Predicting Adverse Drug Reactions Using Drug-Protein Binding Data, Small Molecule Properties, Pathways, and High-Performance Computing Molecular Docking Studies

Montiago LaBute, Jason Lenderman, Xiaohua Zhang

- 05/22/13

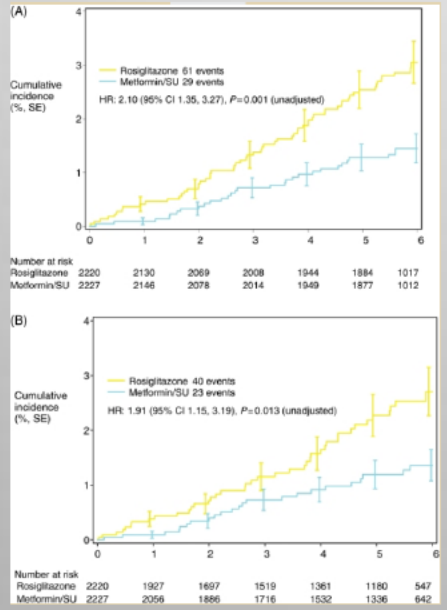
Global Security Principal Directorate

LLNL-PRES-637212X
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Adverse Drug Reactions Are a Serious National Public Health Issue

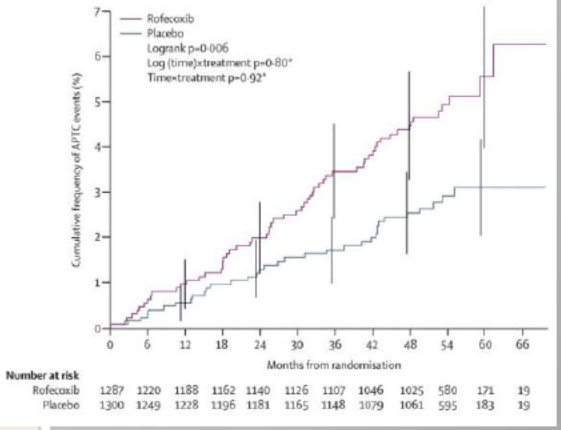
- Adverse drug reactions (ADRs) are rare and complex perturbations of biological pathways by pharmacologically-active small molecules
- ADRs cause 100,000 fatalities and an associated public health cost of $136 billion dollars
  - Comparable to care of cardiovascular disorders and diabetics
- Billions of R&D dollars wasted by pharma companies as drugs present with unexpected ADRs post market or in late stage development (e.g. Avandia and Vioxx)

Image from www.schlesingerlaw.com



Komajda, et al., "Heart failure events with rosiglitazone in type 2 diabetes: data from the RECORD clinical trial" (2010)

http://www.brain-injury-law-center.com/practice-areas/prescription-drugs.html



Baron, et al., "Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial" (2008)

http://pubs.acs.org/cen/coverstory/83/8325/8325vioxx.html

Accurate ADR Prediction using *in Silico* models would have tremendous upside

- Ideally applied before expensive in vitro testing or clinical trials to identify severe ADRs based mainly on what we know about the drug (e.g. protein binding, small molecule) properties,etc.)
- Challenge is to create models to a sufficient accuracy level such that predictions can be trusted; but there are many issues
  - Public data on drugs is small and number of predictors easily exceeds number of training examples
  - ADRs are typically in <10% of the drugs so the classes are highly skewed
  - Publicly available data is small and highly biased towards drugs that made it completely through pipeline w/out severe ADRs
  - Failures may be highly informative, but most likely reside in proprietary databases in pharma companies
  - Useful, if based only on mainly small molecule properties, but these might not be the most predictive variables
  - May need to go up to systems level to get most important predictors
    - Protein:protein interactions
    - Pathways involving interacting protein constituents
    - Phenotypes? Problem is that we would like to make predictions in the absence of trials
  - Should outcomes be aggregated to increase signal strength and mitigate skewed classes issue? How should we do this?

Publicly Available Data To Build Statistical Models (Predictors)

- DrugBank (available University of Alberta, 1480 FDA-approved small molecule drugs
  - Drug-protein associations
    - Targets (if available)
    - Enzymes
    - Transporters/Carriers

DRUGBANK
Open Data Drug & Drug Target Database

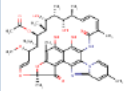
Home | Browse | Search | Downloads | News & Updates | About | Help | Co

Search: Search DrugBank | Search | Help / Advanced

Identification | Taxonomy | Pharmacology | Pharmacoeconomics | Properties | References | Interactions | 0 Comme

targets (2) | enzymes (1) | Show Drugs with Similar Structures for All

Identification

Name	**Rifaximin**
Accession Number	**DB01220** (APRD01218)
Type	small molecule
Groups	approved
Description	Rifaximin is a semisynthetic, rifamycin-based non-systemic antibiotic, meaning that the drug will not pass the gastrointestinal wall into the circulation as is common for other types of orally administered antibiotics. It is used to treat diarrhea caused by E. coli.
Structure	 Download: MOL | SDF | SMILES | InChI Display: 2D Structure | 3D Structure
	Rifamixin

Targets

1. DNA-directed RNA polymerase beta chain

Pharmacological action: **yes**
Actions: **inhibitor**

DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates

Organism class: **bacterial**
UniProt ID: P0ABV2
Gene: rpoB
Protein Sequence: FASTA
Gene Sequence: FASTA
SNPs: SNPJam Report

References:

1. Overington JP, Al-Lazikani B, Hopkins AL: How many drug targets are there? Nat Rev Drug Discov. 2006 Dec;5(12):993-6. Pubmed
2. Imming P, Sinning C, Meyer A: Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov. 2006 Oct;5(10):821-34. Pubmed
3. Vitali B, Turroni S, Dal Piaz F, Candela M, Wasinger V, Brigidi P: Genetic and proteomic characterization of rifaximin resistance in Bifidobacterium infantis BI07. Res Microbiol. 2007 May;158(4):355-62. Epub 2007 Feb 22. Pubmed
4. Ojetti V, Lauritano EC, Barbaro F, Migneco A, Alnora ME, Fontana L, Gabrielli M, Gasbarrini A: Rifaximin pharmacology and clinical implications. Expert Opin Drug Metab Toxicol. 2009 Jun;5(6):675-82. Pubmed
5. Koo HL, Dupont HL, Huang DB: The role of rifaximin in the treatment and chemoprophylaxis of travelers' diarrhea. Ther Clin Risk Manag. 2009;5:841-8. Epub 2009 Nov 2. Pubmed
6. Scarpignato C, Pelosini I: Experimental and clinical pharmacology of rifaximin, a gastrointestinal selective antibiotic. Digestion. 2006;73 Suppl 1:13-27. Epub 2006 Feb 8. Pubmed
7. Pimentel M: Review of rifaximin as treatment for SIBO and IBS. Expert Opin Investig Drugs. 2009 Mar;18(3):349-58. Pubmed
8. Chen X, Ji ZL, Chen YZ: TTD: Therapeutic Target Database. Nucleic Acids Res. 2002 Jan 1;30(1):412-5. Pubmed

2. Orphan nuclear receptor PXR

Pharmacological action: **no**
Actions: **agonist**

Publicly Available Data To Build Statistical Models (Predictors)

- PubChem
  - NCBI database of small molecules and their activities associated with biological assays
  - We use chemical substructure fingerprint data
    - 881-bit feature vector that characterizes the small molecule by indicating the presence ('1') or absence ('0') of particular atoms or chemical structural motifs

PubChem Substructure Fingerprint Description

Section 1: Hierarchic Element Counts - These bits test for the presence or count of individual chemical atoms represented by their atomic symbol.

