bcl::ChemInfo Suite Enables Machine Learning-Based Drug Discovery Using GPUs

Edward W. Lowe, Jr.

Nils Woetzel

May 17, 2012

Outline

- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

Outline

- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

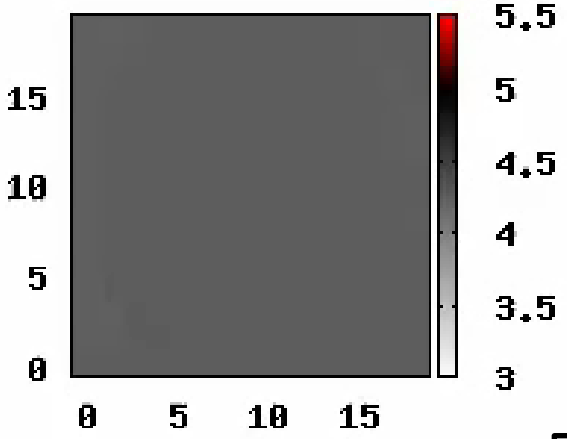
BCL

- BioChemistry Library
  - C++ library for small molecule and protein modeling
  - Machine learning techniques
  - OpenCL GPU-Acceleration

bcl::ChemInfo Goal

MYSQL

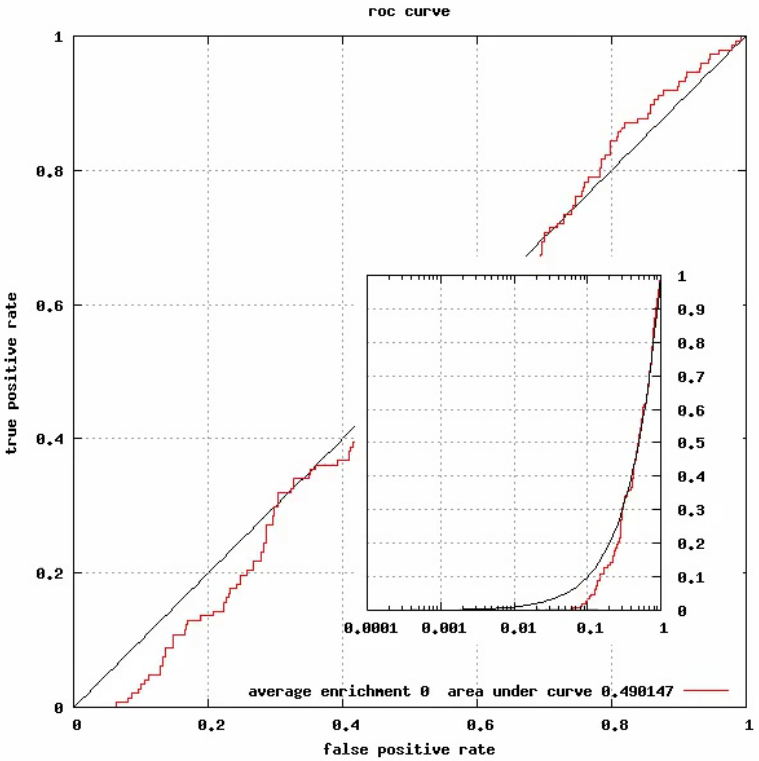
GPU

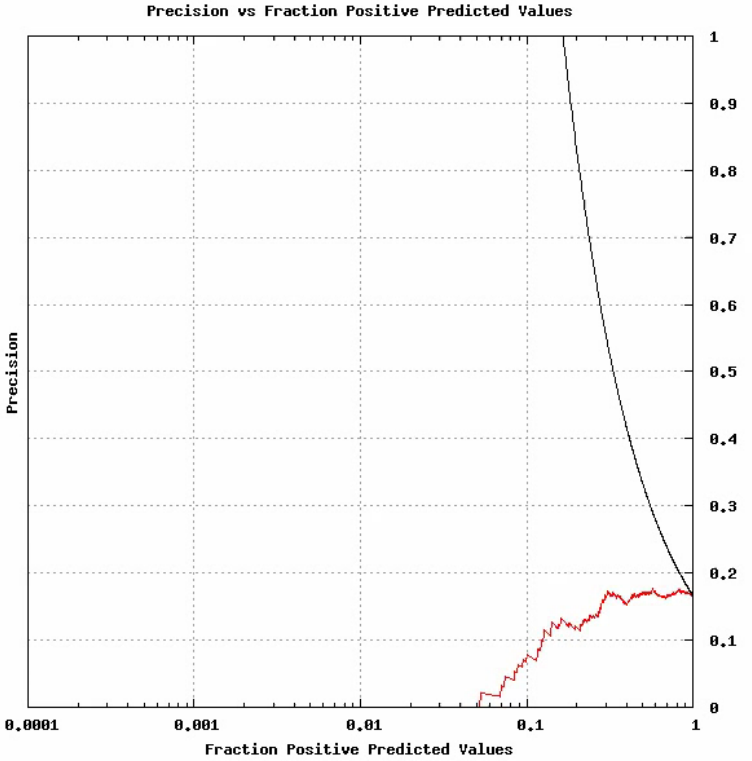


HPC

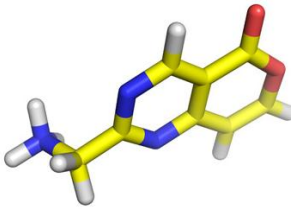
Automation

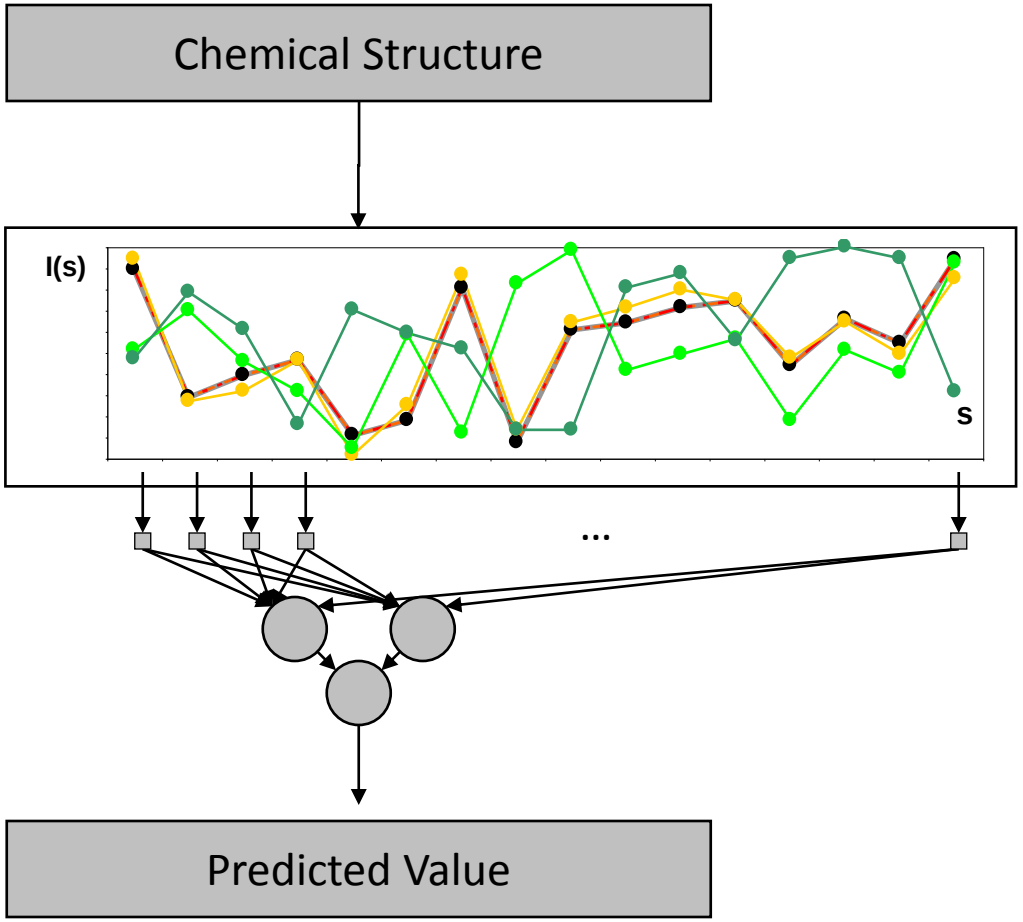
bcl::ChemInfo Goal





Machine Learning Calculates Properties from Numerical Description



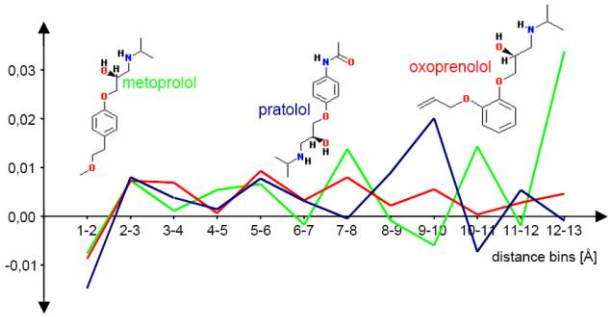


Encoding Chemical Data

