Kinome-wide Activity Models from Diverse High-Quality Datasets

Stephan C. Schürer*,[1] and Steven M. Muskal[2]

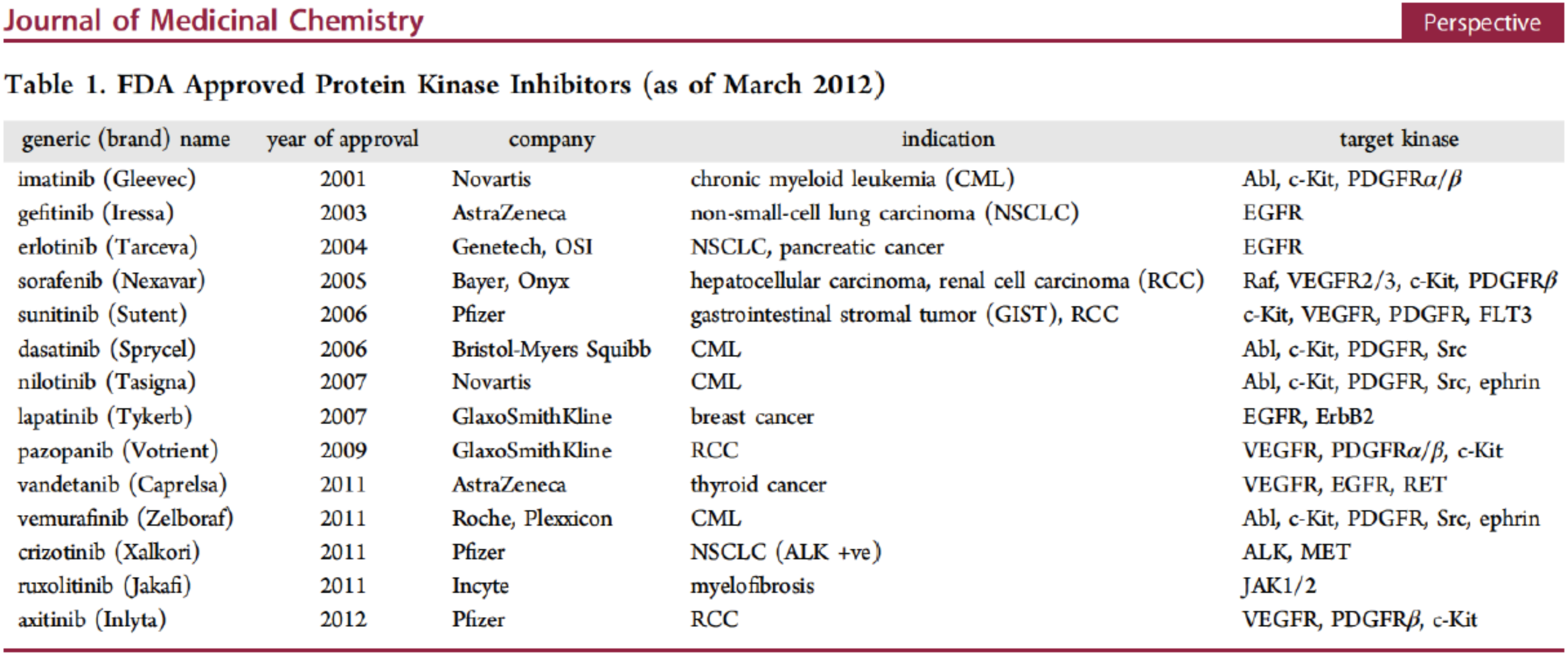
[1]Department of Molecular and Cellular Pharmacology, Miller School of Medicine and Center for Computational Science, University of Miami, Miami, FL 33136, USA.

[2]Eidogen-Sertanty, Inc. 3460 Marron Rd #103-475, Oceanside, CA 92056, USA.

Steven Muskal
Eidogen-Sertanty, Inc
smuskal@eidogen-sertanty.com

Eidogen·Sertanty

FDA Approved Protein Kinase Inhibitors



Journal of Medicinal Chemistry

Perspective

Table 1. FDA Approved Protein Kinase Inhibitors (as of March 2012)

generic (brand) name	year of approval	company	indication	target kinase
imatinib (Gleevec)	2001	Novartis	chronic myeloid leukemia (CML)	Abl, c-Kit, PDGFRα/β
gefitinib (Iressa)	2003	AstraZeneca	non-small-cell lung carcinoma (NSCLC)	EGFR
erlotinib (Tarceva)	2004	Genetech, OSI	NSCLC, pancreatic cancer	EGFR
sorafenib (Nexavar)	2005	Bayer, Onyx	hepatocellular carcinoma, renal cell carcinoma (RCC)	Raf, VEGFR2/3, c-Kit, PDGFRβ
sunitinib (Sutent)	2006	Pfizer	gastrointestinal stromal tumor (GIST), RCC	c-Kit, VEGFR, PDGFR, FLT3
dasatinib (Sprycel)	2006	Bristol-Myers Squibb	CML	Abl, c-Kit, PDGFR, Src
nilotinib (Tasigna)	2007	Novartis	CML	Abl, c-Kit, PDGFR, Src, ephrin
lapatinib (Tykerb)	2007	GlaxoSmithKline	breast cancer	EGFR, ErbB2
pazopanib (Votrient)	2009	GlaxoSmithKline	RCC	VEGFR, PDGFRα/β, c-Kit
vandetanib (Caprelsa)	2011	AstraZeneca	thyroid cancer	VEGFR, EGFR, RET
vemurafinib (Zelboraf)	2011	Roche, Plexxicon	CML	Abl, c-Kit, PDGFR, Src, ephrin
crizotinib (Xalkori)	2011	Pfizer	NSCLC (ALK +ve)	ALK, MET
ruxolitinib (Jakafi)	2011	Incyte	myelofibrosis	JAK1/2
axitinib (Inlyta)	2012	Pfizer	RCC	VEGFR, PDGFRβ, c-Kit

Reference: dx.doi.org/10.1021/jm3003203 I J. Med. Chem. 2012, 55, 6243–6262

Outline

- Kinase Data (KKB)

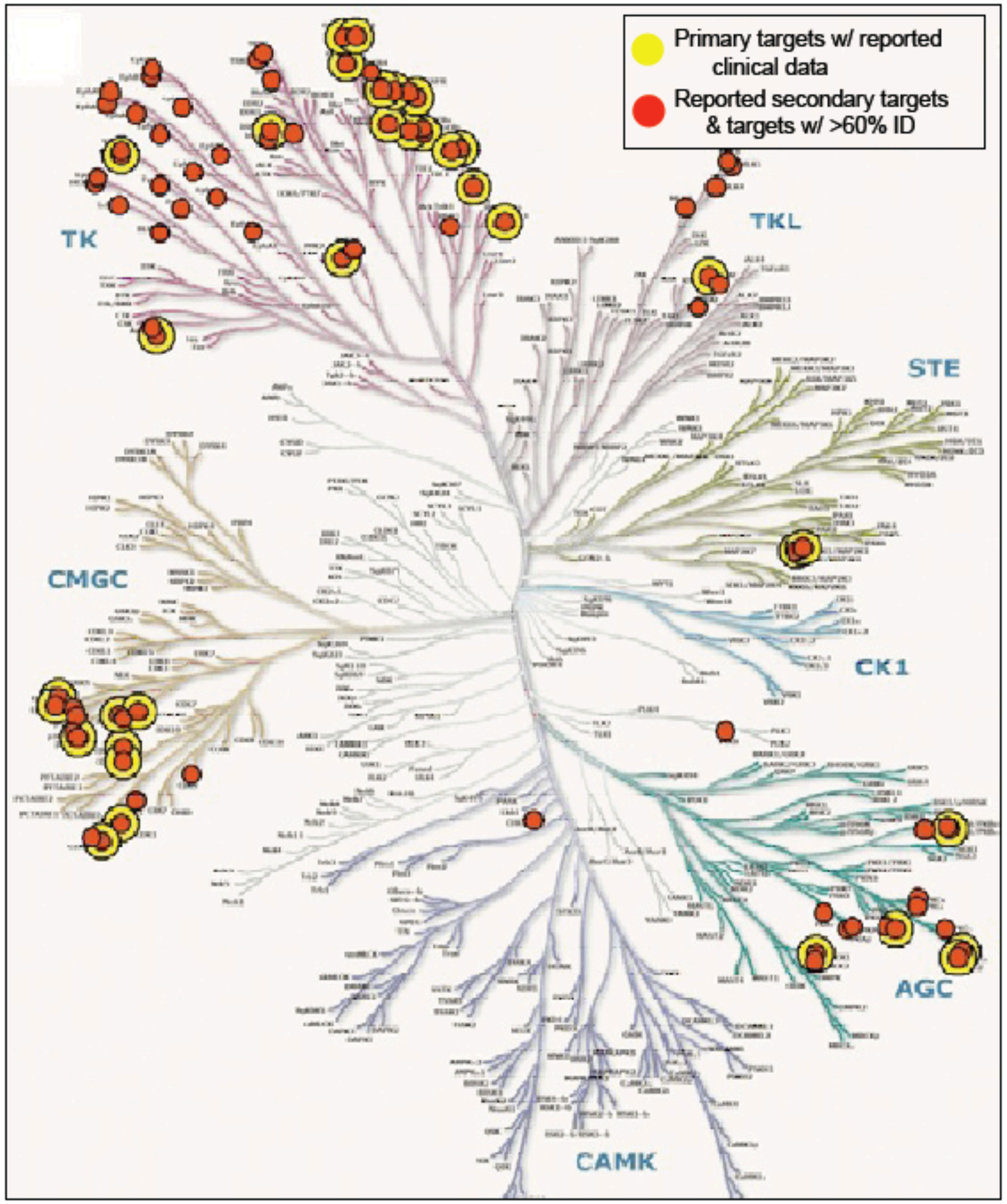
- Regression Models
  - ➡ Conclusions

- Naïve Bayesian Classifier Models
  - ➡ Conclusions

Kinase Knowledgebase (Q2 2012) - "Hot Targets"