Bit Position	Bit Substructure
0	>= 4 H
1	>= 8 H
2	>= 16 H
3	>= 32 H
4	>= 1 Li
5	>= 2 Li
6	>= 1 B
7	>= 2 B
8	>= 4 B
9	>= 2 C
10	>= 4 C
11	>= 8 C
12	>= 16 C
13	>= 32 C
14	>= 1 N

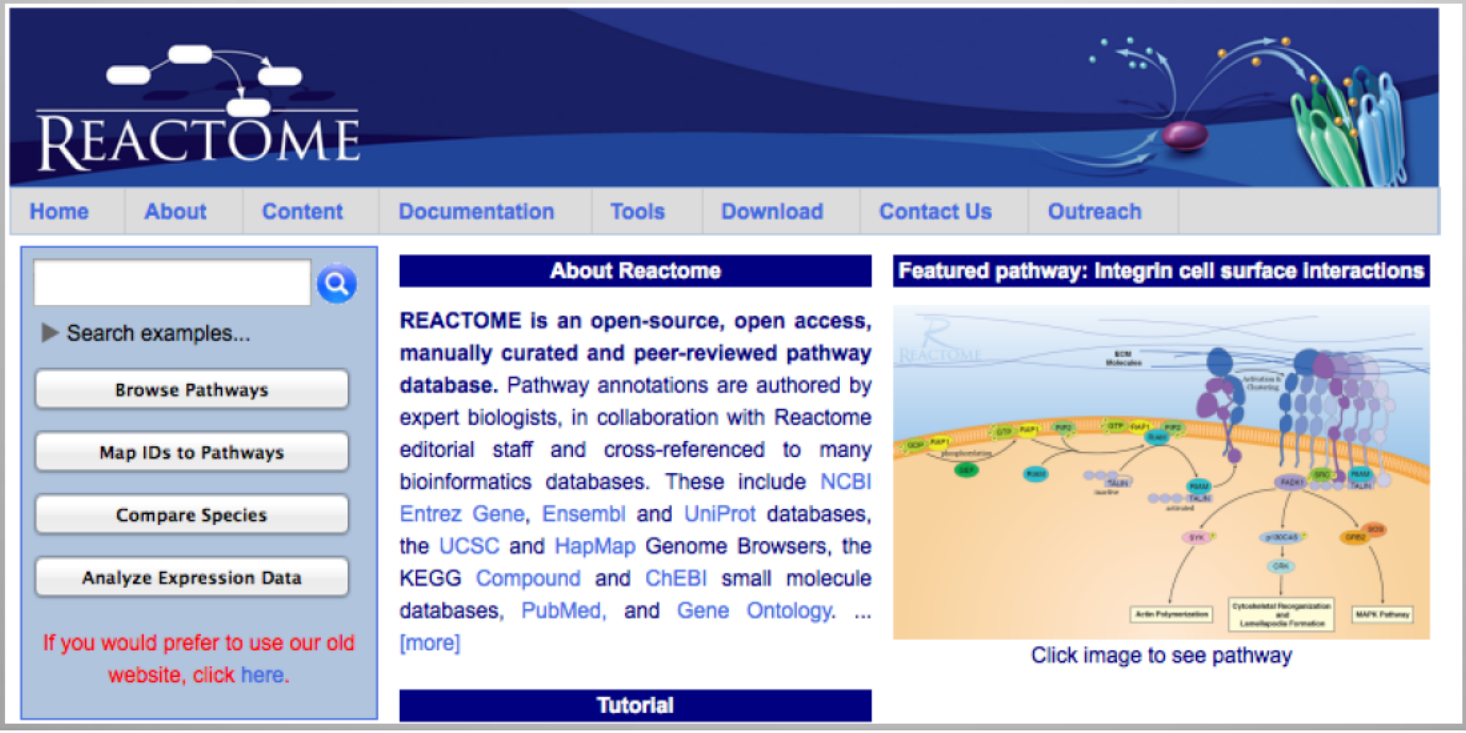
PubChem Substructure Fingerprint Description (cont.)

Section 6: Simple SMARTS patterns (cont.)

Bit Position	Bit Substructure
510	[#1]-N-N-[#1]
511	S=C-N-[#1]
512	C-[As]-O-[#1]
513	S:C:C-[#1]
514	O-N-C-C
515	N-N-C-C
516	[#1]-C=C-[#1]
517	N-N-C-N
518	O=C-N-N
519	N=C-N-C
520	C=C-C:C
521	C:N-C-[#1]
522	C-N-N-[#1]
523	N:C:C-C
524	C-C=C-C
525	[As]-C:C-[#1]
526	Cl-C:C-Cl
527	C:C:N-[#1]
528	[#1]-N-C-[#1]
529	Cl-C-C-Cl
530	N:C-C:C
531	S-C:C-C
532	S-C:C-[#1]
533	S-C:C-N
534	S-C:C-O
535	O=C-C-C
536	O=C-C-N
537	O=C-C-O
538	N=C-C-C
539	N=C-C-[#1]
540	C-N-C-[#1]
541	O-C:C-C
542	O-C:C-[#1]

Publicly Available Data To Build Statistical Models (Predictors)

- Reactome
  - Open-source on-line database of SME-curated biological pathways; organized by species
  - Prime unit is the reaction and entities (nucleic acids, proteins, molecules, complexes, etc.) are hand-mapped to the reaction
  - Reactions are then grouped into biological networks and pathways
  - We make use of UniProt ID (unique protein identifier) mappings to pathway terms for our models



Constructing ADR Health Outcome Groupings

- Used the following data sources
  - SIDER (Currently 996 drugs, 4192 side effects, 99423 interaction pairs)
    - Data on marketed drugs and their reported ADRs
    - Taken from package inserts and publicly available documents
  - MedDRA terms
    - Clinically validated and internationally-used medical terminology
  - Started with ADR Groupings used by Huang et al. (2013) in their ADR prediction work; they used MedDRA system organ classes and ICD-10 diagnosis codes to construct groupings of ADRs with similar mechanisms.
    - We started with the following ten groups defined in the paper:
      - Neoplasms benign, malignant, and unspecified
      - Blood and lymphatic systems disorders
      - Immune system disorders
      - Endocrine disorders
      - Psychiatric disorders
      - Cardiac disorders
      - Vascular disorders
      - Gastrointestinal disorders
      - Hepatobiliary disorders
      - Renal and urinary disorders

ADR Prediction with 1 vs. all L1-Regularized Logistic Regression

- Constructed three distinct (*m x p*) design matrices over a set of *m* = 732 drugs that combine chemical and biological information (both at the pathway and molecular levels)
  - AP – All protein associations for a given drug in DrugBank (p=838)
  - FP – Pubchem Substructure Fingerprints (p=616)
  - Pathways – Used Reactome UniProt IDs protein-to-pathway mappings to transform AP to a pathway representation (p=841), i.e.
    - Fjfjf
- Constructed (*m x k*) response matrix over the *k*=10 ADR groups
  - ADRs were down-selected to include only serious (e.g. high case fatality or hospitalization rates)
- In addition, considered 4 additional combinations for a total of seven design matrices:
  - 1 – AP (p=838)
  - 2 – FP (p=616)
  - 3 – Pathways (p=841)
  - 4 – AP+FP (p=1454)
  - 5 - AP+Pathways (p=1679)
  - 6 – FP+Pathwways (p=1471)
  - 7 – AP+FP+Pathways (p=2295)
- We want proteins, substructures, and pathways to compete on equal footing in the variable selection process