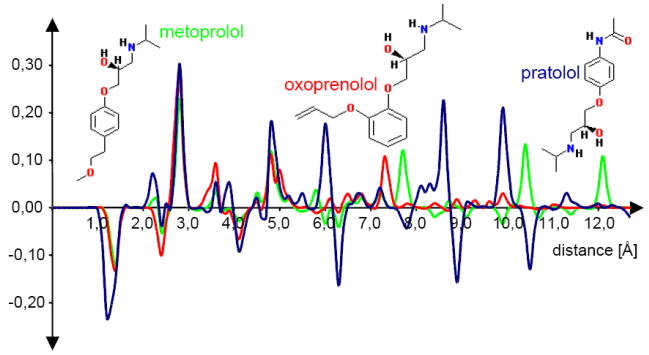
Scalar Descriptors

- Weight
- H-Bond donor
- H-Bond acceptor,
- Topological polar surface area (TPSA)

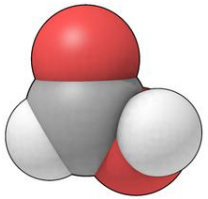
2D/3D Autocorrelation



Radial Distribution Function

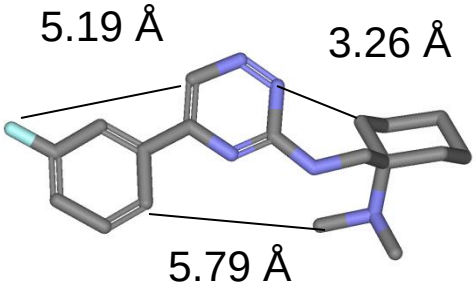


vdWaals Surface Area

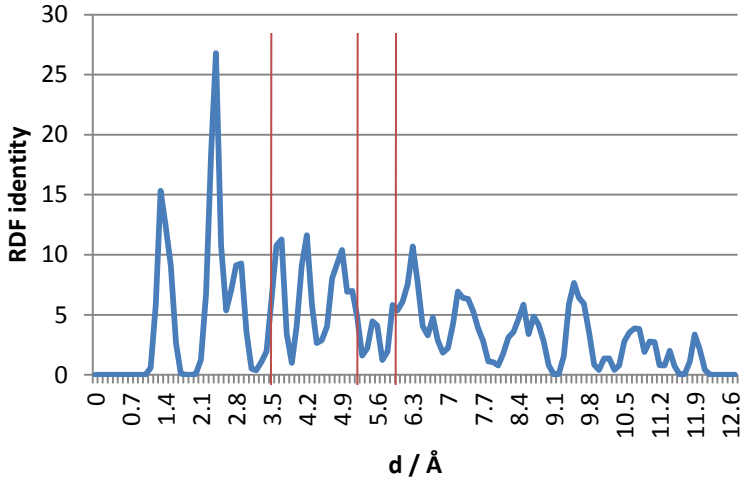


60 descriptor groups → 1284 numerical descriptor values

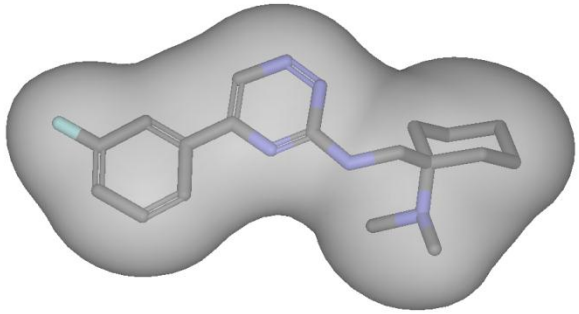
Radial Distribution Functions Describe 3D Shape …



$$g(d) = \sum_{\substack{atom\\pairs\\i,j}} e^{-B(d-d_{ij})^2}$$



‖

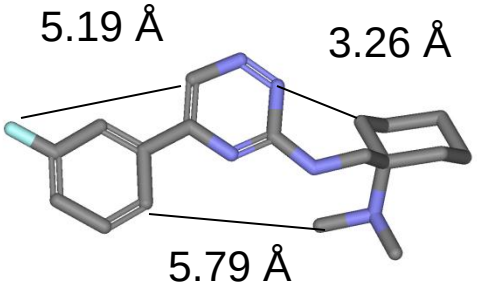


where:

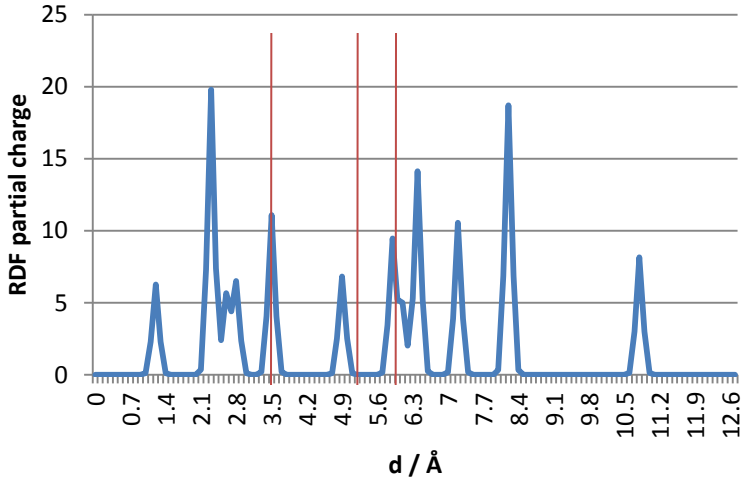
d_{ij} – distance between two atoms

B – temperature factor, here 100

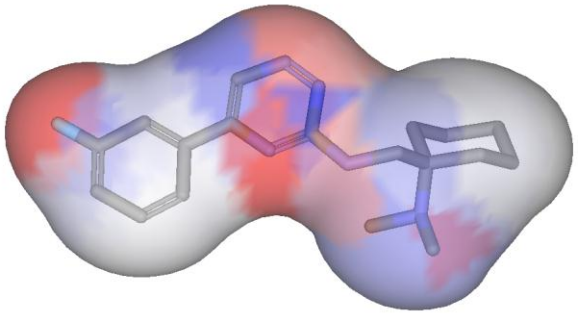
… but can also Encode Chemical Properties such as Partial Charge