Kinase Targets of Clinical Interest

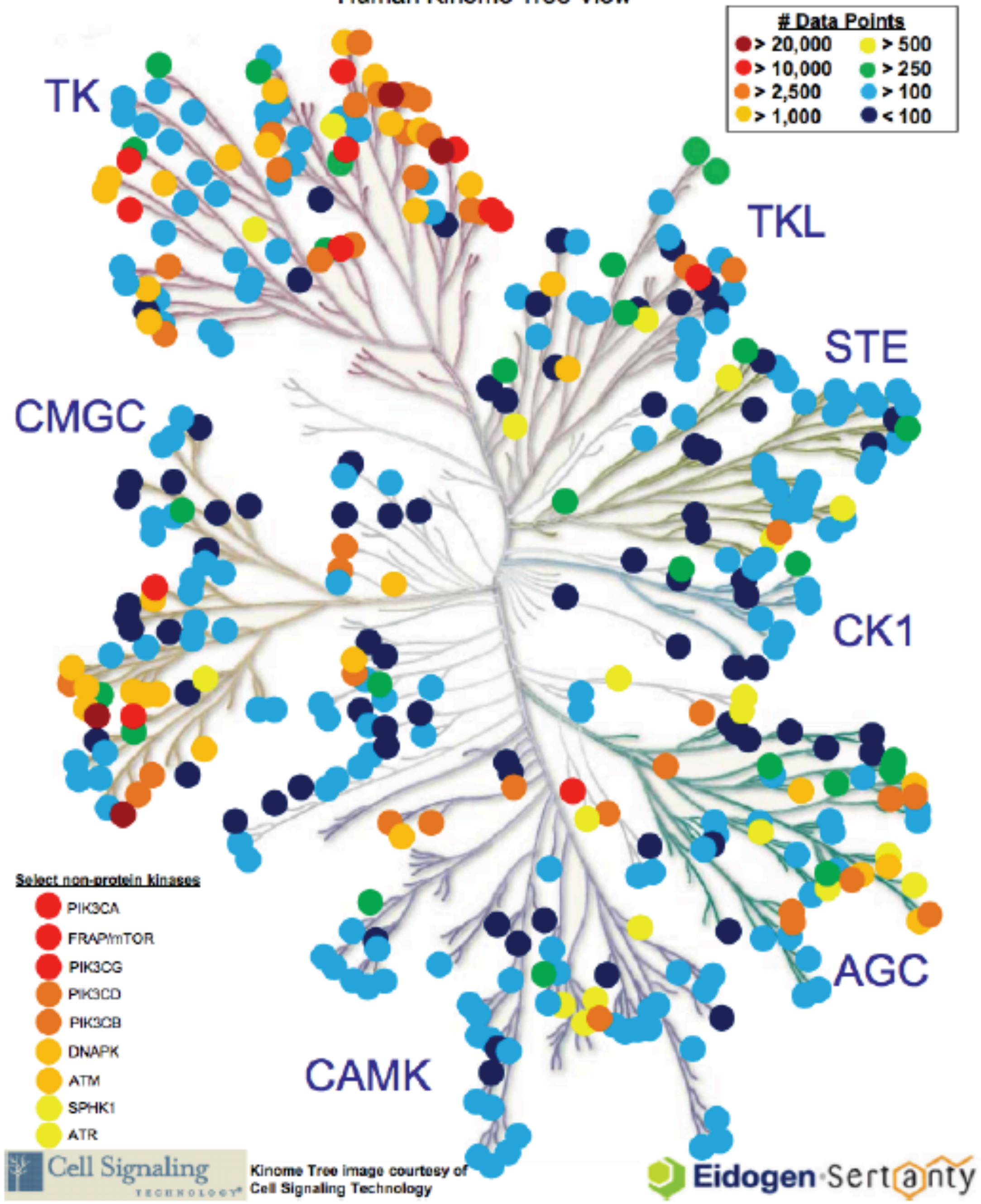
from Vieth *et al. Drug Disc. Today* **10**, 839 (2005).


Primary targets w/ reported clinical data
Reported secondary targets & targets w/ >60% ID
TK
TKL
STE
CMGC
CK1
AGC
CAMK

Eidogen-Sertanty KKB SAR Data Point Distribution

Kinases with Biological Activity Data in Kinase Knowledgebase Q2 2012 Release

Human Kinome Tree View



> 649,000 SAR data points curated from
> 7915 journal articles and patents

Kinase Knowledgebase (KKB) - Q2 2012

Kinase inhibitor structures and SAR data mined from

> 7915 journal articles/patents

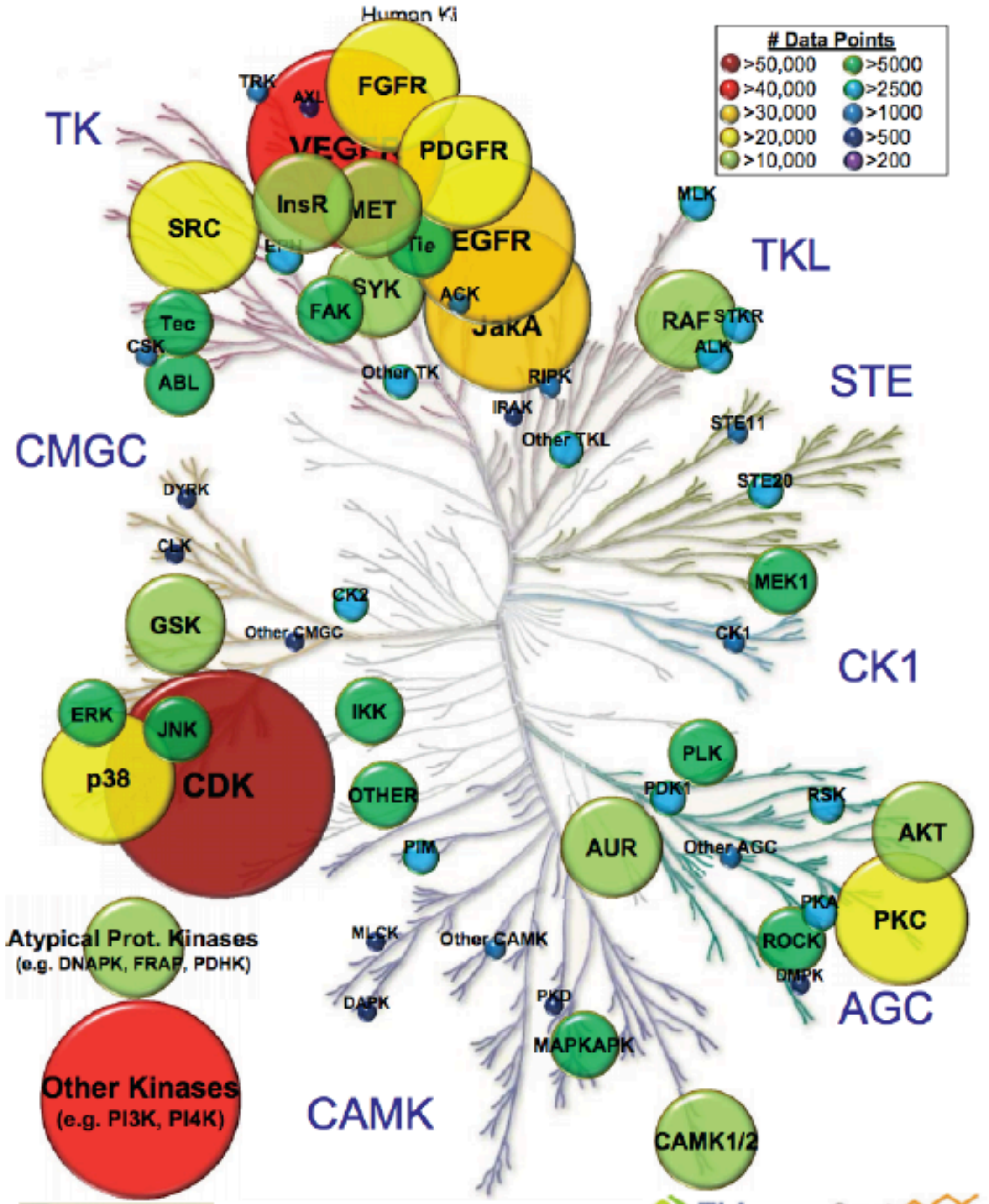
- **KKB Content Summary (Q2 2012):**

  # of kinase targets: **> 480**
  # of SAR Data points: **> 649,000**
  # of **unique** kinase molecules with SAR data: **>241,000**
  # of annotated assay protocols: **>25,472**
  # of all kinase inhibitors (with or without bio-activity data): **> 586,000**

- **KKB Growth Rate:**
  - Average **15-20K** SAR data points added per quarter
  - Average **20-30K** unique structures added per quarter

Kinase Summary Statistics - Q2 2012

Articles covered:	2,307	(+ 30)
Patents and patent applications covered:	5,608	(+ 93)
Total Number of Bio-activity data points:	649,384	(+ 31,602)
Total Number of unique molecules:	586,610	(+ 8601)
Total Number of unique molecules w/ assay data:	241,680	(+ 8601)
Total Number of assay protocols:	25,472	(+ 322)



