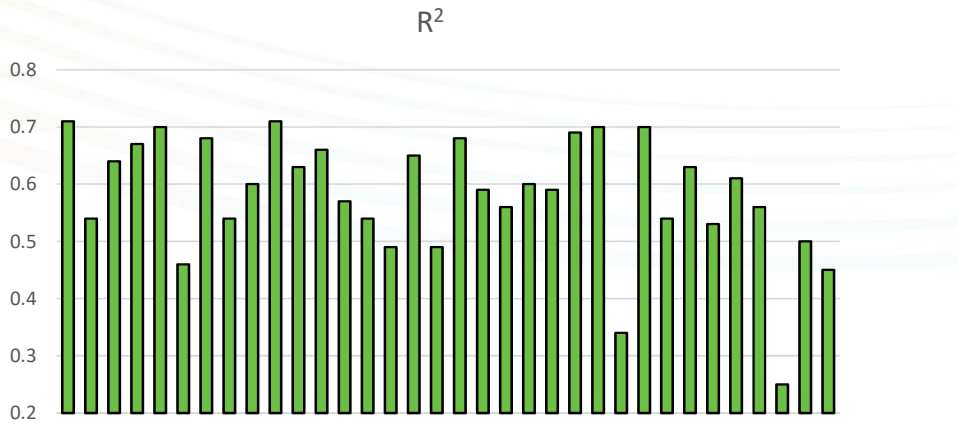
Other Target name
Carbonic anhydrase II
Thrombin
Serotonin transporter
Norepinephrine transporter
Glucocorticoid receptor
Androgen Receptor
HERG
Neuronal acetylcholine receptor; alpha4/beta2
Neuronal acetylcholine receptor protein alpha-4 subunit combined
Glutamate NMDA receptor; GRIN1/GRIN2B
ABL Kinase
Aurora-B Kinase
Glycogen synthase kinase-3 beta

Current Model Panel

Summary stat's

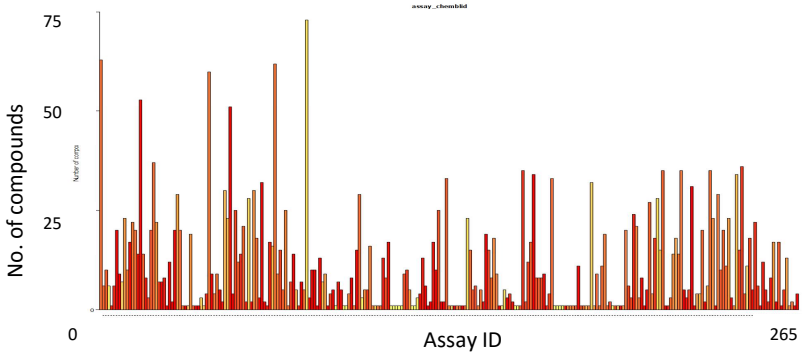
- All regression models for pK_i data have RMSE values below 1 log unit
- The statistical success of the models is surprisingly good considering the mixing of the data
  - Note these R^2 are not best fit lines but from lines of equivalence rooted through the origin
  - In general the R^2 values for the models built on IC_{50} values were lower & RMSE higher
- Attempts to improve the models by more local modelling were attempted
  - Eg Adenosine A3