$$g(d) = \sum_{\substack{atom\\pairs\\i,j}} A_i A_j \, e^{-B(d-d_{ij})^2}$$



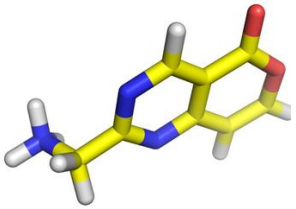
‖

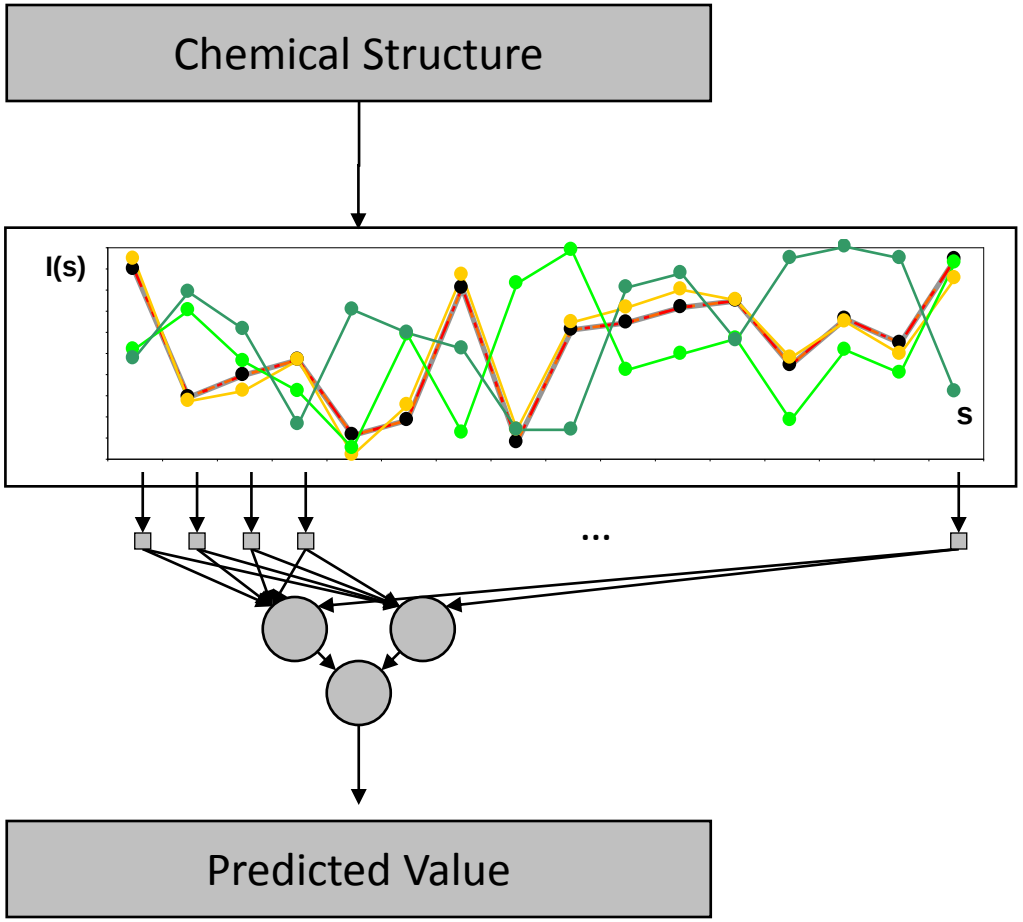


where:

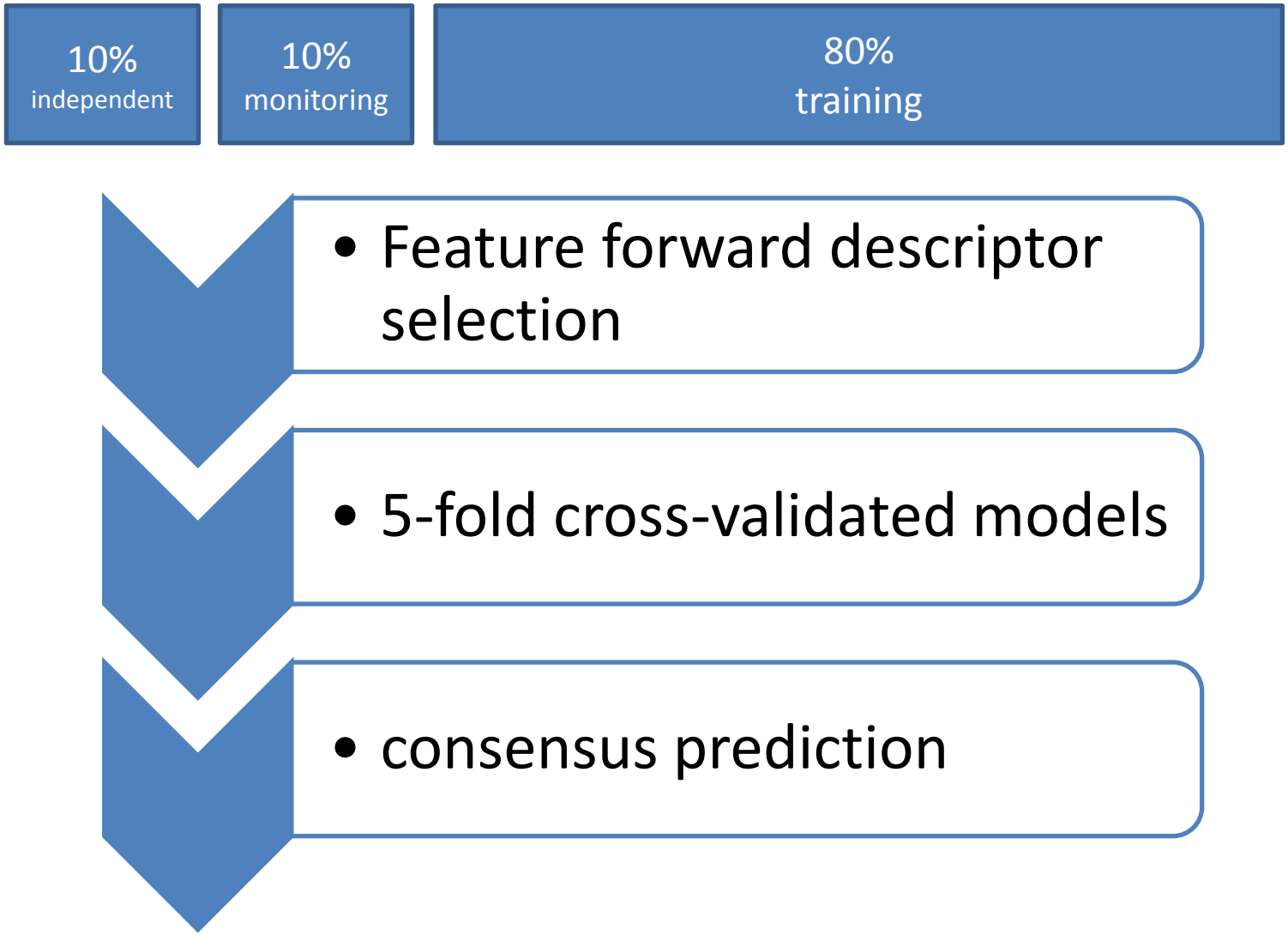
- d_{ij} – distance between two atoms
- A_i , A_j – atom properties, here lone pair electro negativity
- B – temperature factor, here 100