Cell Signaling

Kinome Tree image courtesy of Cell Signaling Technology

Eidogen·Sertanty

Targets with largest increase in Data Points in Q2-12

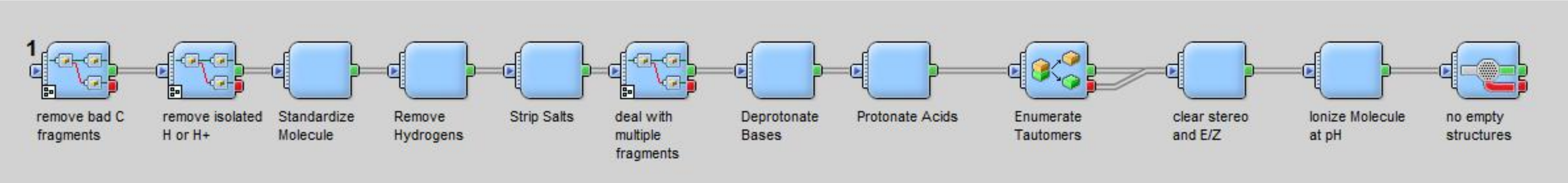
Target	# Data Points added
FGFR3	4626
KDR	4482
FGFR1	4446
FLT3	3047
TTK	1634
FGFR2	1428
FGFR4	1364
PIK3CA	1254
PIK3CD	955
JAK3	920
MTOR	827
JAK2	533
PTK2	463
RPS6KB1	425
JAK1	387
ALK	361
AKT1	357
ROCK2	335
SYK	305
BRAF	268
GSK3B	251
LRRK2	219
EGFR	211
BTK	197
TYK2	188
IRAK4	178
PIK3CB	158
PIK3CG	147
PIM1	147
IKBKB	130
CDK2	117
MAPK1	108
ERBB2	94
CSF1R	73
MET	72
TGFBR1	68
PLK1	55
PIM3	54
CDK9	52

Kinase Data Used in this Study - Q4 2009 [Q2 2012]

- Biological Activity Data Points: > 437,000 [> 649K]
- Unique kinase molecules w/assay data: >162,000 [> 241K]
- Unique kinase molecules patents/articles: > 507,000 [> 586K]
- Number of unique kinase targets with assay data: 394 [480]
- Number of annotated assay protocols: 20,593 [25,472]

Data Pre-Processing

- Starting point: KKB-Q2 2009
- Only enzymatic (homogeneous) assays with defined target
- Only high quality data (IC50, Ki, Kd)
- Standardized chemical structures (salt forms, stereochemistry, E/Z geometry, tautomers, ionization)
- Kinase target Entrez Gene names and SwissProt accessions
- 233,667 unique data points (411 kinases)
- 126,114 unique chemical structures



KinomeScan Data (Experimental Validation Data)

- NIH HMS LINCS DataBase (Harvard Medical School LINCS center) http://lincs.hms.harvard.edu/resources/software/hms-lincs-database/
  - The LINCS program develops a library of molecular signatures based on gene expression and other cellular changes in response to perturbing agents across a variety of cell types using various high-throughput screening approaches
- 25,064 total datapoints downloaded:
  - 60 unique compounds (43 with defined/known chemical structure) against 486 targets
  - Kinase activity screened at 10 µM concentration
  - Targets mapped to KKB targets by UniProt accessions
  - Data not in KKB
- Result: 4,796 datapoints from 43 compounds

Outline

- Kinase Data (KKB)

- Regression Models
  - ➡ Conclusions

- Naïve Bayesian Classifier Models
  - ➡ Conclusions

Quantitative Regression Models

- K nearest neighbors (kNN) [20 nn, Gaussian weighting 0.5]
- Partial least squares (PLS) [components restricted to 20]
- Works best with congeneric series
- Sensitive to outliers and noise
- 168 kinase datasets have ≥ 20 Dps
- PipelinePilot - ECFP4 (circular) fingerprints as descriptors
- Full 10 fold cross validation for both PLS and kNN (20 repetitions), report R^2, q^2, RMSE

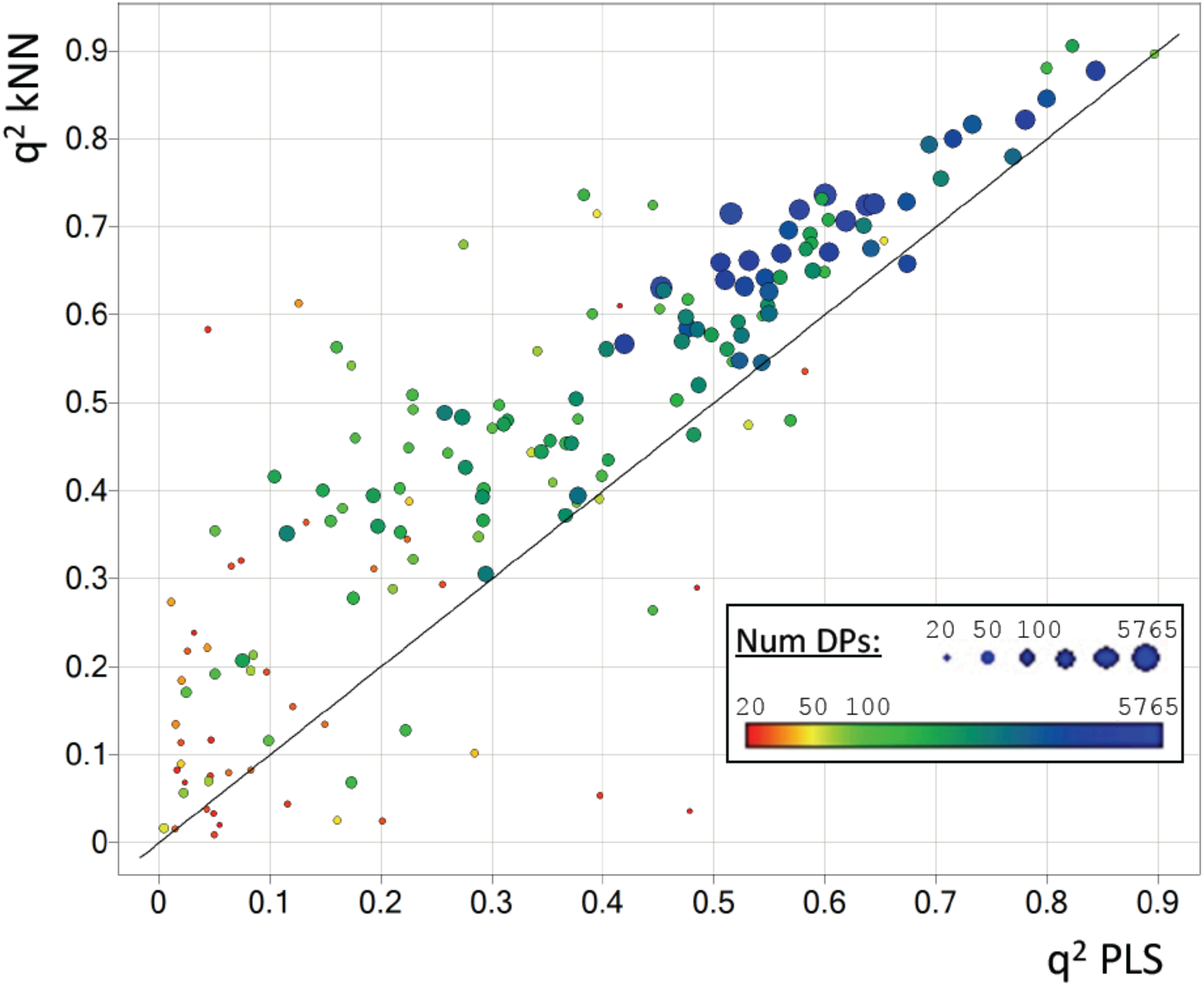


Figure 5. Quantitative regression models developed from 168 kinase datasets. q^2 values of kNN vs. PLS regression models and the number of kinase activity data points indicated by the circle size and color