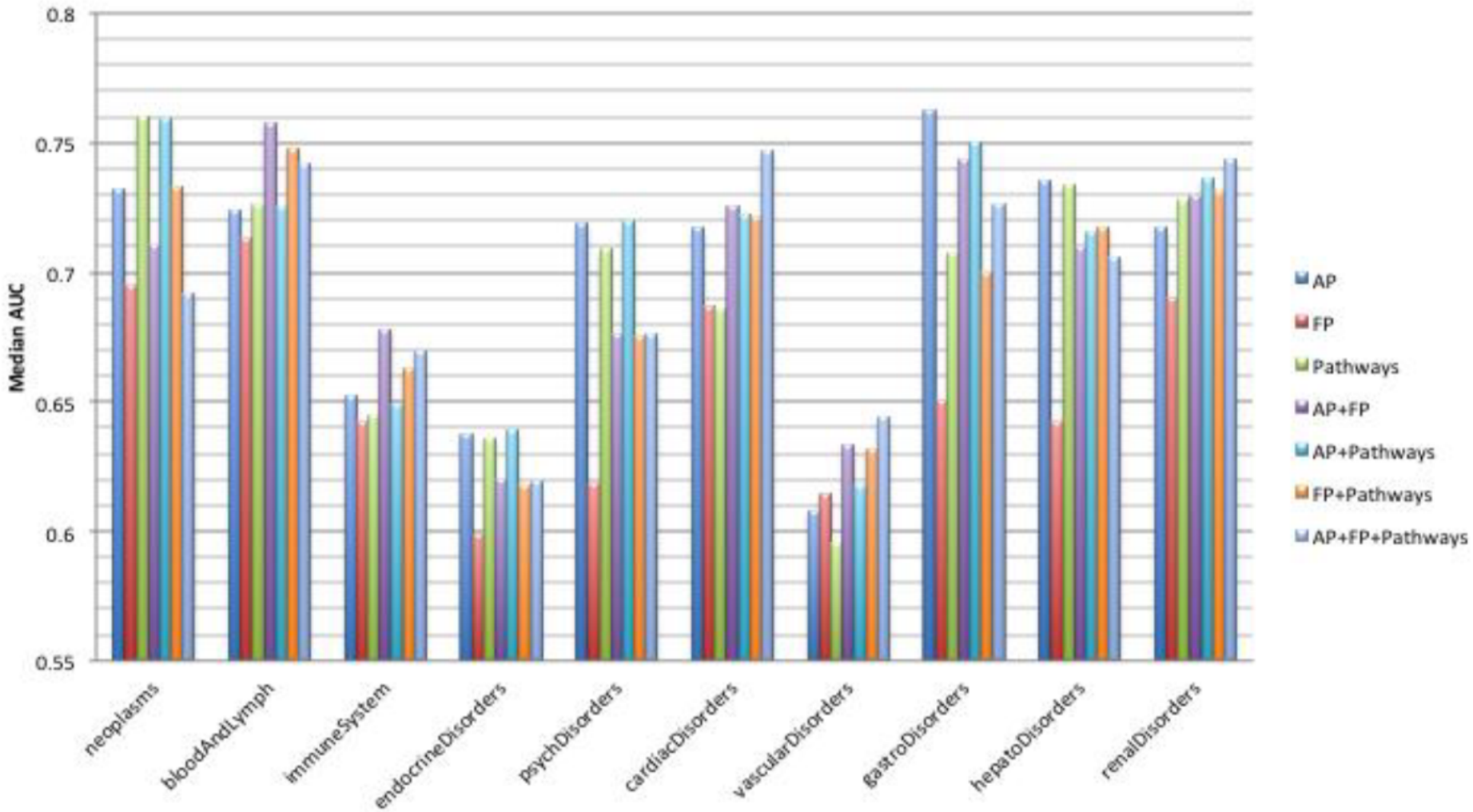
ADR Prediction with 1 vs. all L1-Regularized Logistic Regression

- We are working in a regime where *m<p* or *m<<p*
- Logistic regression using L1 regularization has been shown to work well for this regime and has become a standard workhorse for this problem, especially in biological/genomic applications where number of possible predictors are huge

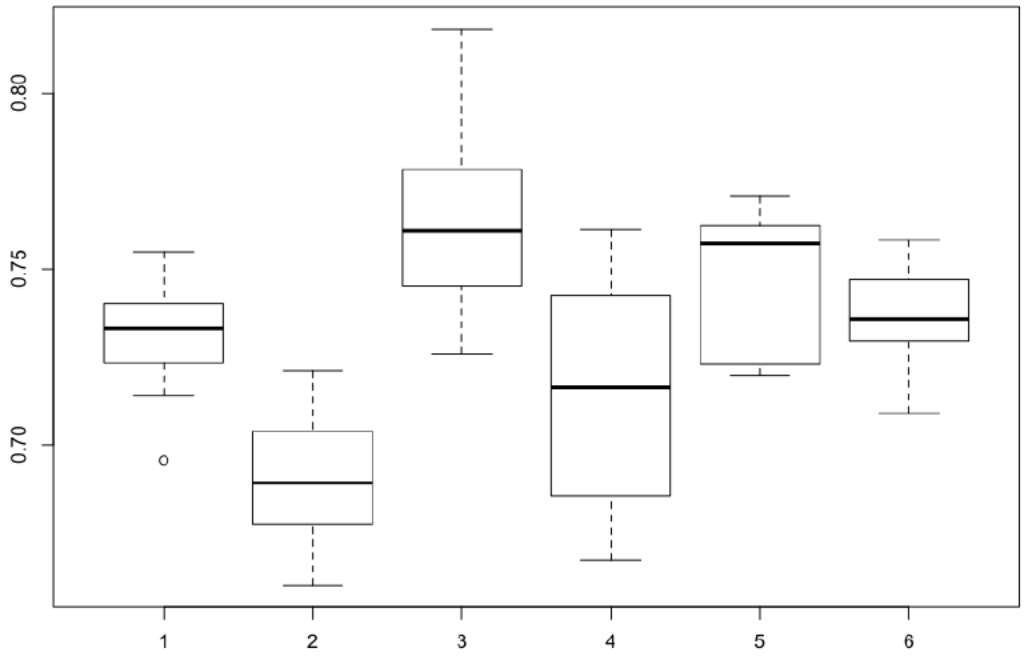
$$p(y_i = 1 | x_i; \beta) = \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)} \qquad L(\beta) = \sum_{i=1}^{m} y_i \ln p_i + (1 - y_i) \ln(1 - p_i) + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \arg\max_{\beta} L(\beta)$$

- During parameter estimation, L1 regularization drives the beta coefficient of variables that contribute little predictive value to a model to zero, more effectively than L2 regularization
  - Behaves similarly to step-wise selection and L2, but without the discontinuities of the step-wise method
- We perform one vs. all Logistic regression on all ten classes in turn using the R package 'glmnet' of Hastie and co-workers
- Given the sparsity of data, we did ten-fold cross-validation, finding optimal lambda (L1) parameters. The cost function is maximization of the area under the receiver-operator characteristic curve (AUC)
  - We repeat this ten times and report the median value for each ADR group

ADR Prediction with L1-regularized logistic regression (Results)