Machine Learning Calculates Properties from Numerical Description

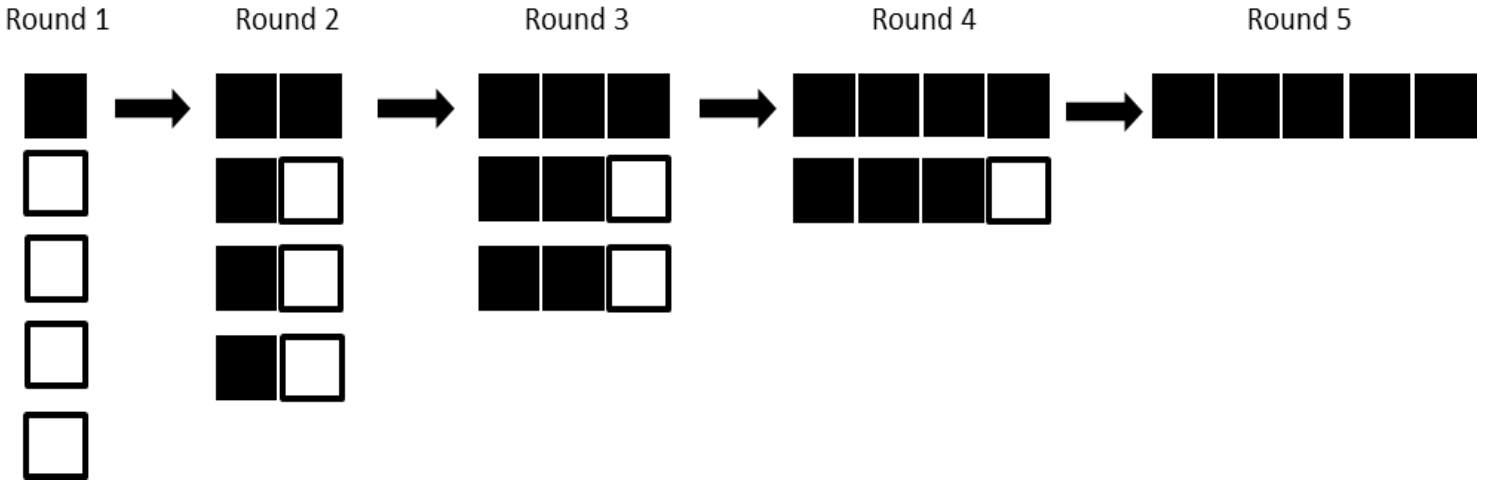




Protocol For Model Training



Forward Feature Selection



$$cv * \frac{n(n+1)}{2} = 9150$$

GPU Performance

Data Set ID	Actives	Inactives
884	3,438	7,066
893	5,398	65,259
1445	883	206,897

ML Method	884	893	1445
ANN	109/1 (109)	1151/10 (115)	3660/32 (114)
SVM	14/0.4 (35)	145/5 (29)	441/14 (32)
KNN	7/0.4 (18)	714/25 (29)	6118/90 (68)

GPU Performance

Data Set ID	Actives	Inactives
884	3,438	7,066
893	5,398	65,259
1445	883	206,897

Similarity Measure	884	893	1445
Tanimoto	53/0.2 (265)	147/0.55 (267)	3.4/0.02 (170)
Cosine	47/0.2 (235)	150/0.53 (**283**)	3.5/0.02 (175)
Dice	52/0.2 (260)	145/0.54 (269)	3.9/0.02 (195)
Euclidean	27/0.2 (138)	95/0.51 (186)	2.3/0.01 (230)
Manhattan	20/0.2 (100)	56/0.52 (108)	1.6/0.01 (160)

bcl::ChemInfo Suite

- Molecule -> feature vectors (descriptors)
- Feature Selection (FFS, BFS, ISA, PCA*; PBS)
- Diverse objective functions
- ANN*, SVR*, kNN*, Kohonen, DT
- MYSQL
- Model analysis
- Virtual Screening
- Similarity Analysis*

Note: * = GPU-accelerated

Lowe Jr, E.W., et al. *GPU-Accelerated Machine Learning Techniques Enable QSAR Modeling of Large HTS Data*. in *Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. 2012. San Diego, CA: IEEE.

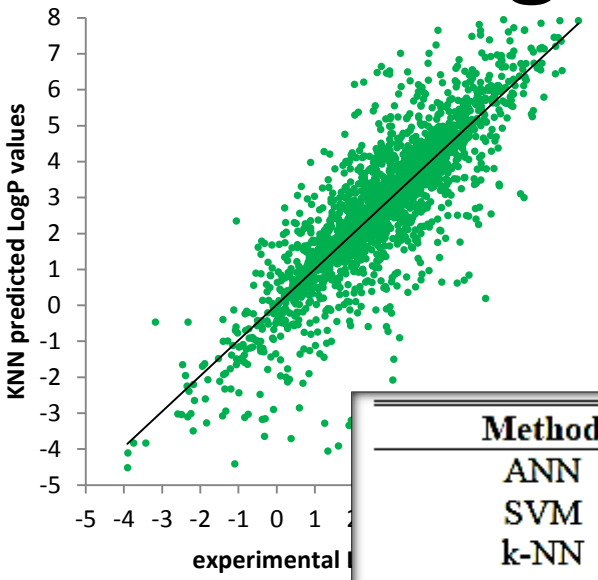
Outline

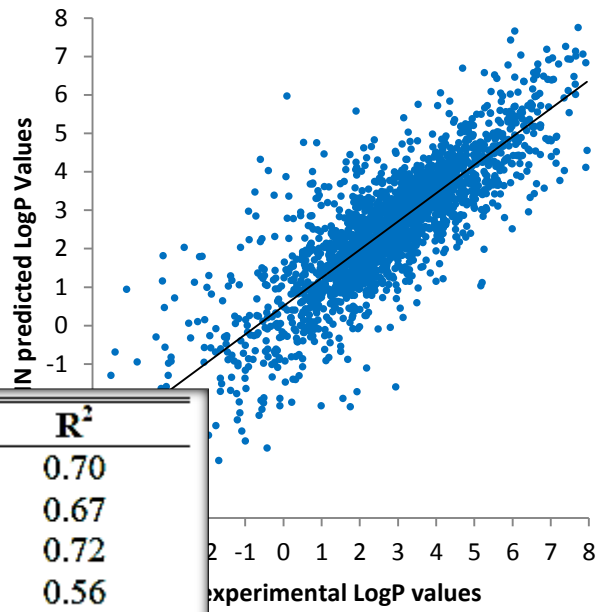
- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

logP Prediction

- Metric of hydrophobicity
- Important for molecule fate
- Logarithm of the octanol-water partition coefficient
- 22,500 compounds from MDDR, reaxys, SciFinder