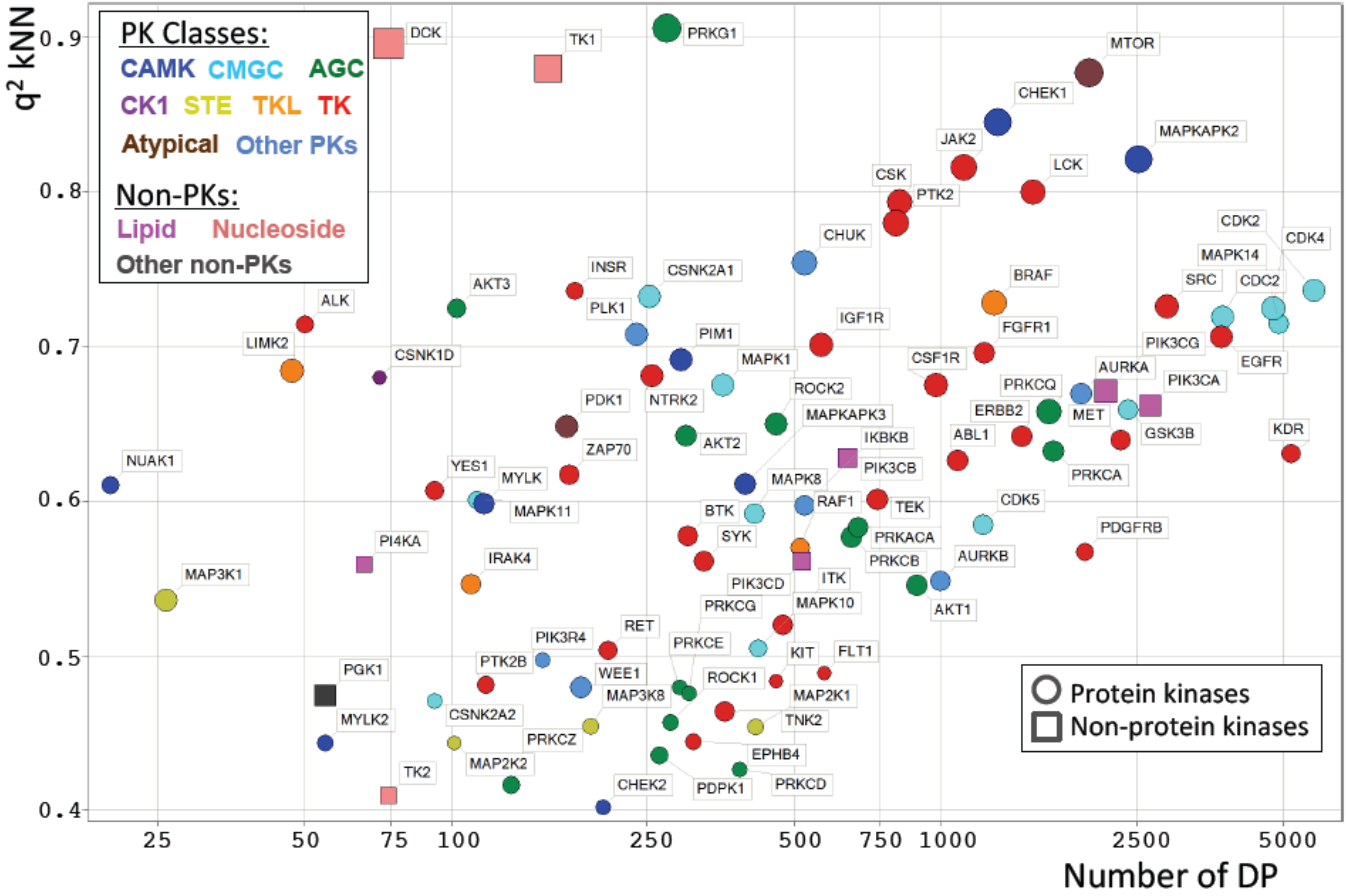
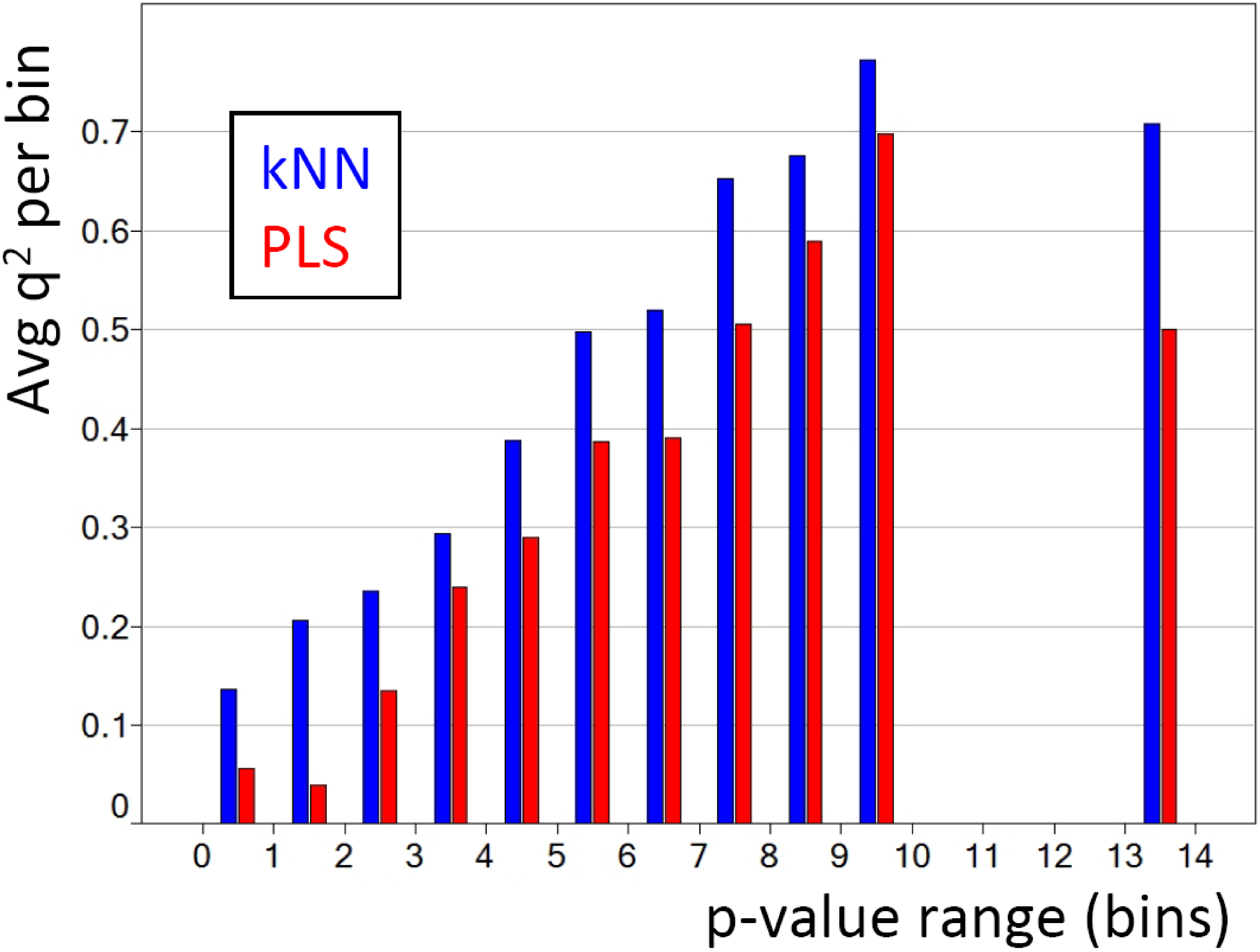


Figure 7. kNN and PLS activity predictors for 91 kinases (q^2 kNN > 0.4 and q^2 PLS > 0.25) by the number of data points; datasets include protein and non-protein kinases and all major kinase groups. kNN q^2 is shown by number of (unique) structure-data points. Symbol indicates protein vs. non-protein kinase, size is scaled by PLS q^2, colored by kinase group, and annotated by HUGO kinase gene symbol.


Avg q² per bin
kNN
PLS
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
p-value range (bins)

Conclusions - Regression Models

- Work well for many kinase data sets
- kNN performs slightly better than PLS
- Larger numbers of data points improve both PLS and kNN models
- Best results for kinases with ≥ 50 data points
- Regression models improve with increased activity range

Outline

- Kinase Data (KKB)

- Regression Models
  - ➡ Conclusions

- Naïve Bayesian Classifier Models
  - ➡ Conclusions

Laplacien-modified Naïve Bayesian Classifiers

- Scale linearly, work in high-dimensional spaces (no over-fitting), good for structurally diverse cmpds, multiple activity classes, robust to outliers

- Define data sets by unique kinase gene IDs with active compounds defined as pIC50 ≥ 6

- 189 kinase data sets with at least 10 active molecules

- Data were treated in two ways:
  - Known Active - Known Inactive (KA-KI)
  - Presumed Inactive (PI): 126,114 unique chemical structures - N_{inact}

- ECFP4 (circular) fingerprints

- Leave-one-out cross-validation and repetitive train/test evaluation measuring ROC and enrichment factors