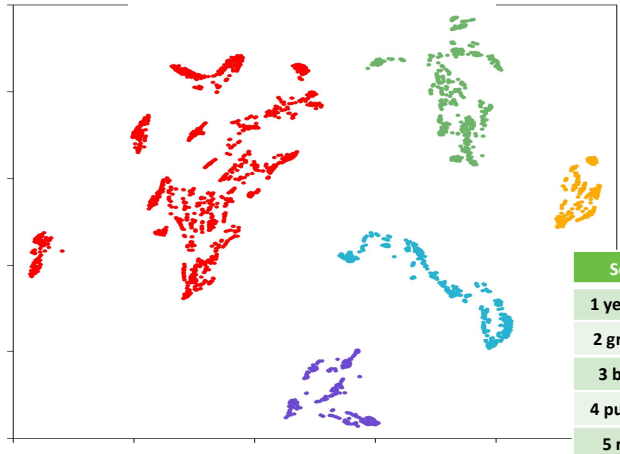


Adenosine A3 Data Set Modelling

- The original ChEMBL pK_i data set had 3740 rows which was de-duplicated to leave 2788 unique compounds
- Remove the highly variable data points
  - Original work had the most active value retained (ie worse case scenario)
  - New work removed all points if the results differed by >0.5 log units
- New de-duplicated set had 2745 unique compounds and highly similar results
  - R^2 = 0.63; RMSE = 0.65
- Produce a 'local' model with a single assay..?
  - Over 260 "assays" in this set
  - The large numbers of "different" assays mean amalgamation of data is inevitable.
- Produce models for local subsets of the data..?
  - Only partially successful as some models performed worse than the original model
    - Set 3, R^2 of 0.74, RMSE of 0.55
    - Set 2, R^2 of 0.75, RMSE of 0.53
  - Each local model was too local to generalise



Adenosine ChemSpace



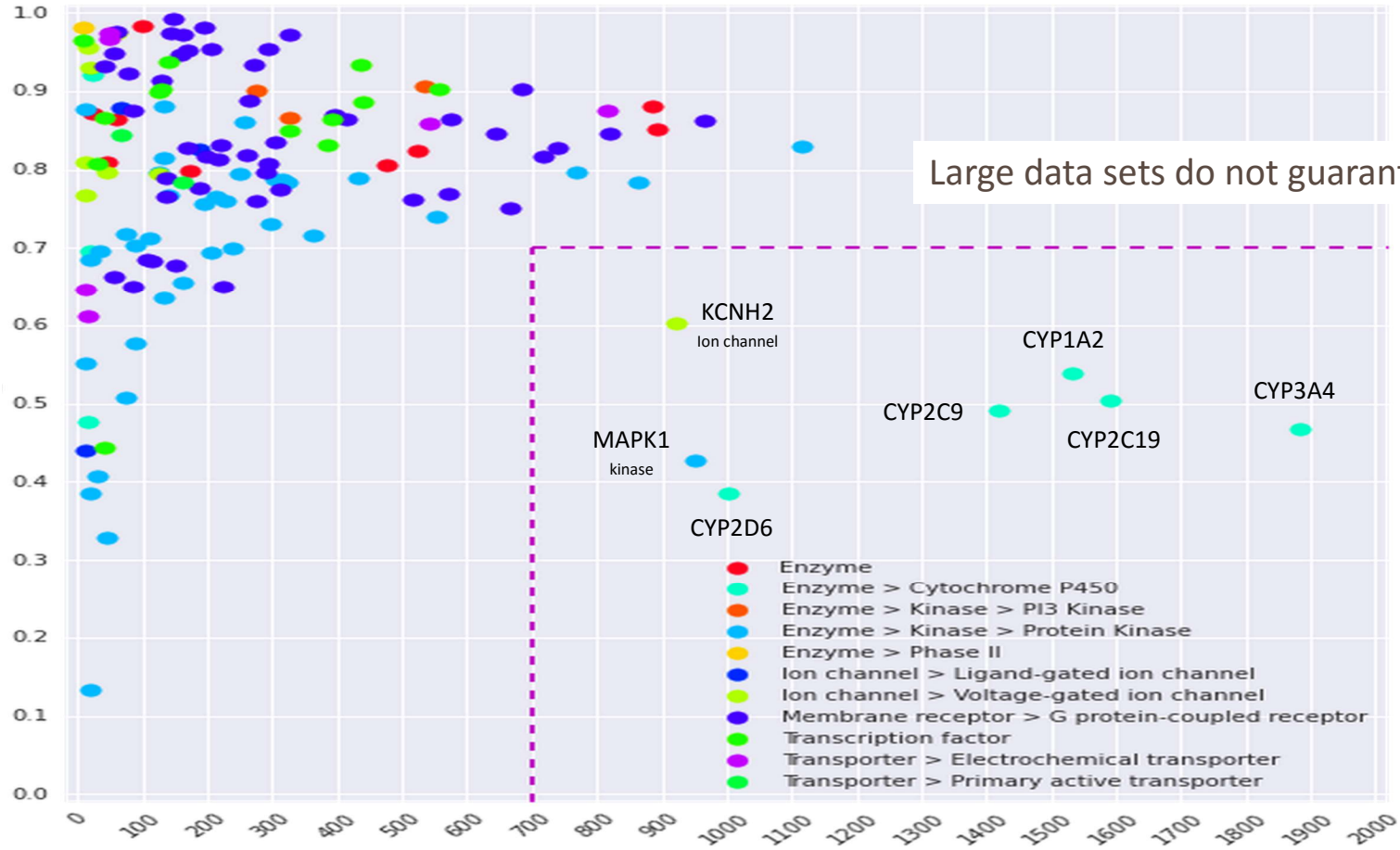
Set	Size
1 yellow	221
2 green	540
3 blue	431
4 purple	299
5 red	1254