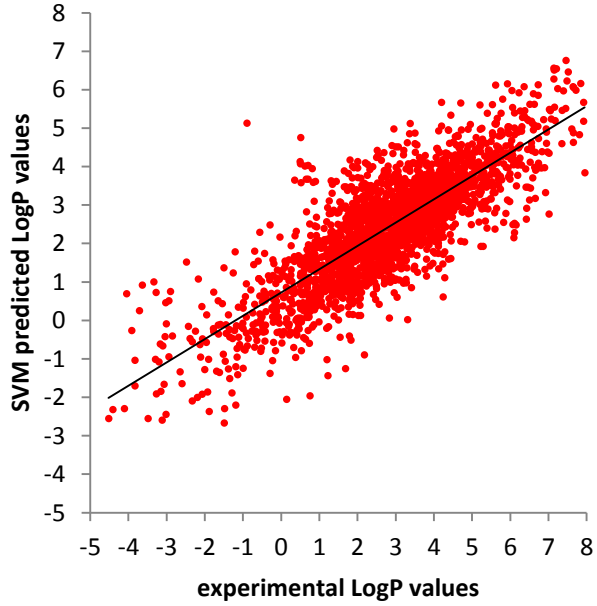
logP Prediction

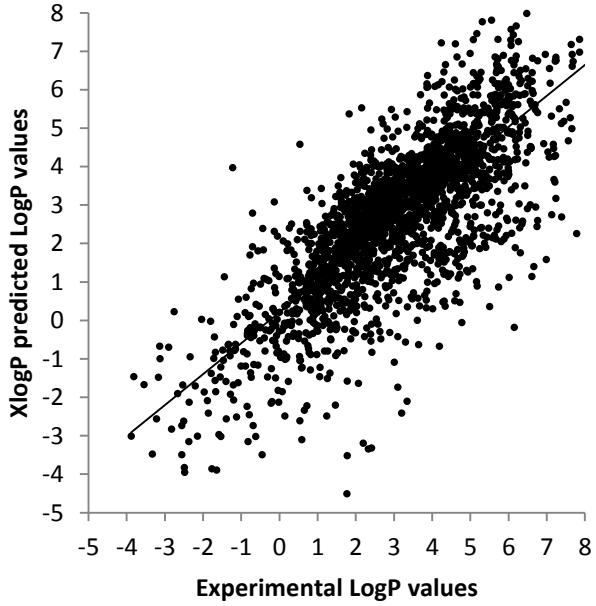




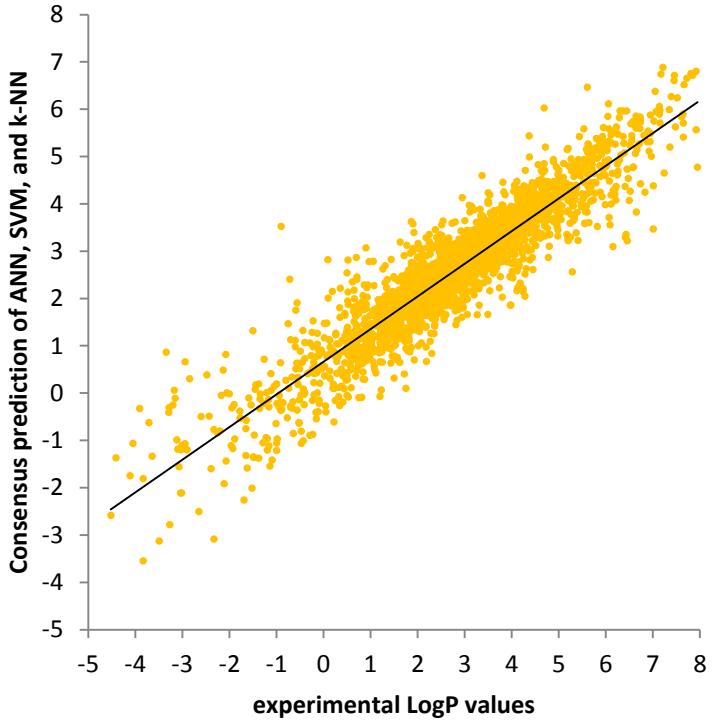
Method	RMSD	R^2
ANN	1.20	0.70
SVM	1.21	0.67
k-NN	1.03	0.72
XlogP	1.41	0.56

XLogP





ANN+SVM+KNN



Method	RMSD	R^2
ANN/SVM	1.04	0.8
ANN/k-NN	1.06	0.81
SVM/k-NN	0.99	0.77
ANN/SVM/k-NN	0.86	0.86

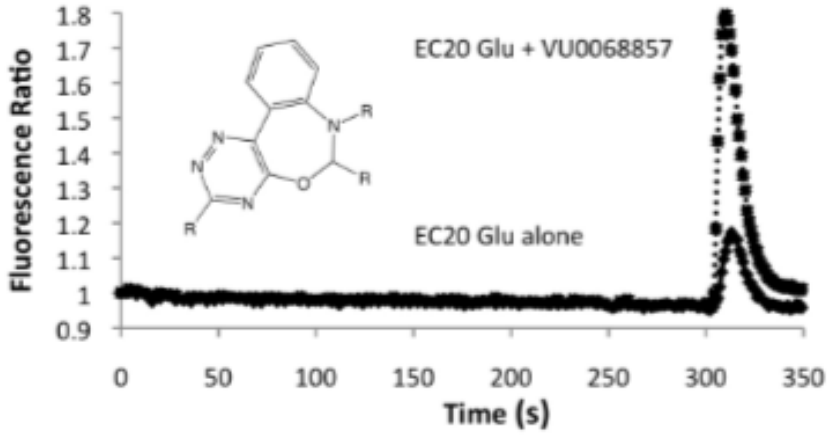
Lowe, E.W., Jr., et al., *Comparative Analysis of Machine Learning Techniques for the Prediction of LogP*, in *SSCI 2011 CIBCB - 2011 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*2011: Paris, France

Outline

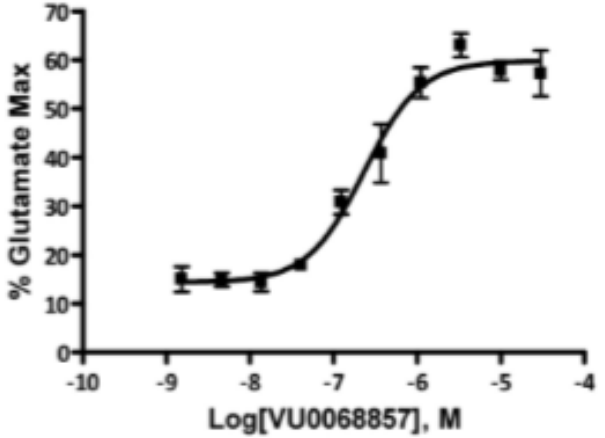
- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

High-Throughput Screen yields 1387 PAMs and 345 NAMs of $mGluR_5$

a)



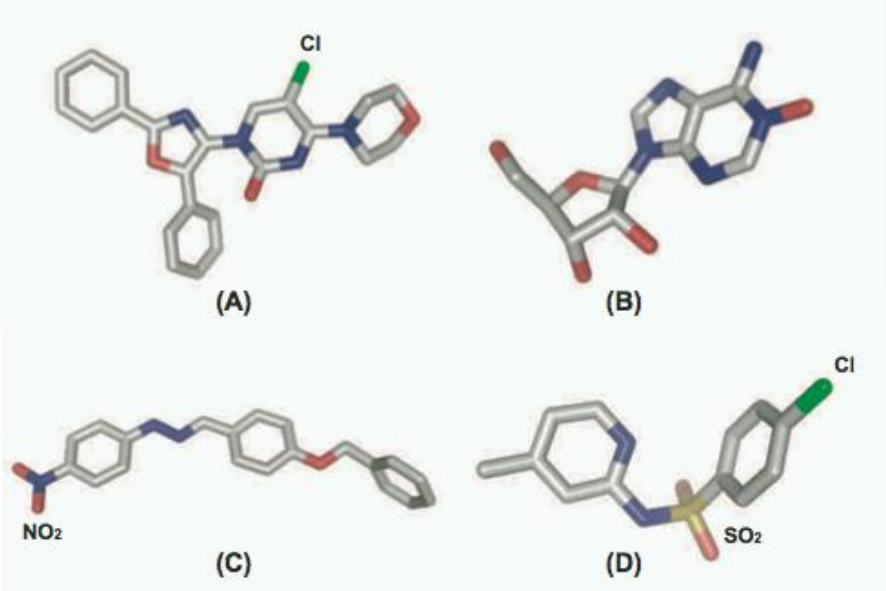
b)