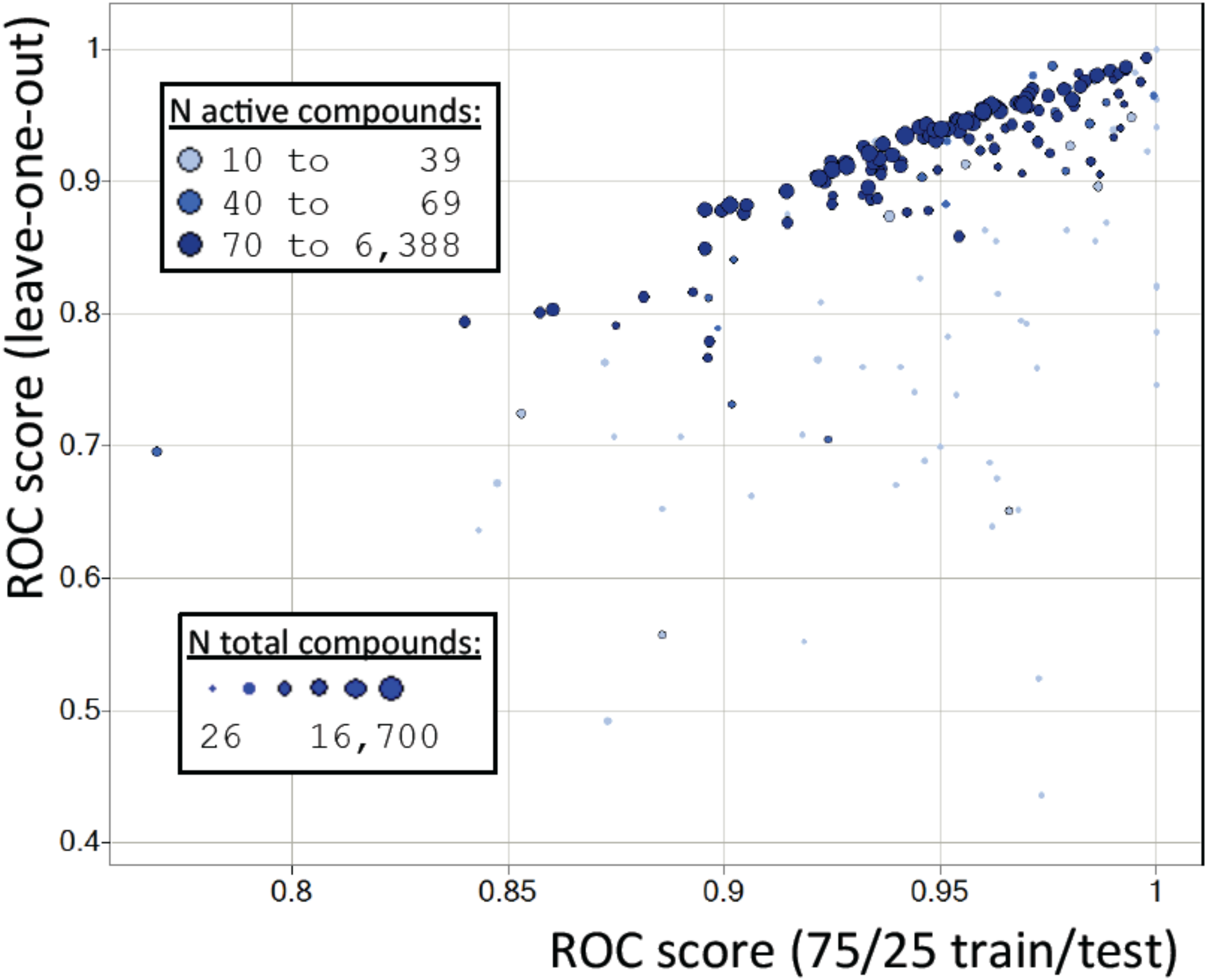


Figure S1. ROC score of leave-one-out vs. randomized 75/25 (train/test) cross validation for 187 kinase KA-KI classifiers with at least 10 active samples. Dot size is scaled by the number of total compounds in each dataset and colored by the number of actives. 129 classifiers are shown that correspond to datasets with at least 40 actives and 111 with at least 70 actives. ROC scores for randomized 75/25 training/test validations are averages of 10 repetitions. Compare table S1.

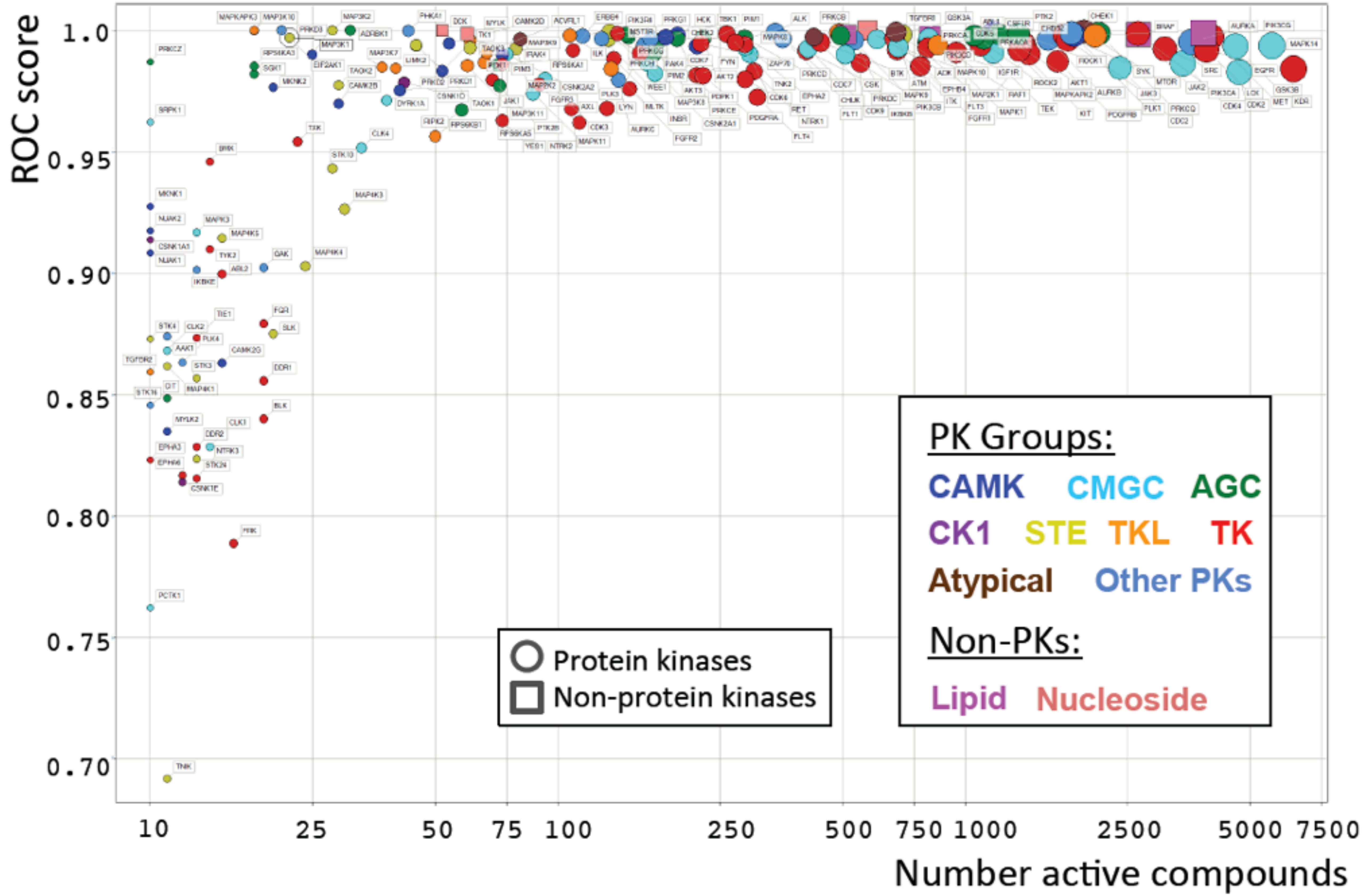


Figure S2. Characterization of all 189 kinase KA-PI protein and non-protein kinase classifiers, including all major protein kinase groups. ROC scores are shown as a function of active samples. Shape by protein vs. non-protein kinase, color-coded by kinase group, scaled by number of active data points, annotated by HUGO kinase gene symbol. ROC scores increase significantly for classifiers based on datasets with more than 25 active compounds.

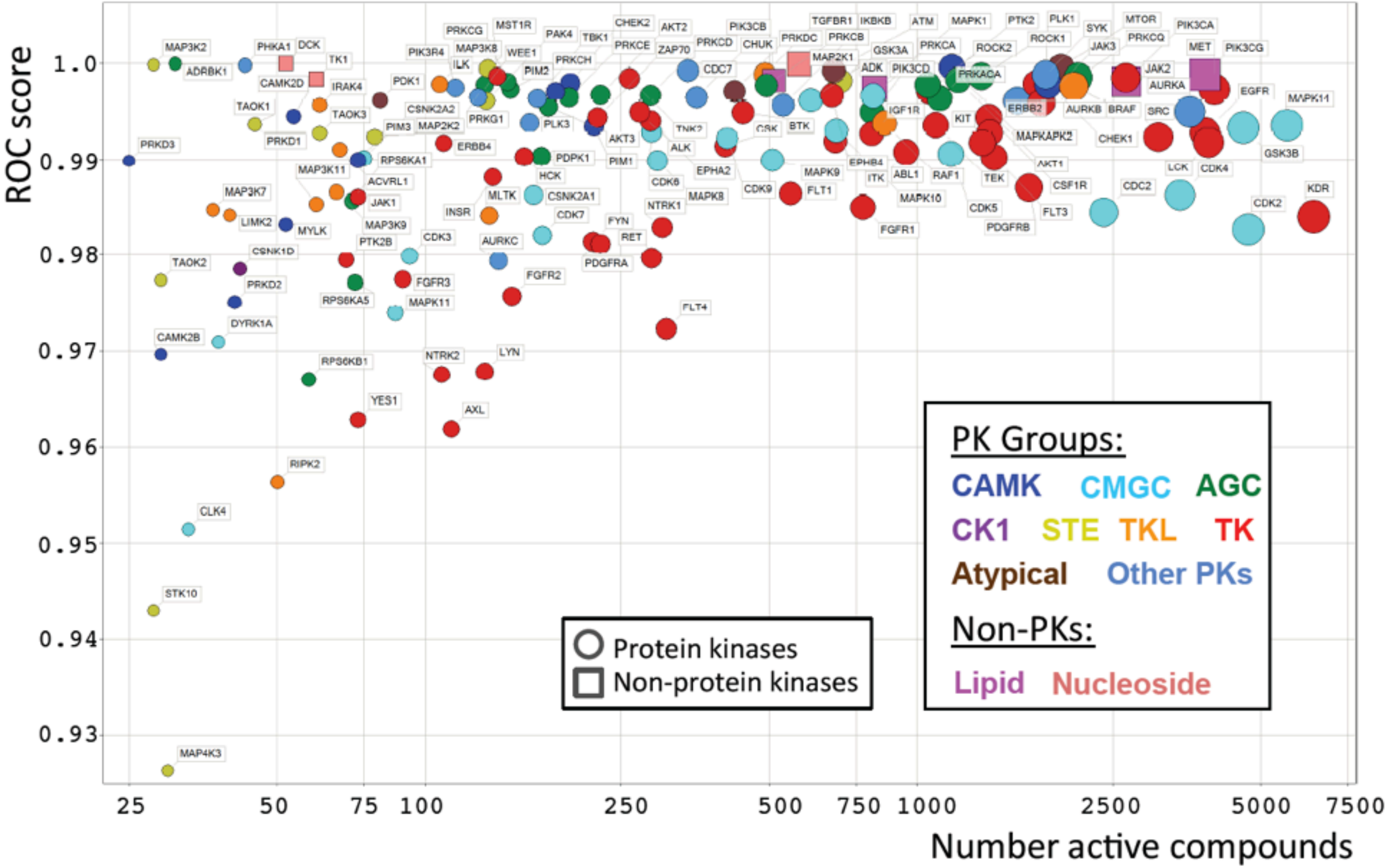


Figure 1. Characterization of 141 kinase KA-PI protein and non-protein kinase classifiers, including all major protein kinase groups. ROC scores are shown as a function of active samples. Shape by protein vs. non-protein kinase, color-coded by kinase group, scaled by number of active data points, annotated by HUGO kinase gene symbol.