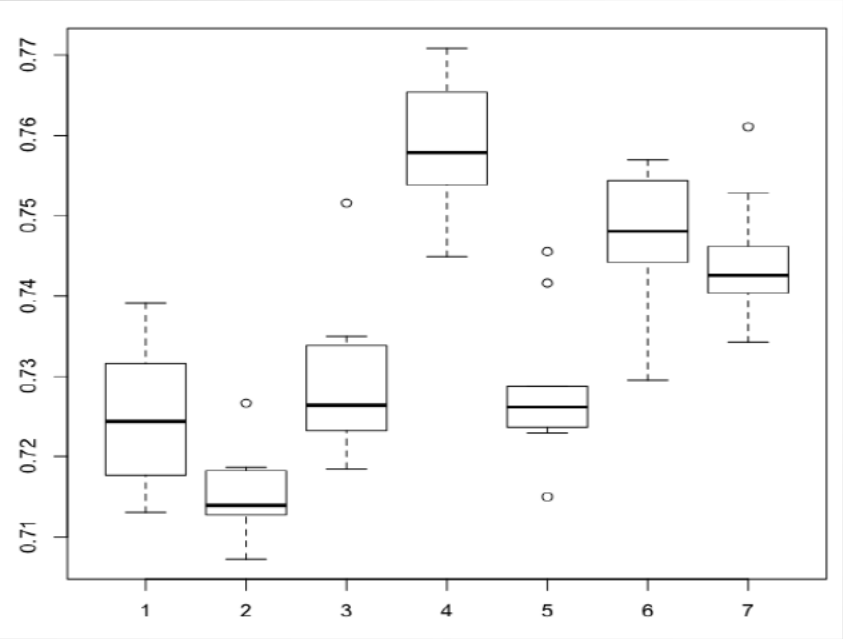


Neoplasms



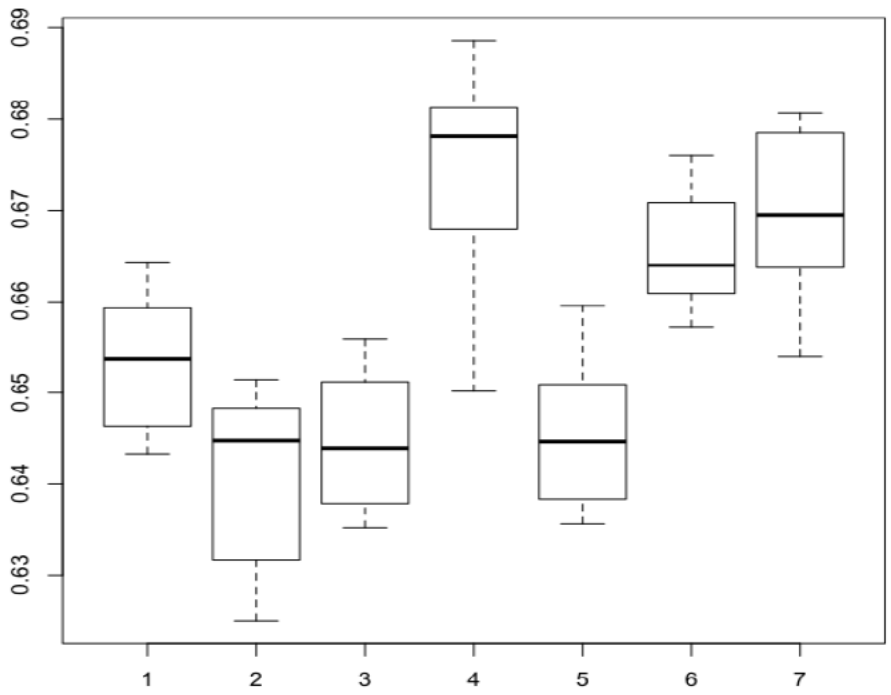

```
"CRMPs_in_Sema3A_signaling"
"Transport_of_nucleotide_sugars"
"Metal_ion_SLC_transporters"
"Generation_of_second_messenger_molecules"
"TRAF3-dependent_IRF_activation_pathway"
"Keratan_sulfate_biosynthesis"
"Synthesis_of_PG"
"PI_Metabolism"
"Platelet_degranulation"
"Regulation_of_Water_Balance_by_Renal_Aquaporins"
"Synthesis_of_UDP-N-acetyl-glucosamine"
"Signaling_by_BMP"
"N-Glycan_antennae_elongation"
"Calnexin/calreticulin_cycle"
"G_alpha_(s)_signalling_events"
"Synthesis_of_substrates_in_N-glycan_biosythesis"
"Muscarinic_acetylcholine_receptors"
"Androgen_biosynthesis"
"CHL1_interactions"
"Nuclear_signaling_by_ERBB4"
```

bloodAndLymph



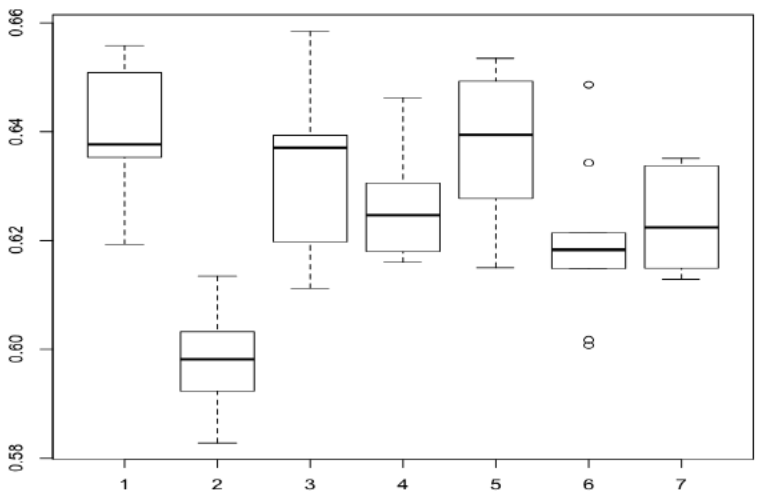

```
"3_beta-hydroxysteroid_dehydrogenase/Delta_5-->4-isomerase_type_II"
"Opioid_receptor,_sigma_1"
"Dihydropyrimidine_dehydrogenase_[NADP+]"
"Tubulin_alpha-3_chain"
"Ceramide_glucosyltransferase"
"Arylsulfatase_A"
"Thyroid_peroxidase"
"Thymidylate_synthase"
"Cl-C:C-O-C"
"Oxygen-insensitive_NADPH_nitroreductase"
"Glutathione_S-transferase_A2"
"Interleukin-1_beta"
"Glutamate_receptor_ionotropic,_kainate_1"
"FK506-binding_protein_1A"
"50S_ribosomal_protein_L10"
"Toll-like_receptor_7"
"Low-density_lipoprotein_receptor-related_protein_2"
"DNA-directed_RNA_polymerase_subunit_beta'"
"Sc1c(Cl)cccc1"
"Nischarin"
```

immuneSystem



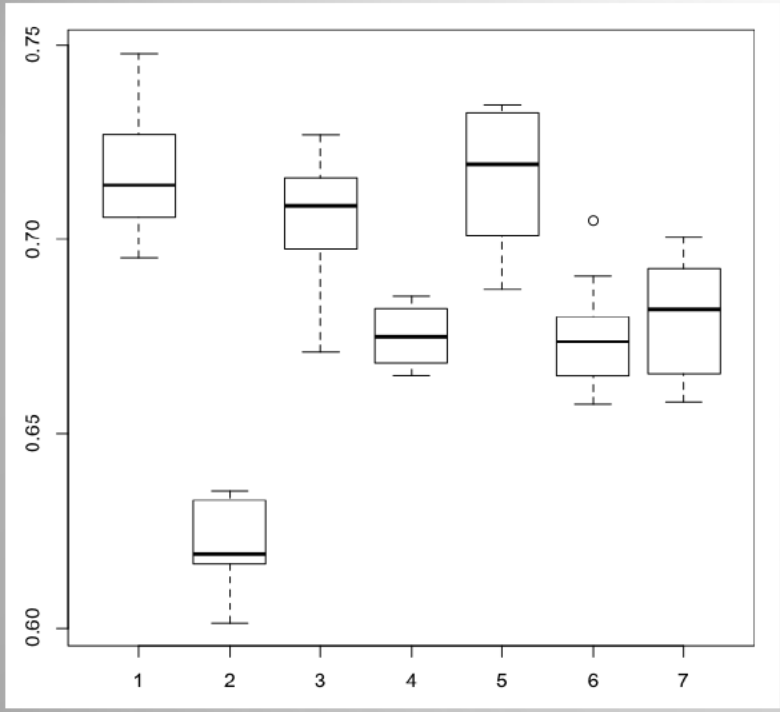

```
"O=N-C:C-N"
"C(~F)(:C)"
"Sc1c(N)cccc1"
"Type-1_angiotensin_II_receptor"
"Nc1cc(Cl)ccc1"
"Penicillin-binding_protein_1A"
"C(~C)(~H)(~P)"
"Solute_carrier_organic_anion_transporter_family_member_1B1"
"Angiotensin-converting_enzyme"
">=1Al_"
"O=C-C-C=O"
"Prostaglandin_G/H_synthase_1"
"FK506-binding_protein_1A"
"Voltage-dependent_P/Q-type_calcium_channel_subunit_alpha-1A"
"Cytochrome_P450_2C9"
"Cytochrome_P450_1A2"
"Sc1c(Cl)cccc1"
"Mineralocorticoid_receptor"
"O=C-C-C-C"
"C(-N)(=C)"
```