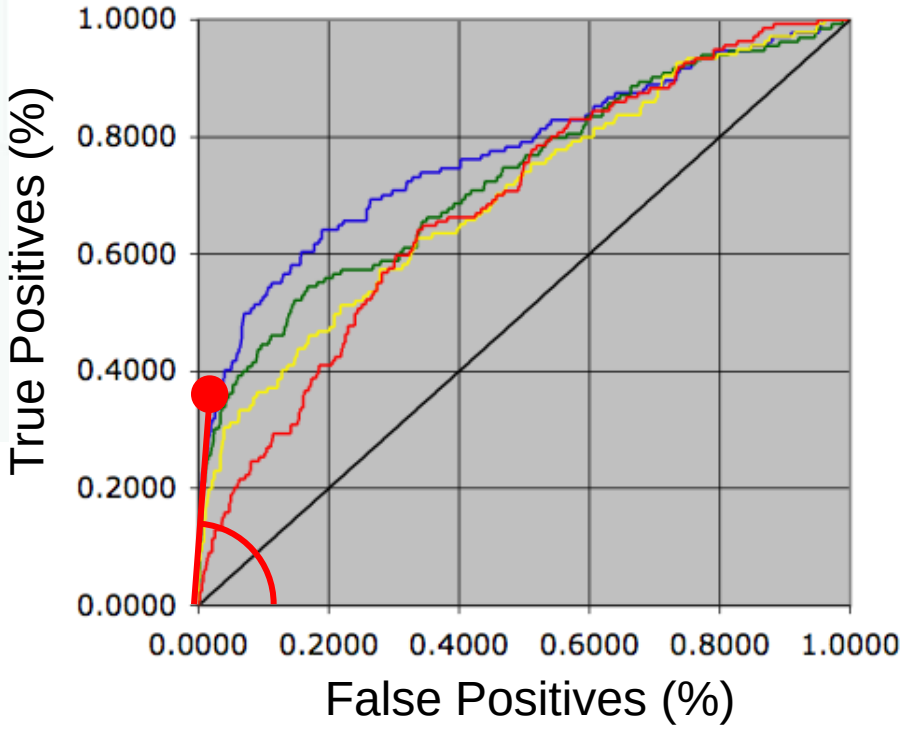
- 150,000 compounds were tested for allosteric modulation of $mGluR_5$
- 1,387 (0.94%) compounds were verified as PAMs of $mGluR_5$
- 345 (0.23%) compounds were verified as NAMs of $mGluR_5$.

Niswender, C. M.; Johnson, K. A.; Luo, Q.; Ayala, J. E.; Kim, C.; Conn, P. J.; Weaver, C. D. *Mol Pharmacol* ***2008, 73, 1213-24.***

Virtual Screen for Highly Active Compounds and Novel Leads



vHTS Training Optimization (ROC curves)

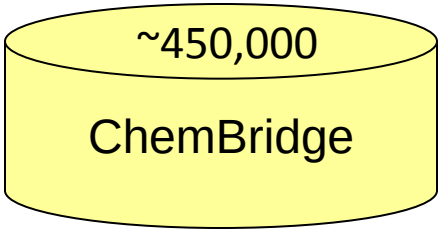


- A) True positive
- B) False negative
- C) False positive
- D) True negative

- Enrichment of Active Compounds by 43x

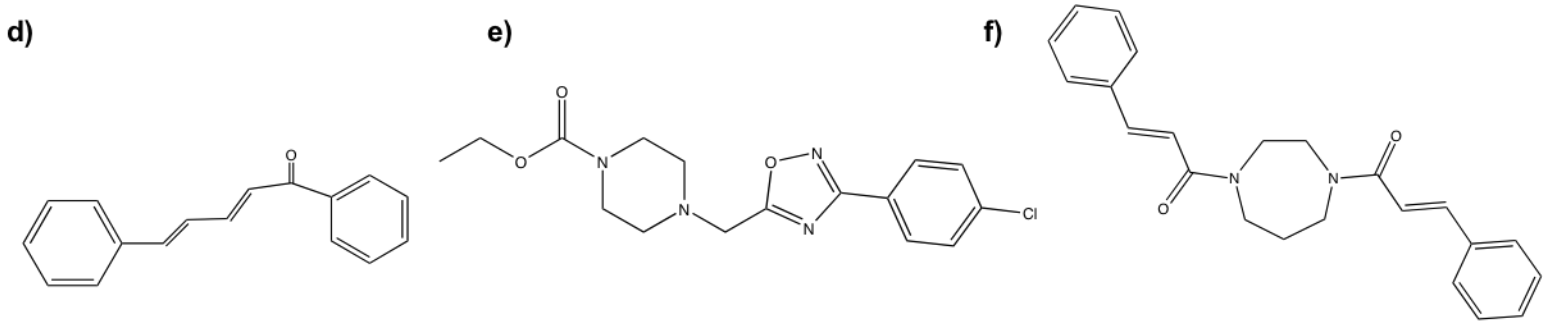
$$Enrichment = \frac{TP/(TP+FP)}{P/(P+N)}$$

Experimental Results $mGluR_5$ Positive Allosteric Modulators



824 Compounds predicted with EC_{50} < 1μM by QSAR model

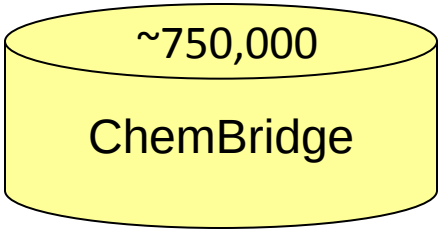
232 Compounds (28.1%) were confirmed as $mGlur_5$ PAMs
Enrichment = 28.1% / 0.96% = 30



Non-trivial scaffold modifications with mGluR5 PAM activity

Mueller, R., et al., *Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening.* ACS Chemical Neuroscience, 2010. **1**(4): p. 288-305.

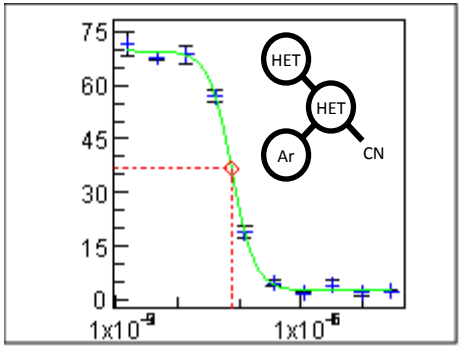
Experimental Results $mGluR_5$ Negative Allosteric Modulators



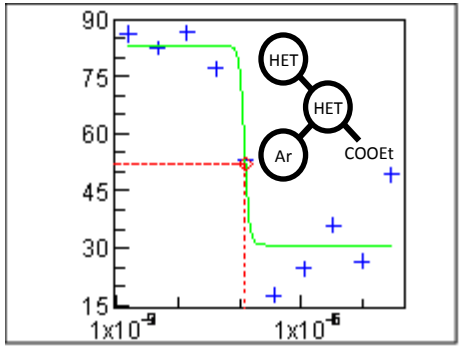
749 Compounds with novel Scaffolds predicted
with EC_{50} < 10μM by QSAR model

12 Compounds (3.6%) were confirmed as $mGlur_5$ NAMs
Enrichment = 3.6% / 0.23% = 16

VU0240790-4
EC_{50} = 75 nM



VU0360620-1
EC_{50} = 124 nM



Mueller, R., et al., *Discovery of 2-(2-Benzoxazoyl amino)-4-Aryl-5-Cyanopyrimidine as Negative Allosteric Modulators (NAMs) of Metabotropic Glutamate Receptor 5 (mGlu5): From an Artificial Neural Network Virtual Screen to an In Vivo Tool Compound.* ChemMedChem, 2012. **7**(3): p. 406-414.