* At least 25 actives

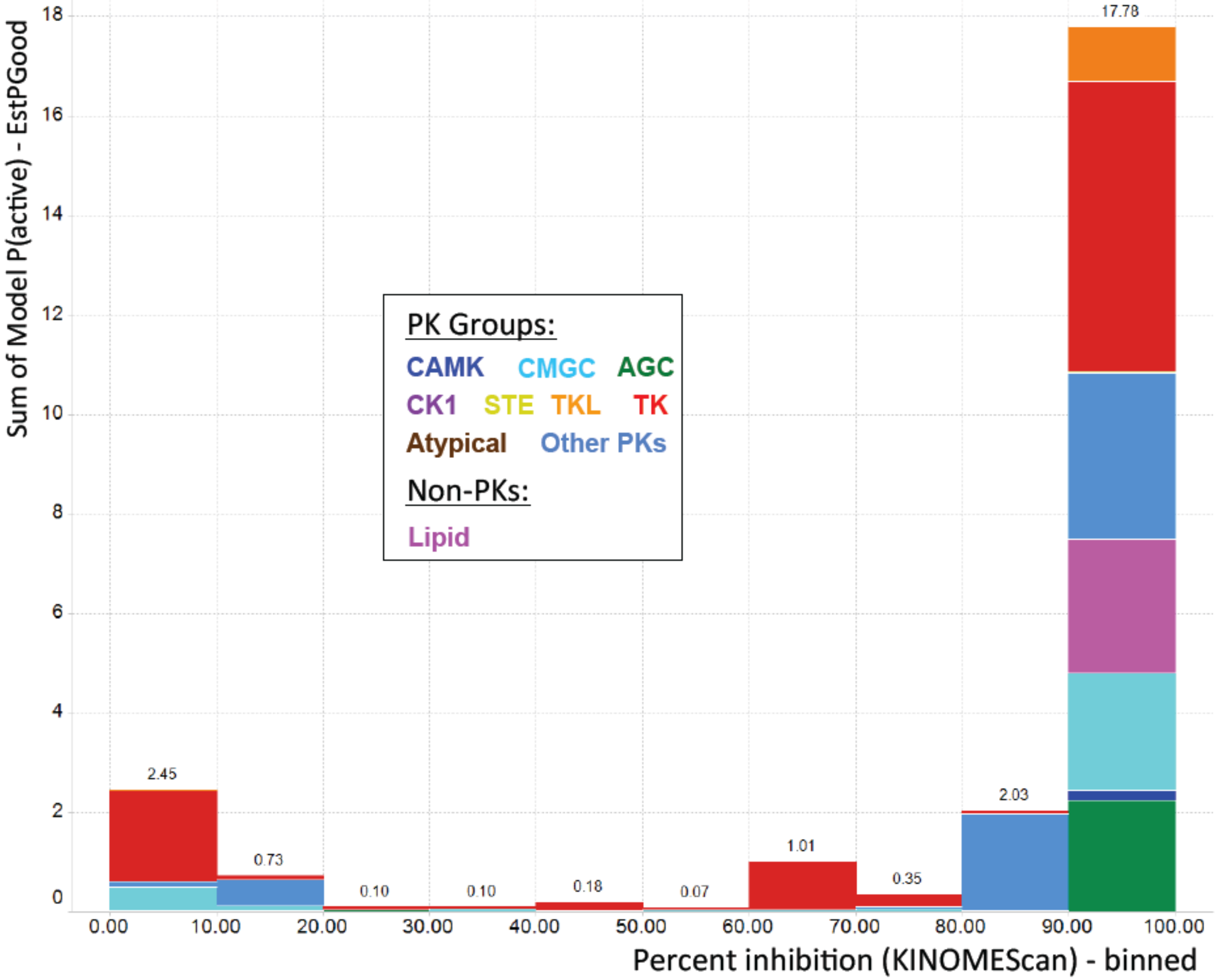


Figure 9. Aggregated predicted probabilities (EstPGood) of compounds being against kinases (based on the KA-PI classifiers) as a function of the actual KINOMEScan percent inhibition ranges; by category of kinase group and protein vs. non-protein kinase; 4,796 activity data points for 43 compounds mapped to KA-PI models (not all compounds tested against the same number of targets).

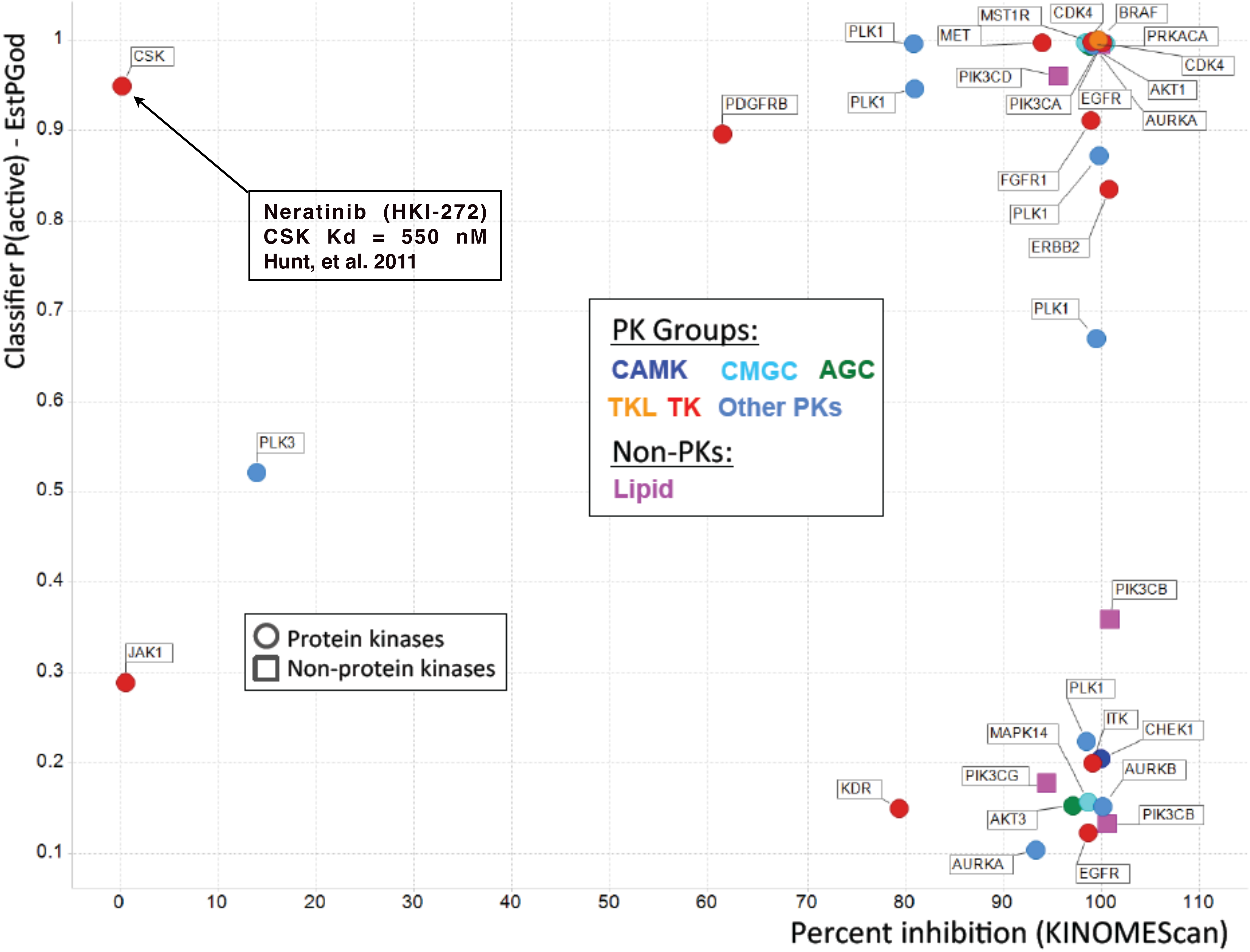


Figure 8. Probability (EstPGood) of compounds being active against a kinase based on KA-PI kinase classifiers and actual KINOMEScan percent inhibition values (at 10 μM); compare supporting table S6. Kinase classified by groups and protein vs. non-protein kinases.