endocrineDisorders



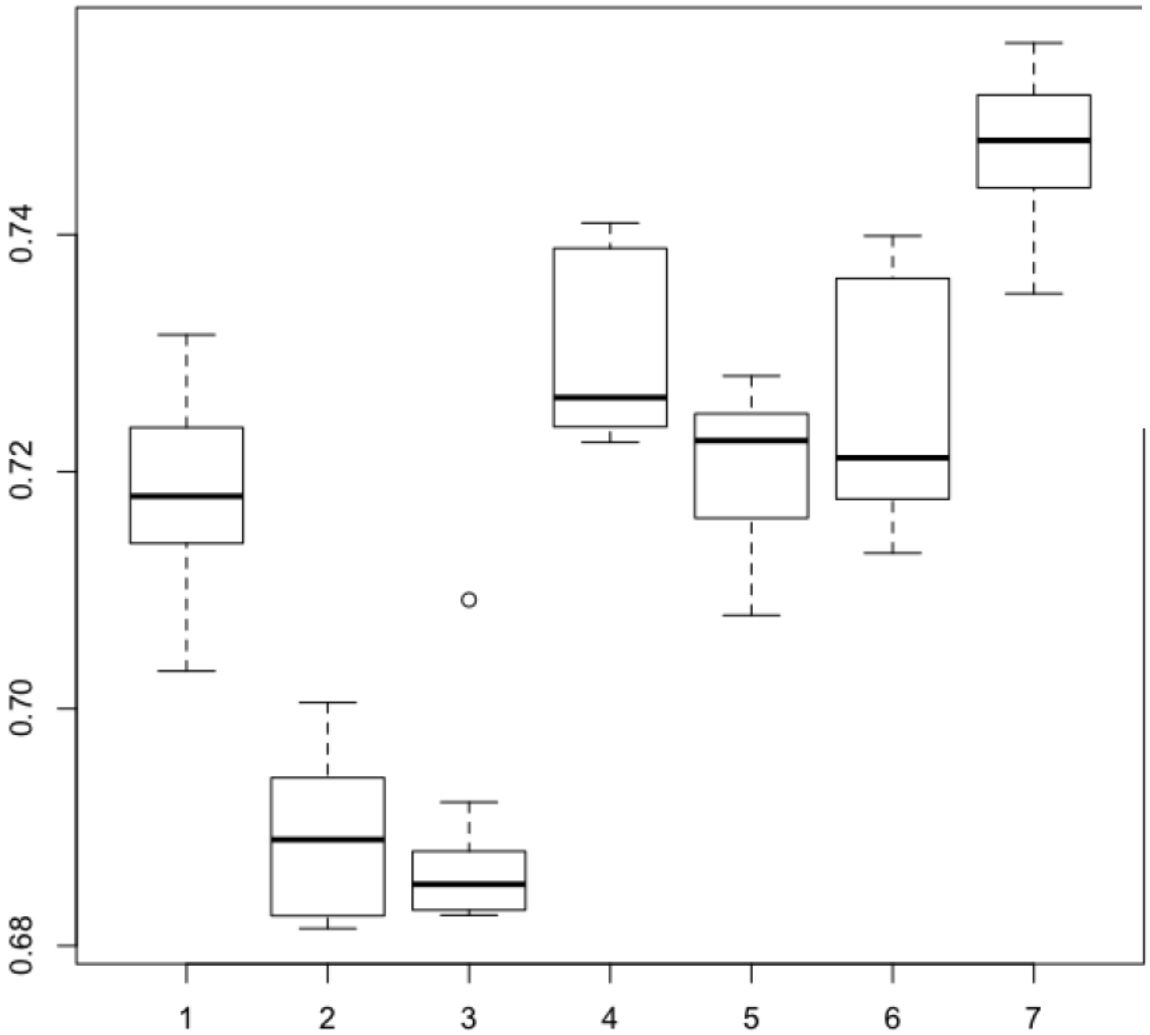

```
"Protein_tyrosine_kinase_2_beta"
"DNA_polymerase"
"Voltage-dependent_calcium_channel_subunit_alpha-2/delta-3"
"ATP-binding_cassette_sub-family_B_member_8,_mitochondrial"
"Integrin_beta-2"
"Follicle-stimulating_hormone_receptor"
"Tyrosine-protein_phosphatase_non-receptor_type_1"
"Somatostatin_receptor_type_1"
"Integrin_alpha-IIb"
"Probable_pyruvate-flavodoxin_oxidoreductase"
"Metabotropic_glutamate_receptor_5"
"Tumor_necrosis_factor_ligand_superfamily_member_11"
"ADP/ATP_translocase_1"
"Antithrombin-III"
"Glycogen_synthase_kinase-3_beta"
"High_affinity_interleukin-8_receptor_A"
"Retinal_rod_rhodopsin-sensitive_cGMP_3',5'-cyclic_phosphodiesterase_subunit_gamma"
"Nuclear_factor_NF-kappa-B_p105_subunit"
"DNA_gyrase_subunit_A"
"Inosine-5'-monophosphate_dehydrogenase_1"
```

psychDisorders



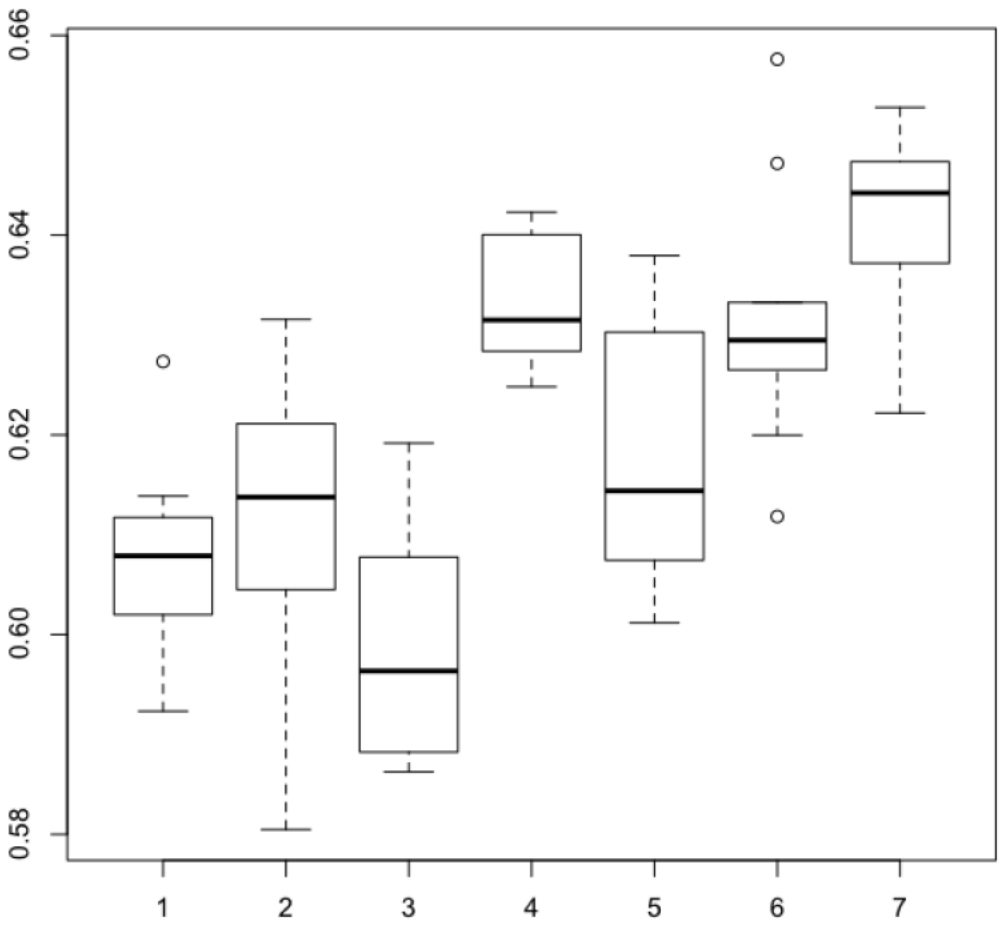

```
"Fatty-acid_amide_hydrolase"
"3_beta-hydroxysteroid_dehydrogenase/Delta_5-->4-isomerase_type_II"
"Sialidase-2"
"UDP-glucuronosyltransferase_2B17"
"Beta-lactamase_PSE-2"
"Cannabinoid_receptor_2"
"DNA_gyrase_subunit_A"
"DNA_polymerase"
"Glycogen_synthase_kinase-3_beta"
"Somatostatin_receptor_type_1"
"Aldose_reductase"
"ATP-binding_cassette_sub-family_B_member_8,_mitochondrial"
"Voltage-dependent_P/Q-type_calcium_channel_subunit_alpha-1A"
"Gamma-aminobutyric_acid_type_B_receptor_subunit_2"
"26S_proteasome_non-ATPase_regulatory_subunit_2"
"UDP-glucuronosyltransferase_1-4"
"Glutathione_S-transferase_A2"
"Spectrin_beta_chain,_brain_1"
"DNA_polymerase"
"Alanine_aminotransferase_1"
```