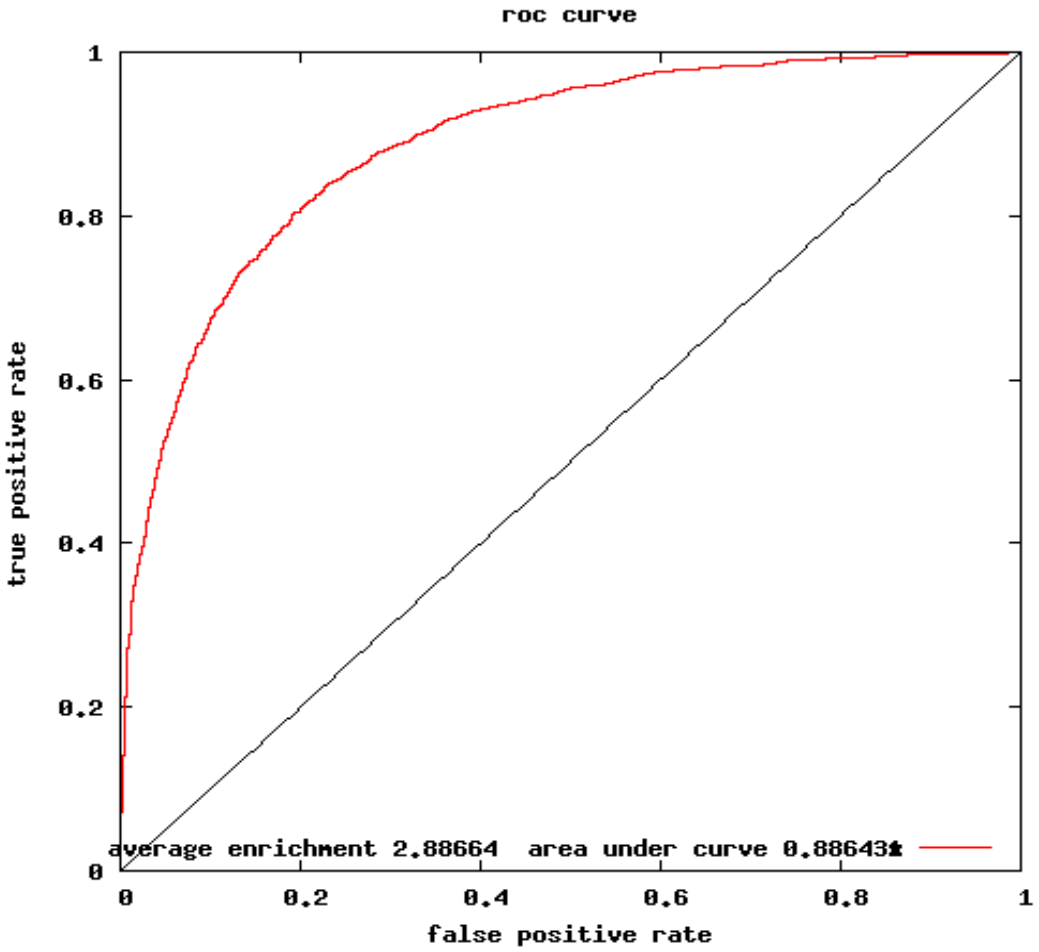
Outline

- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

CYP 3A4

- Metabolism of xenobiotics
- Oxidizes largest range of substrates of all CYPs
- Present in largest quantity in liver
- Involved in metabolism of ½ of the drugs used today
- Activates many toxins
- 3,438 actives
- 7,066 inactive

CYP3A4 Model Performance



Method	Average Enrichment	Number Features
ANN	2.78	298
SVM	2.67	392
KNN	2.78	73
Kohonen	2.71	94
DT	1.43	332
ANN/KNN/ Kohonen	**2.89**	*

$$Enrichment = \frac{TP/(TP+FP)}{P/(P+N)}$$

Outline

- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

Malaria

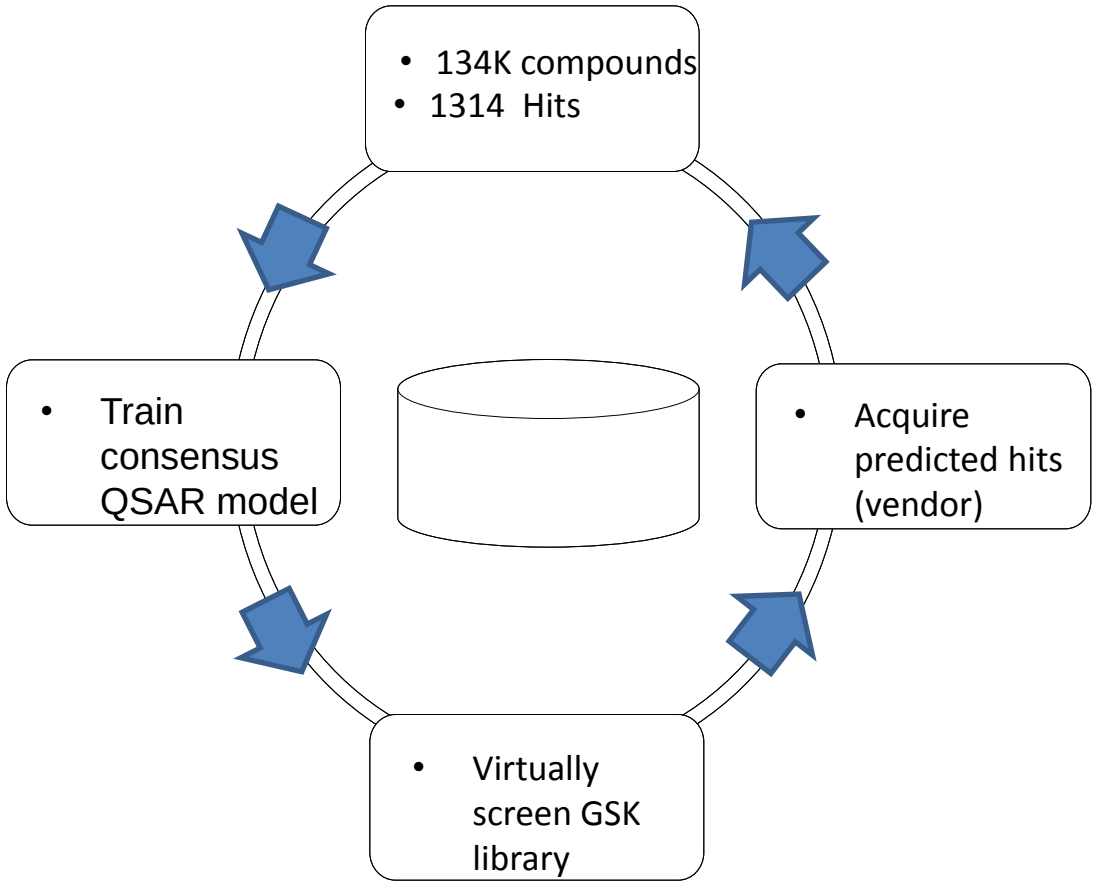
- parasitic disease
- high fevers, flu-like symptoms, anemia
- 250 million cases of fever and ~1 million deaths annually

- Parasite digests hemoglobin
  - free heme toxic to host cells
  - parasite crystallizes heme to hemozoin