Conclusions - Classification Models

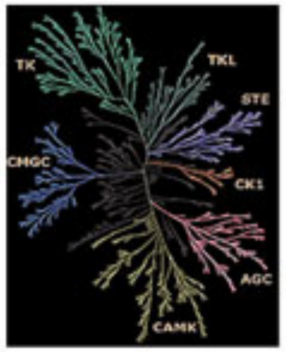
- Work very well for known actives - known inactives (KA-KI)
- Relevant and applicable for real world, highly unbalanced data sets (KA-PI)
- Leave-one-out ROC is a good guide of model quality
- Naïve Bayesian classification is excellent for the majority of kinases (>140)
- Performance increases markedly with ≥ 50 active compounds

➡ Very useful for virtual screening and rapid profiling

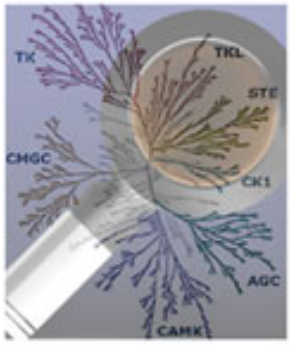
Eidogen's iPhone, iPad, and Android Apps

iProtein

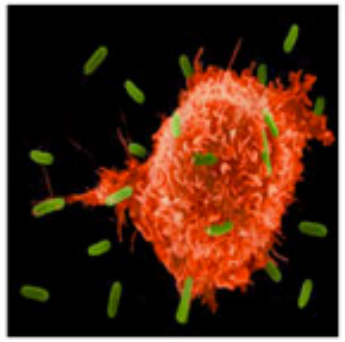
iKinase



iKinasePro



iOncology



Mobile Reagents

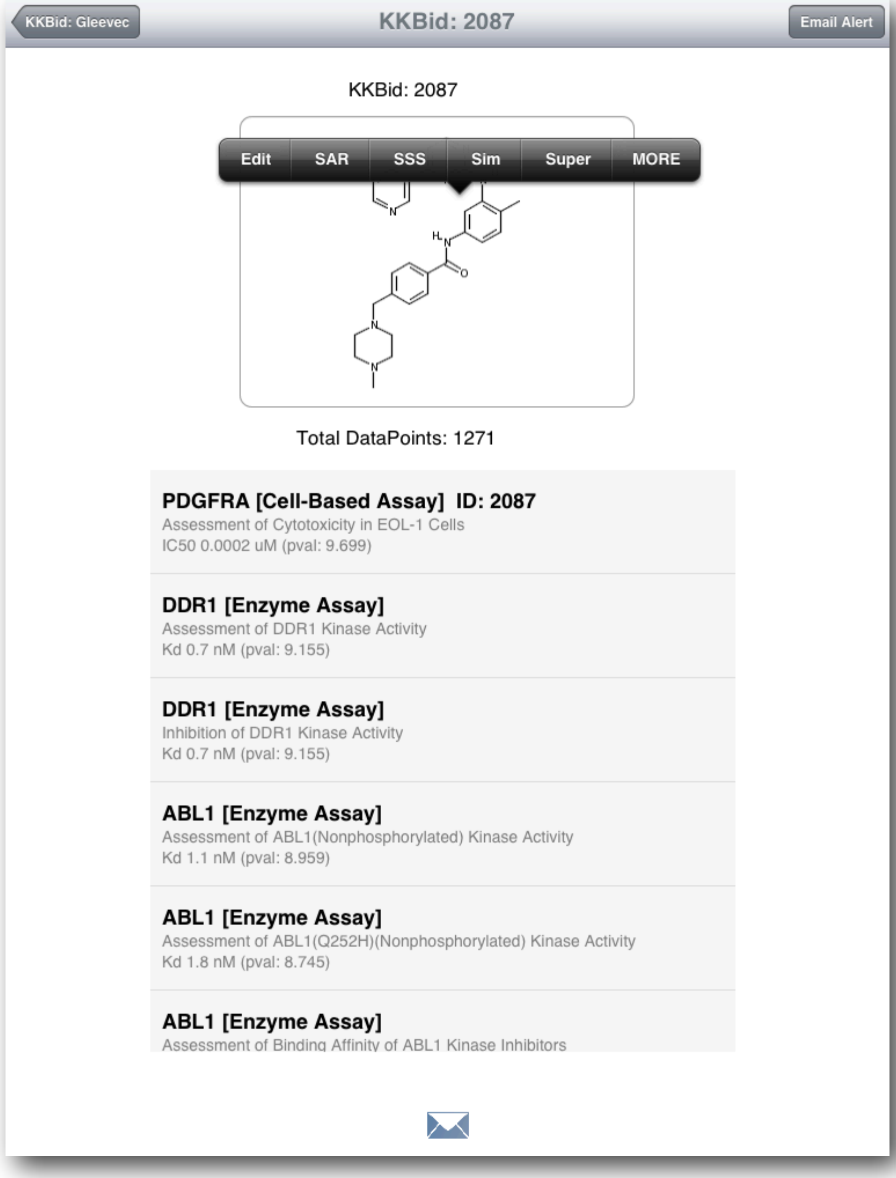
Reaction101

Yield101

SPRESImobile

See: eidogen.com, kinasedb.com or kinasedata.com

MobileApps Support Real Scientific Workflows



KKBid: 2087

Total DataPoints: 1271

PDGFRA [Cell-Based Assay] ID: 2087
Assessment of Cytotoxicity in EOL-1 Cells
IC50 0.0002 uM (pval: 9.699)

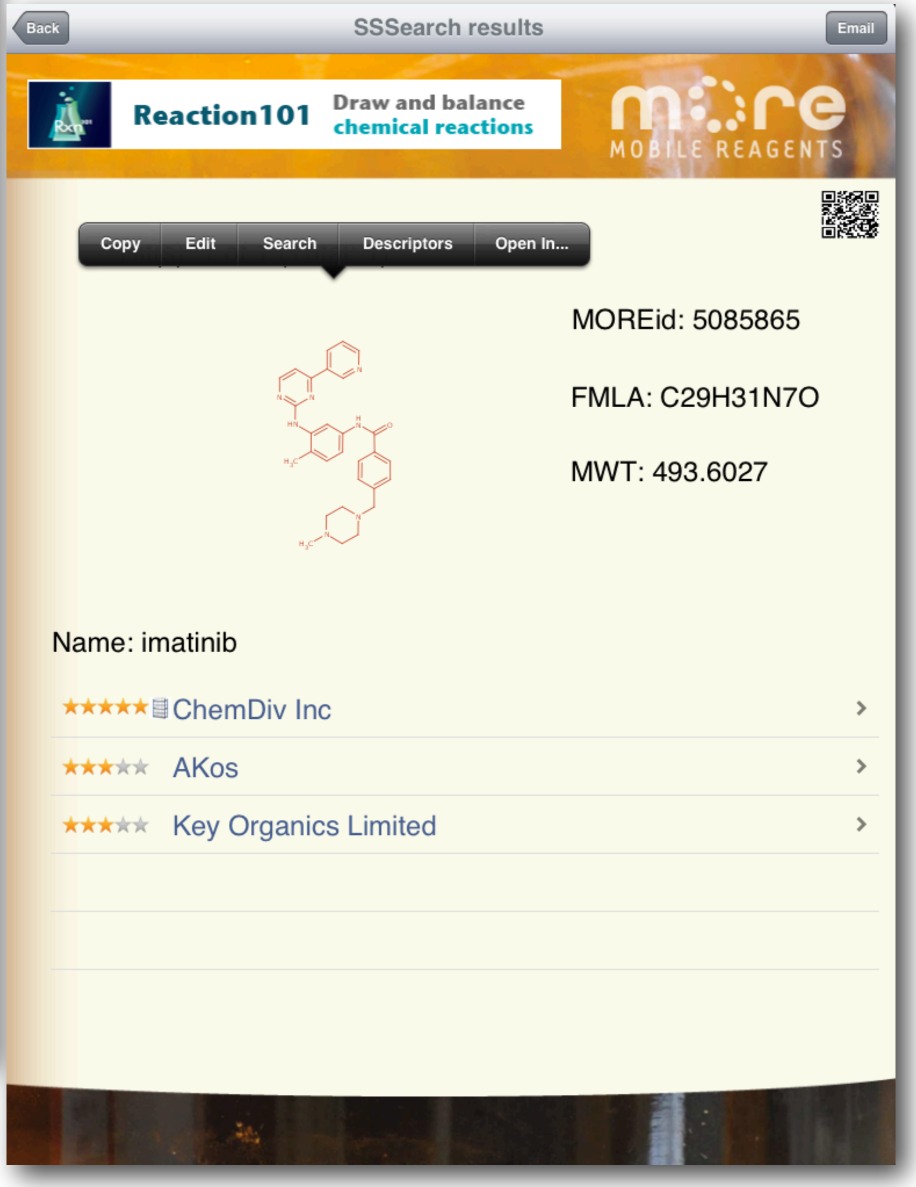
DDR1 [Enzyme Assay]
Assessment of DDR1 Kinase Activity
Kd 0.7 nM (pval: 9.155)

DDR1 [Enzyme Assay]
Inhibition of DDR1 Kinase Activity
Kd 0.7 nM (pval: 9.155)

ABL1 [Enzyme Assay]
Assessment of ABL1(Nonphosphorylated) Kinase Activity
Kd 1.1 nM (pval: 8.959)

ABL1 [Enzyme Assay]
Assessment of ABL1(Q252H)(Nonphosphorylated) Kinase Activity
Kd 1.8 nM (pval: 8.745)