cardiacDisorders



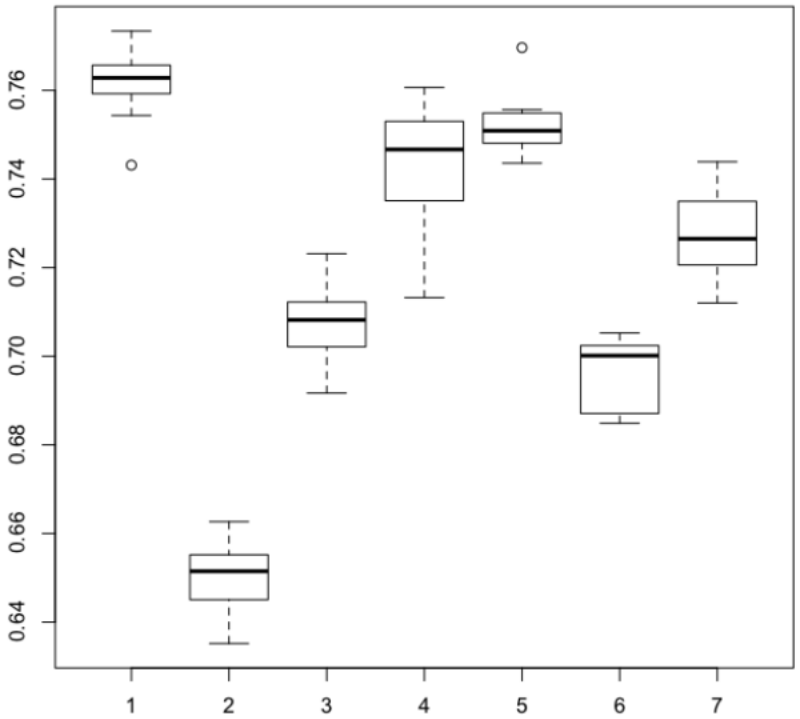

```
"DNA_polymerase"
"Cam-PDE_1_activation"
"Atrial_natriuretic_peptide_receptor_A"
"cGMP-inhibited_3',5'-cyclic_phosphodiesterase_A"
"Beta-lactamase_PSE-2"
"Penicillin-binding_protein_2"
"C(~C)(~C)"
"Catechol_O-methyltransferase"
"Thymidine_kinase"
"High_affinity_copper_uptake_protein_1"
"Sc1cc(Cl)ccc1"
"Gene_Expression"
"Peroxisome_proliferator-activated_receptor_alpha"
"Late_Phase_of_HIV_Life_Cycle"
"Translocator_protein"
"Glutathione_S-transferase_A1"
"Sodium_channel_protein_type_4_subunit_alpha"
"O-C-C=N"
"The_activation_of_arylsulfatases\t"
"Calnexin/calreticulin_cycle"
```

vascularDisorders



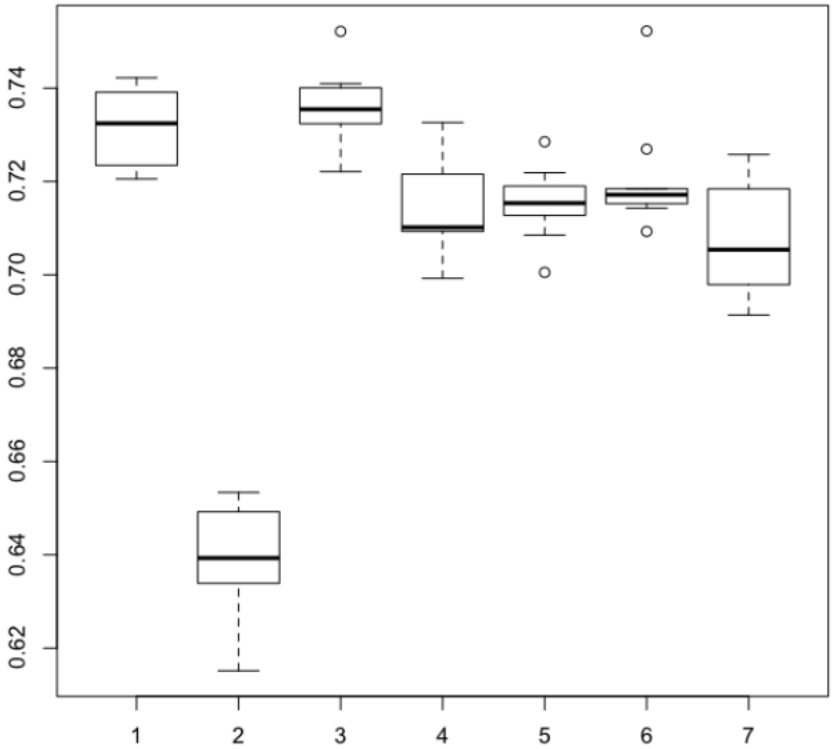

```
"Dihydropyrimidine_dehydrogenase_[NADP+]"
"Estrogen_receptor_beta"
"Prothrombin"
"Mannitol_dehydrogenase"
"3_beta-hydroxysteroid_dehydrogenase/Delta_5-->4-isomerase_type_II"
"3-phosphoinositide-dependent_protein_kinase_1"
"1,3-beta-glucan_synthase_component_FKS1"
"Gamma-aminobutyric_acid_type_B_receptor_subunit_2"
"Nuclear_factor_NF-kappa-B_p105_subunit"
"5-hydroxytryptamine_receptor_1E"
"Phospholipase_A2,_membrane_associated"
"Glutamate_receptor_ionotropic,_kainate_1"
"Oxygen-insensitive_NADPH_nitroreductase"
"Tumor_necrosis_factor_ligand_superfamily_member_11"
"Glycosyltransferase_GtfA"
"Elastin"
"Opioid_receptor,_sigma_1"
"Copper-transporting_ATPase_2"
"Beta-lactamase"
"26S_proteasome_non-ATPase_regulatory_subunit_2"
```