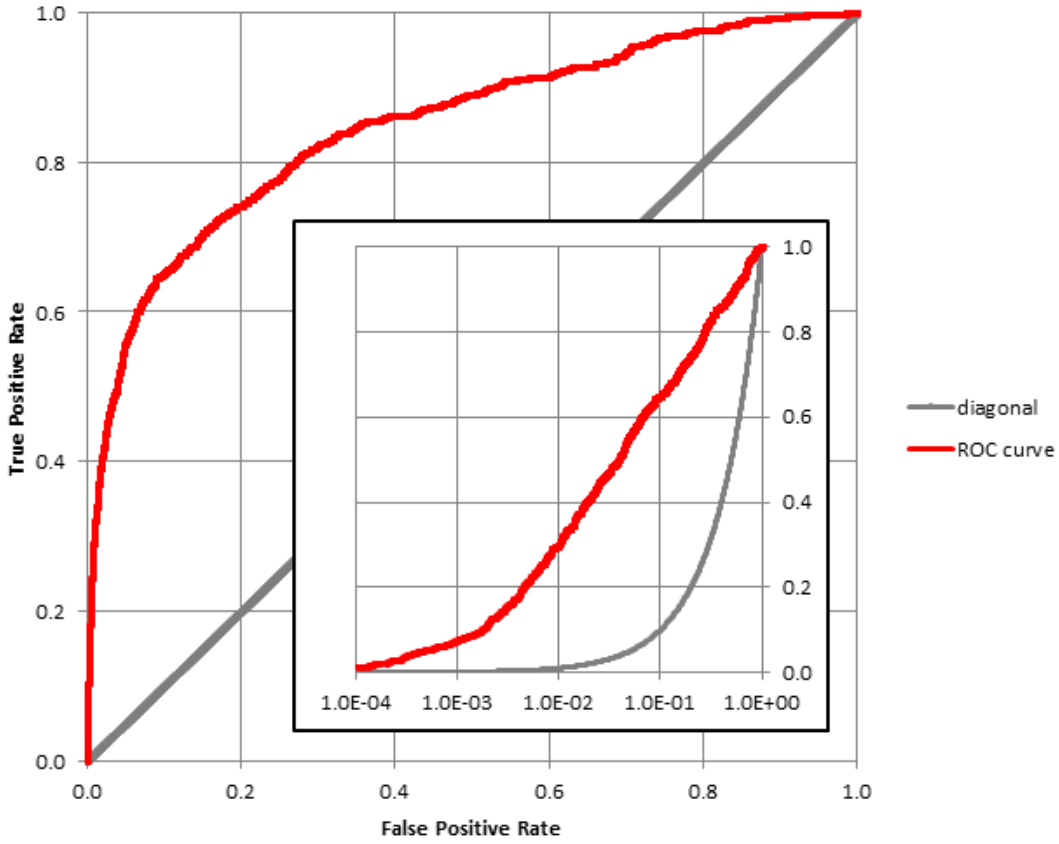
- hemozoin crystallization target of Malaria therapeutics

Malaria Model Optimization Workflow

- ~134,000 compounds screened for inhibition of hemozoin crystallization
- 1,314 inhibitors were found



Malaria Model Performance



Quality Measures:

Integral under ROC Curve	RMSD
0.85	0.29

Enrichments for different cutoffs:

Cutoff (False Positive Rate)	Enrichment
top 1%	33.2
top 2%	27.1
top 5%	19.0

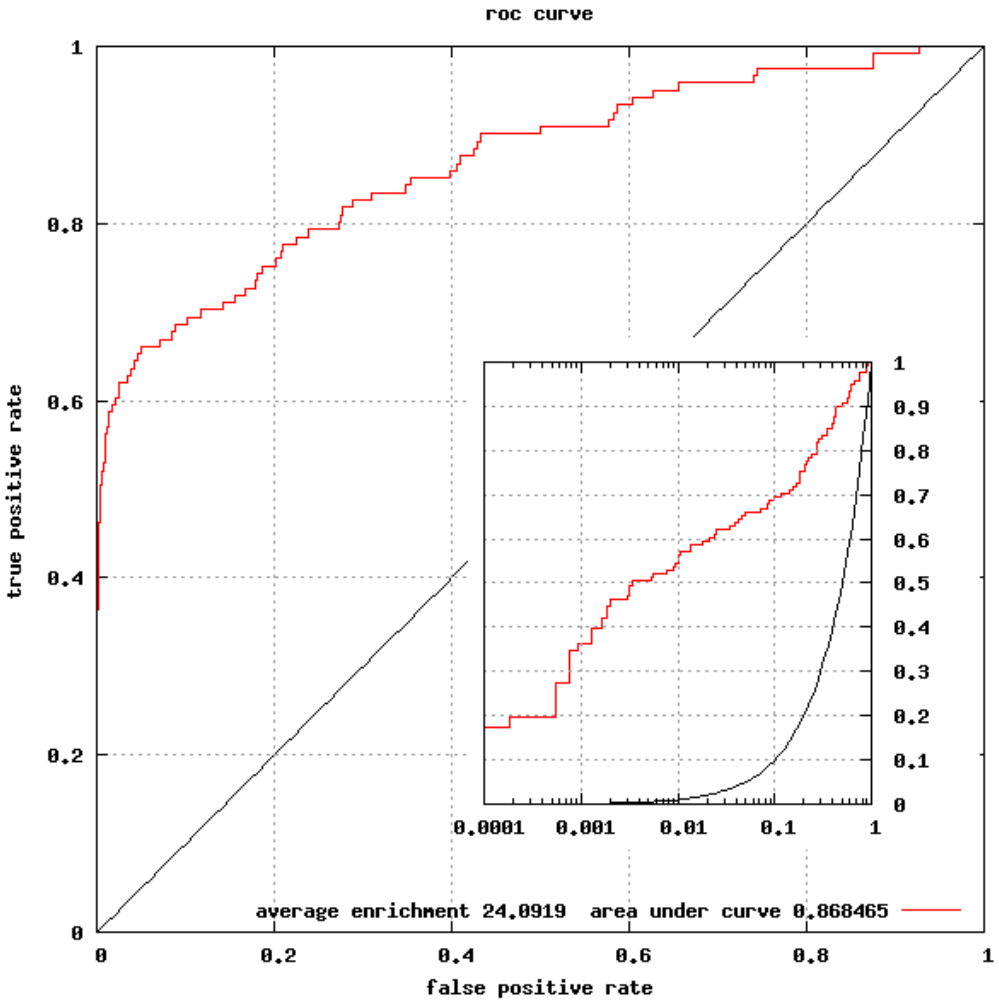
Outline

- Machine Learning Cheminformatics Framework
- QSPR
  - logP
- QSAR
  - $mGluR_5$
  - CYP 3A4
  - Malaria
  - KRas

KRas

- GTPase
- Indicated in Cancer
  - Leukemia
  - Colon
  - Pancreatic
  - Lung

KRas NMR Fragment Screen



- ~10k fragments screened
- hits with K_d's = 242
- Virtual screen of PubChem and Chembridge (~40m)
- Rank-list top 2500

Acknowledgments

Nils Woetzel
Mariusz Butkiewicz
Ralf Mueller
Matthew Spellings
Albert Omlor
Zollie White
Jens Meiler

Collaborators
Conn
Wright
Fesik

www.meilerlab.org

Funding
NIH 5T90DA022873-02 (Integrative Training in Therapeutic Discovery; PI: Marnett)
NIH 1R21MH082254 and 1R01MH090192 (NIMH; PI: Meiler)
NSF OCI-1122919 (Transformative Computational Science using CyberInfrastructure; PI: Lowe)