EBI – Naive Bayes Modelling

- Use ECFP4 radial descriptors with Multi-Class Naïve Bayes (MCNB) methodology to produce a profile of active/inactive predictions
- Data set used is the combined actives from all targets
- Split of training:test is 75:25% and the modelling is run with 10 different splits
- Activity at a target is a molecule with any activity <10uM
- Any known inactive or unknown activity is assumed to be inactive

EBI – Naive Bayes Modelling

Mean predictive success (over the 10 runs) for known actives

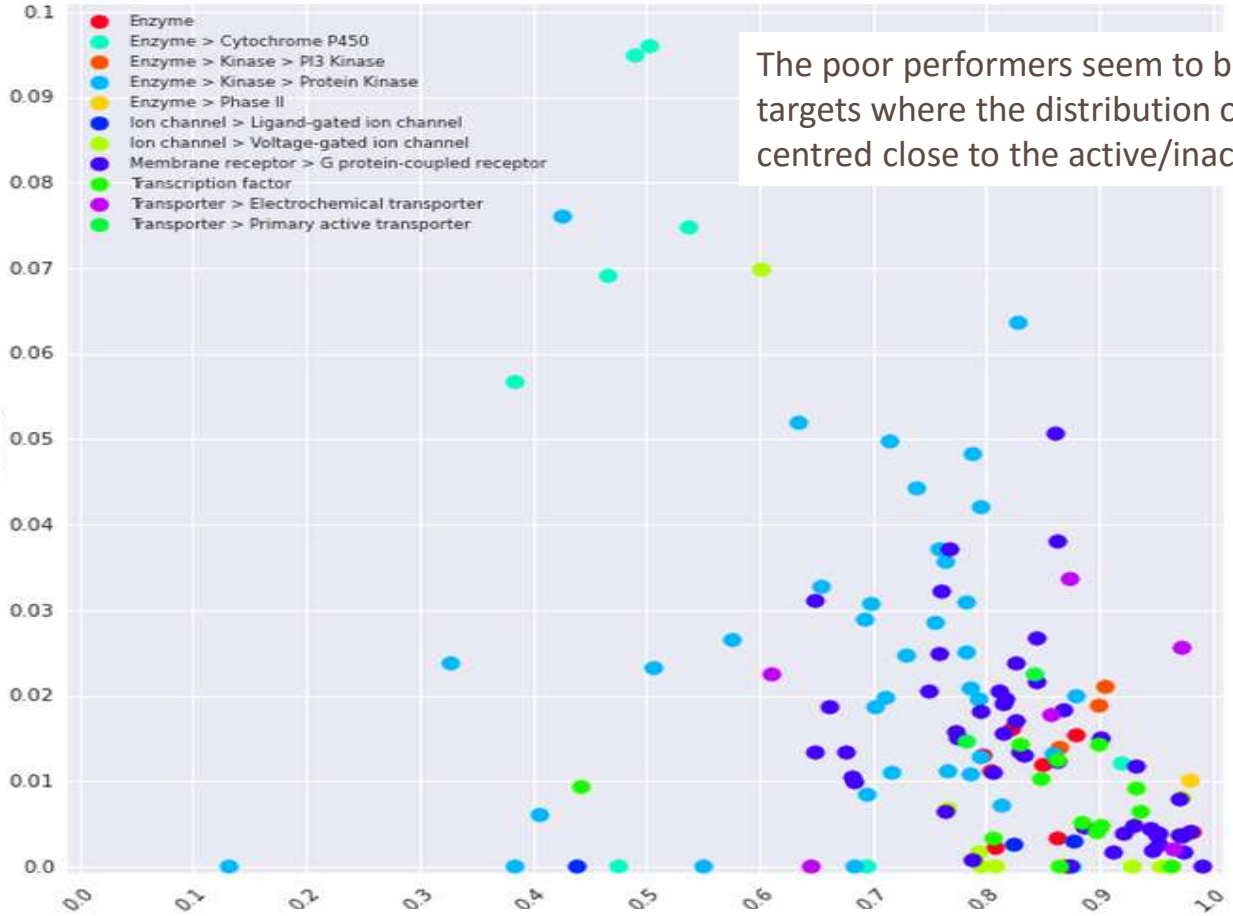


Large data sets do not guarantee success

Number of actives in the target test set

EBI – Naive Bayes Modelling

Mean false prediction for known inactives

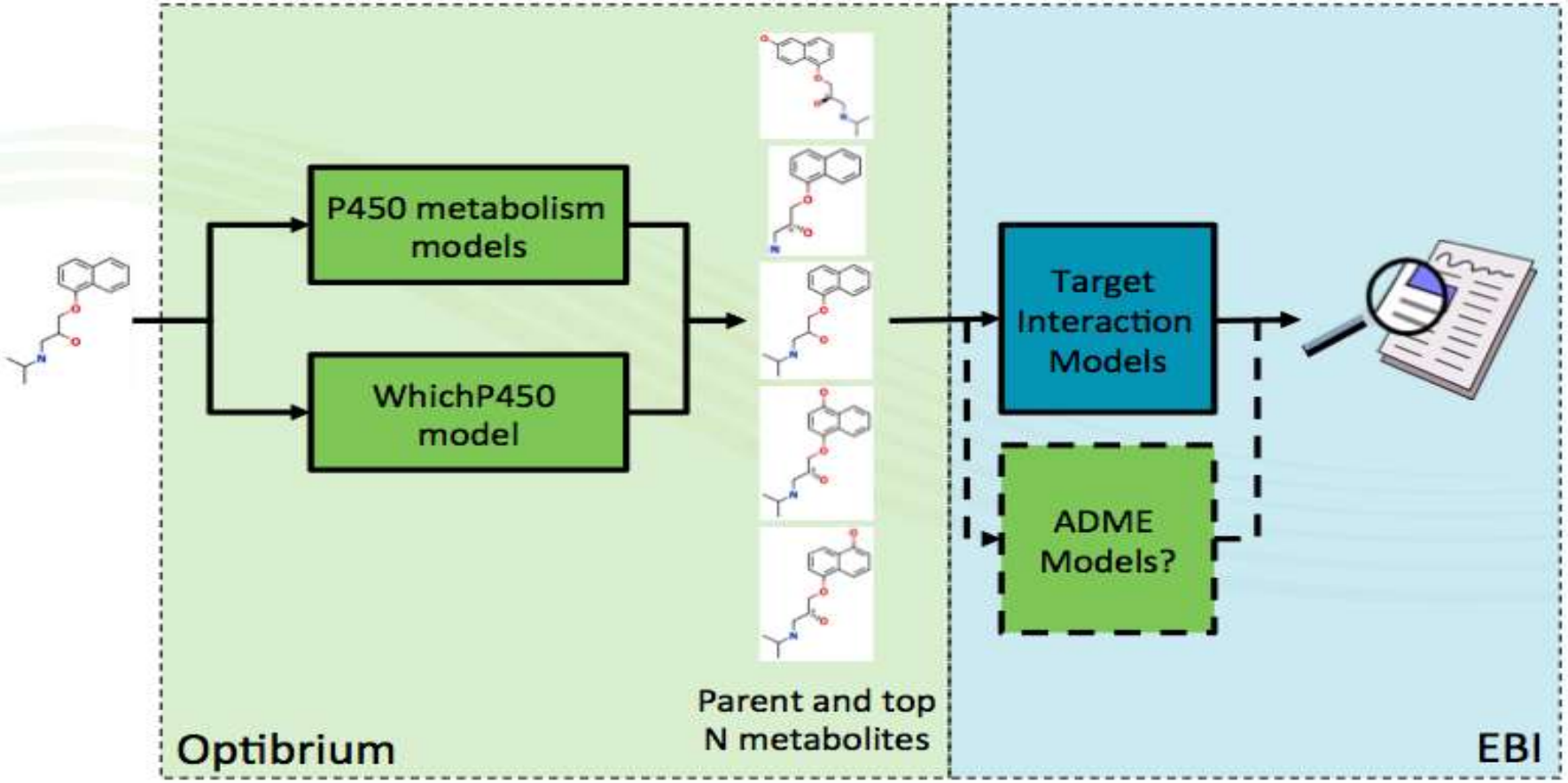


The poor performers seem to be those counter targets where the distribution of activities is more centred close to the active/inactive boundary.

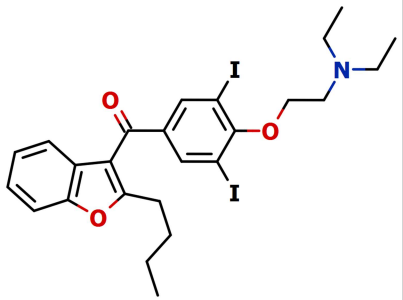
Note the different scales on the axes