gastroDisorders



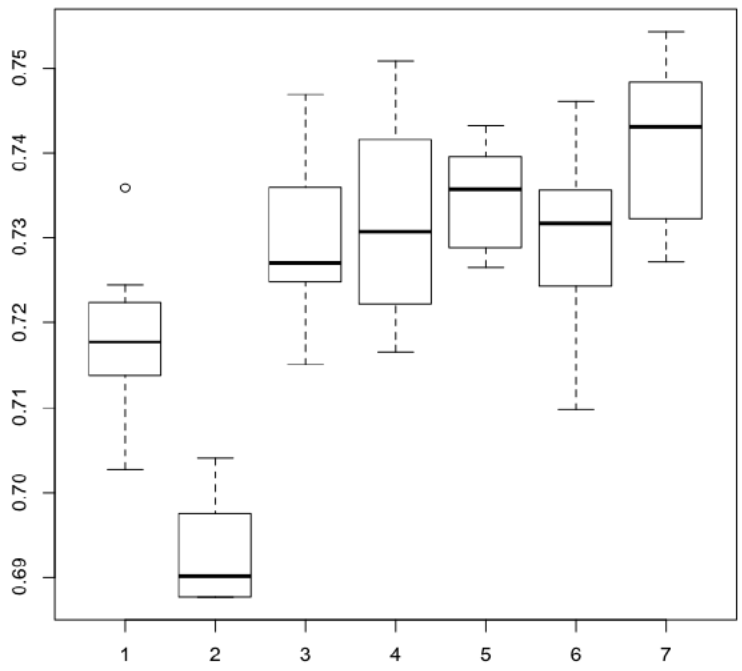
"DNA_topoisomerase_4_subunit_A"
"Angiotensin-converting_enzyme"
"Farnesyl_pyrophosphate_synthetase"
"Prostaglandin_G/H_synthase_2"
"Penicillin-binding_protein_2"
"DNA_polymerase"
"Solute_carrier_family_12_member_3"
"Reverse_transcriptase"
"Solute_carrier_family_12_member_1"
"Dihydropyrimidine_dehydrogenase_[NADP+]"
"Prostaglandin_E2_receptor_EP2_subtype"
"Peroxiredoxin-5,_mitochondrial"
"Peptidoglycan_synthetase_ftsI"
"Beta-lactamase_PSE-2"
"Nischarin"
"Neuraminidase"
"Branched-chain-amino-acid_aminotransferase,_cytosolic"
"Thymidylate_synthase"
"Gamma-aminobutyric_acid_type_B_receptor_subunit_2"
"Beta-lactamase"

hepatoDisorders




```
"RNA_Polymerase_I,_RNA_Polymerase_III,_and_Mitochondrial_Transcription"
"Cysteine_formation_from_homocysteine"
"Expression_of_DEC1_(BHLHE40,_BHLHB2)"
"Expression_of_ACOX1"
"Calnexin/calreticulin_cycle"
"Latent_infection_of_Homo_sapiens_with_Mycobacterium_tuberculosis"
"Phosphorylation_of_ITAMs_of_Ig-alpha_(CD79A)_and_Ig-beta_(CD79B)"
"Expression_of_PLIN2"
"Nuclear_signaling_by_ERBB4"
"Glycoprotein_hormones"
"Smooth_Muscle_Contraction"
"Type_I_hemidesmosome_assembly"
"Regulated_proteolysis_of_p75NTR"
"Expression_of_APOA2"
"Hemostasis\t"
"PERK_regulated_gene_expression"
"The_activation_of_arylsulfatases\t"
"Activation_of_Gene_Expression_by_SREBP_(SREBF)"
"Eicosanoids"
"Expression_of_Acetyl_CoA_Carboxylase_1_(ACACA,_ACC1)"
```

renalDisorders

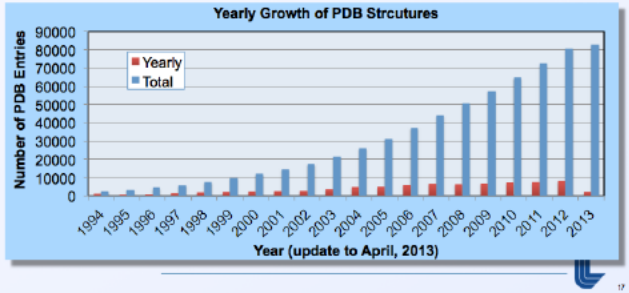



```
"3_beta-hydroxysteroid_dehydrogenase/Delta_5-->4-isomerase_type_II"
"Thymidylate_synthase"
"Amiloride-sensitive_sodium_channel_subunit_gamma"
"Acetylcholine_receptor_subunit_alpha"
"1,3-beta-glucan_synthase_component_FKS1"
"Somatostatin"
"Potassium_voltage-gated_channel_subfamily_KQT_member_1"
"Thymidine_kinase"
"Vascular_endothelial_growth_factor_A"
">=1I_"
"Tissue-type_plasminogen_activator"
"Activation_of_Gene_Expression_by_SREBP_(SREBF)"
"Potassium_channel_subfamily_K_member_1"
"Cytidine_deaminase"
"Alkaline_phosphatase,_placental-like"
"Voltage-dependent_calcium_channel_gamma-1_subunit"
"Gamma-aminobutyric_acid_type_B_receptor_subunit_2"
"Solute_carrier_family_22_member_16"
"26S_proteasome_non-ATPase_regulatory_subunit_2"
"Solute_carrier_family_2,_facilitated_glucose_transporter_member_2"
```

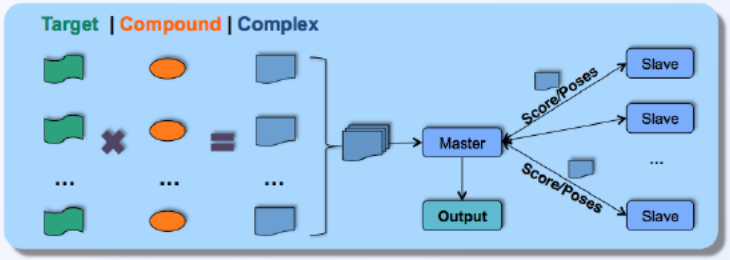
Results using drug-protein docking scores calculated using VinaLC, corrected with free energy method, as predictors of ADRs

Why do we need high performance computing in drug design?

- 1 ligand docking into 1 target takes ~1 min. Rescore 1 pose takes >10 min.
- Drug-like compounds ~10^{60} possibilities.
- Druggable targets ~10% human genome/ off-target prediction.

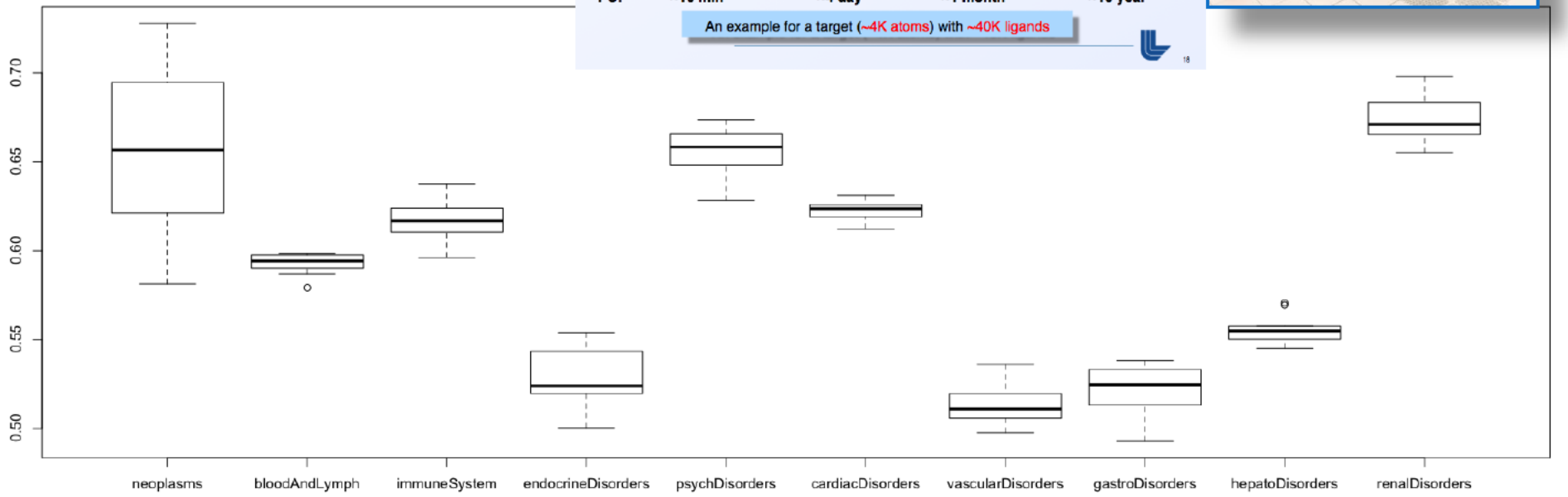


Our current workflow of in-house docking & rescoring protocol saves time



	Prep protein target	Prep ligand compound	docking	Free energy rescoring
HPC:	~10 min @ 1 CPU	~1 hr @ 100 CPU	~1 hr @ 700 CPU	~5 hr @ 15K CPU
PC:	~10 min	~4 day	~1 month	~10 year