Mean predictive success for known actives

Integration



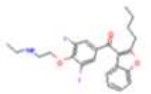
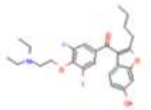
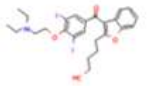
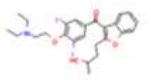
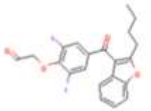
Current Output



Amiodarone

- HUGO gene symbols for those parent/metabolites believed to be active are returned

- Metabolite suggestions also include the other half of any change; like N-dealkylation to produce an aldehyde which can be flagged

	CYP2D6_R	CYP3A4_R	activities	alert	lability	other_smiles	smiles	mol	other_mol
0	45.804700	49.389600		True	bile	CC=O	CCCCc1oc2ccccc2c1C(=O)c1cc(I)c(OCCNCC)c(I)c1		
1	4.038930	0.158629	ACHE;CA2;CYP2C8;ESR1;ESR2	False	Mod Labile	None	CCCCc1oc2cc(O)ccc2c1C(=O)c1cc(I)c(OCCN(CC)CC)c(I)c1		None
2	1.049790	0.294910	CA2	False	Mod Labile	None	CCN(CC)CCOc1c(I)cc(C(=O)c2c(CCCCO)oc3ccccc23)cc1I		None
3	0.281715	0.218902	ADRA2B	False	Mod Labile	None	CCN(CC)CCOc1c(I)cc(C(=O)c2c(CCC(C)O)oc3ccccc23)cc1I		None
4	0.089519	0.342148		True	Mod Labile	CCNCC	CCCCc1oc2ccccc2c1C(=O)c1cc(I)c(OCC=O)c(I)c1		

Future Improvements

- P450 regioselectivity :
  - more isoforms ***including plant CYPs***
  - more pathways covered {eg dehalogenations}
- Expansion of non-CYP metabolism :
  - better suggestion of metabolites formed {more than one step considered}
  - Phase II metabolism considered
  - other mechanisms considered {Aldehyde Oxidase, FMO's, Reductases etc}
- WhichP450 :
  - production implementation
  - more isoforms
- Toxicity targets :
  - more relevance {linking to AOPs, cf ChEMBL vs in-house data}
  - Optibrium - more targets {if the data allows}
  - EBI - creation of target specific models rather than MCNB {in order to use more of the definite inactives rather than presumed inactives}

Acknowledgements

- Matthew Segall, Jon Tyzack, & Rasmus Leth
- Francis Atkinson, Ines Smit

- Hepatic and Cardio Toxicity Systems modelling
  - The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under the grant agreement no 602156