An example for a target (~4K atoms) with ~40K ligands



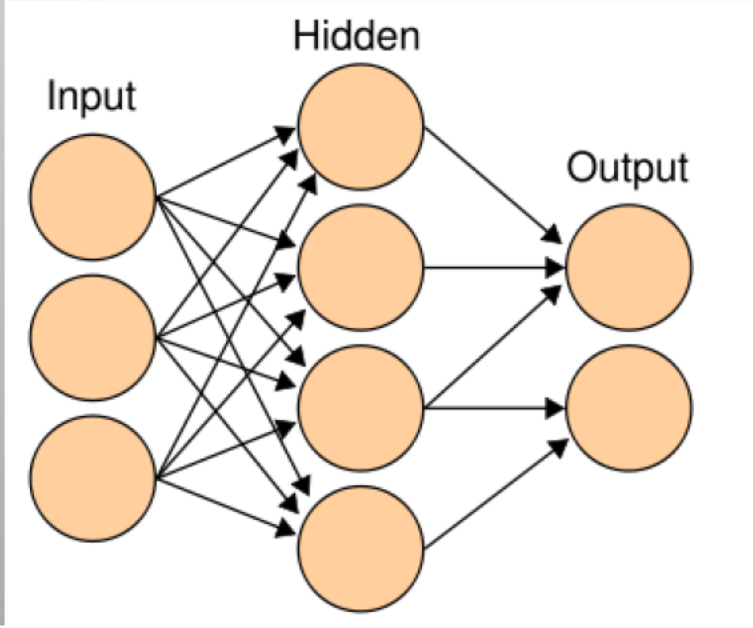
Shows promise; Initial protein panel needs to include more targets through protein homology modeling

Side effect profile prediction with neural networks

- Use single-hidden layer of 50 nodes that allow for non-linear mixing of predictors
- L2-regularization insufficient in the presence of large number of irrelevant variables, i.e. number of samples to train the model well scales linearly with irrelevant variables (Ng 2004)
- Used Principal Component Analysis to scale down to a "reasonable" number of predictors ~70 (chosen so we have an order of magnitude more samples than predictors)
- Estimated NN weights using a quasi-Newton optimizer with BFGS updating of the Hessian as implemented in R-package 'ucminf'
- Gradients calculated using the backpropagation method

Drug-protein binding (AP)

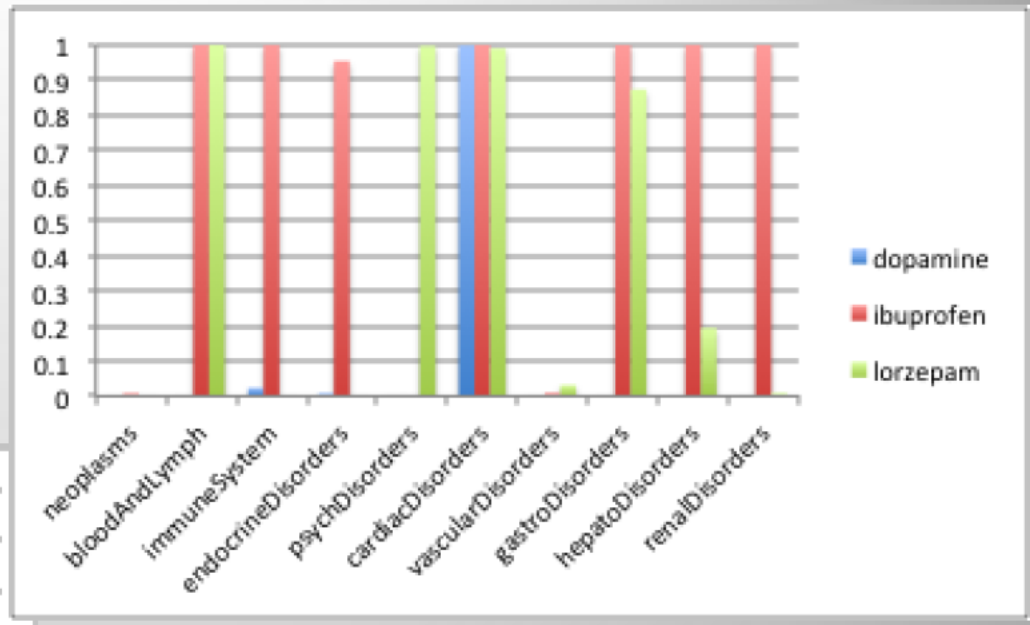
Samples projected onto first 70 PCs

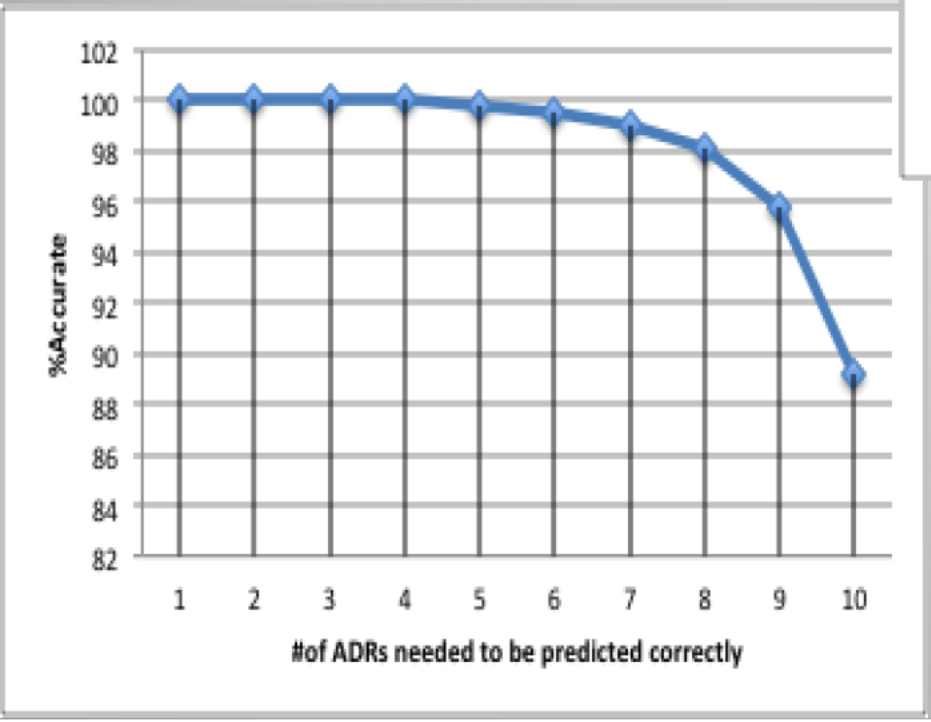


Ten ADR outcomes

Preliminary Side Effect Profile Results On Training Set for AP dataset

Side effect profiles predicted for 3 drugs in the training set. Vertical axis is probability that drug will have one or more ADRs in that group





Accuracy here is defined as the number of *n* side effects per drug predicted correctly in the training set

n is varied from 1 to the maximum of 10