ABL1 [Enzyme Assay]
Assessment of Binding Affinity of ABL1 Kinase Inhibitors



MOREid: 5085865

FMLA: C29H31N7O

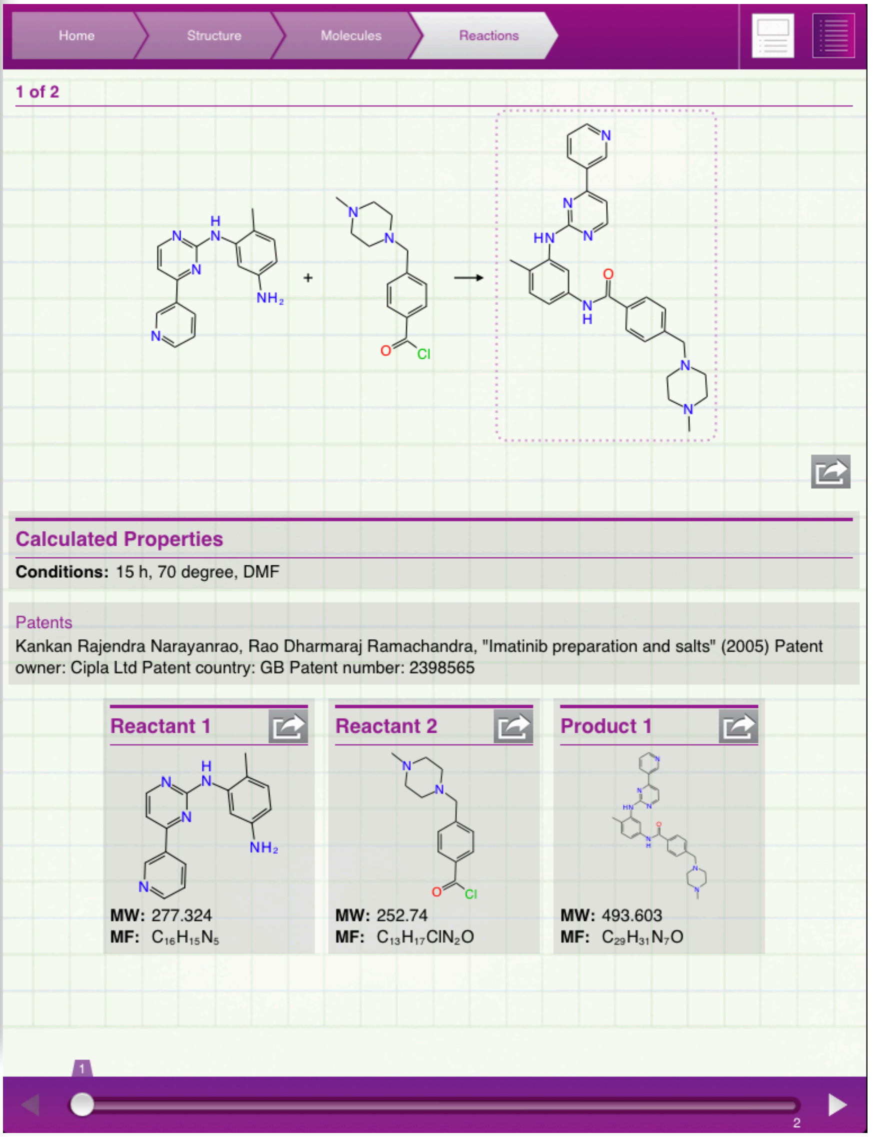
MWT: 493.6027

Name: imatinib

★★★★★ ChemDiv Inc

★★★★★ AKos

★★★★★ Key Organics Limited



Calculated Properties

Conditions: 15 h, 70 degree, DMF

Patents

Kankan Rajendra Narayanrao, Rao Dharmaraj Ramachandra, "Imatinib preparation and salts" (2005) Patent owner: Cipla Ltd Patent country: GB Patent number: 2398565

Reactant 1	Reactant 2	Product 1
MW: 277.324 **MF:** $C_{16}H_{15}N_5$	**MW:** 252.74 **MF:** $C_{13}H_{17}ClN_2O$	**MW:** 493.603 **MF:** $C_{29}H_{31}N_7O$

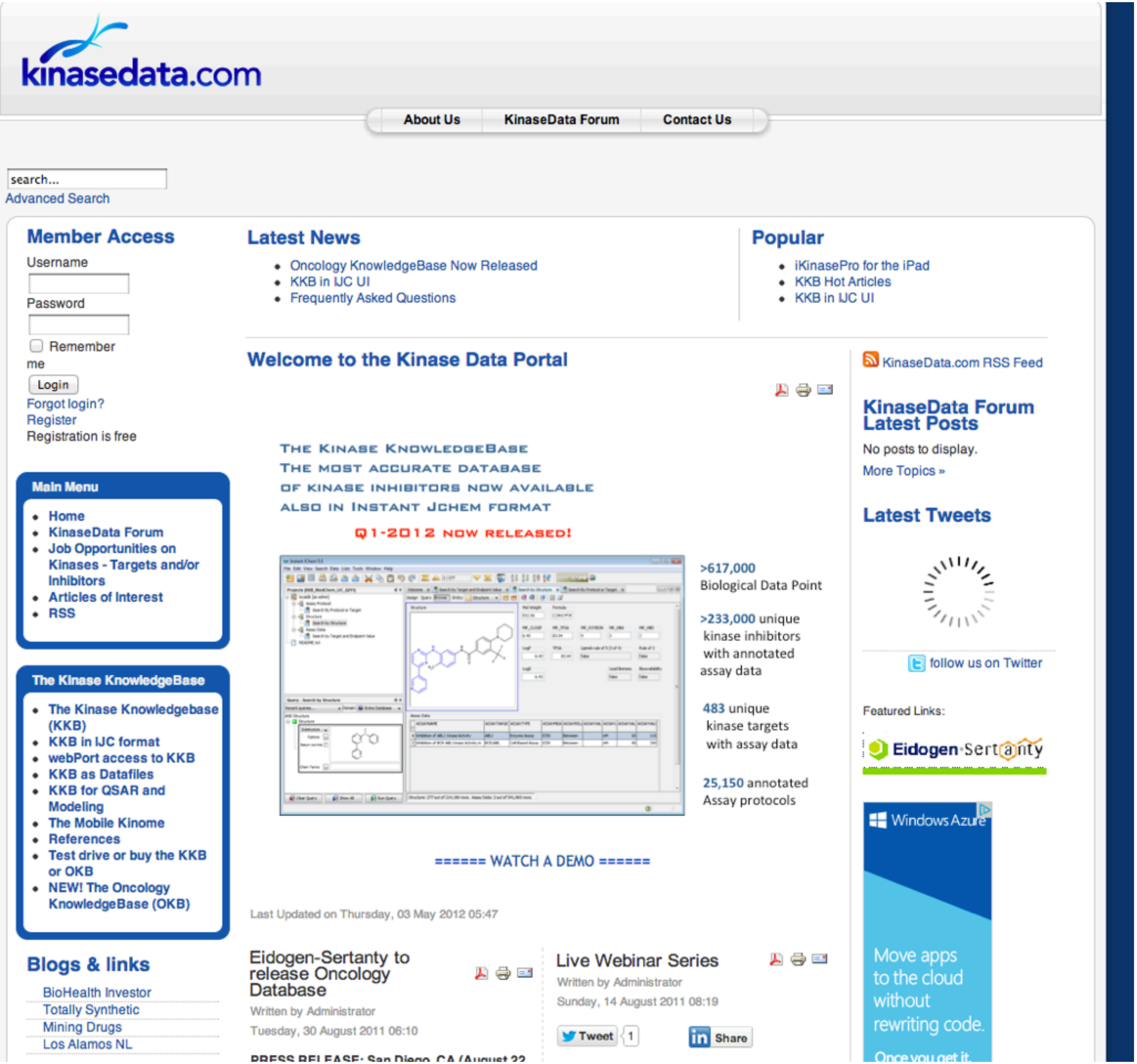
Bioactivity searching (e.g. kinase SAR)

Commercial availability

Synthesis planning

Acknowledgements

- Dr. Rajan Sharma and Prof. Stephan Schurer
  - iKinase iKinasePro iOncology
- Dr. Maurizio Bronzetti
  - Mobile Reagents Reaction101 SPRESImobile
- Dr. Alex Clark
  - MMDSLib: Mobile Reagents iProtein
  - Reaction101 Yield101 SPRESImobile
- Dr. Tony Yuan
  - JSDraw: iKinasePro Mobile Reagents iProtein
- Dr. Peter Löw, Dr. Josef Eiblmaier, et al.
  - SPRESImobile



kinasedata.com

About Us | KinaseData Forum | Contact Us

search...

Advanced Search

Member Access

Username

Password

□ Remember me

Login

Forgot login?

Register

Registration is free

Main Menu

- Home
- KinaseData Forum
- Job Opportunities on Kinases - Targets and/or Inhibitors
- Articles of Interest
- RSS

The Kinase KnowledgeBase

- The Kinase Knowledgebase (KKB)
- KKB in IJC format
- webPort access to KKB
- KKB as Datafiles
- KKB for QSAR and Modeling
- The Mobile Kinome
- References
- Test drive or buy the KKB or OKB
- NEW! The Oncology KnowledgeBase (OKB)

Blogs & links

- BioHealth Investor
- Totally Synthetic
- Mining Drugs
- Los Alamos NL

Latest News

- Oncology KnowledgeBase Now Released
- KKB in IJC UI
- Frequently Asked Questions

Popular

- iKinasePro for the iPad
- KKB Hot Articles
- KKB in IJC UI

Welcome to the Kinase Data Portal

THE KINASE KNOWLEDGEBASE
THE MOST ACCURATE DATABASE
OF KINASE INHIBITORS NOW AVAILABLE
ALSO IN INSTANT JCHEM FORMAT

Q1-2012 NOW RELEASED!

>617,000 Biological Data Point

>233,000 unique kinase inhibitors with annotated assay data

483 unique kinase targets with assay data

25,150 annotated Assay protocols

====== WATCH A DEMO ======

Last Updated on Thursday, 03 May 2012 05:47

Eidogen-Sertanty to release Oncology Database

Written by Administrator

Tuesday, 30 August 2011 06:10

PRESS RELEASE: San Diego, CA (August 22

Live Webinar Series

Written by Administrator

Sunday, 14 August 2011 08:19

Tweet 1 | Share

KinaseData.com RSS Feed

KinaseData Forum Latest Posts

No posts to display.

More Topics »

Latest Tweets

follow us on Twitter

Featured Links:

Eidogen-Sertanty

Windows Azure

Move apps to the cloud without rewriting code.

Once you get it

Naïve Bayesian with Laplacien Correction

Naïve Bayes (features are conditionally independent):

$$P(C \mid F_1, \ldots, F_n) = k \prod_{i=1}^{n} \frac{P(C \mid F_i)}{P(C)}$$

$$P(C) = \frac{A}{T}$$

$$P(C \mid F_i) = \frac{A_{Fi}}{T_{Fi}}$$

A: Number of active samples
T: Total number of sample
A_{Fi}: Active sample with feature Fi
T_{Fi}: Total sample with feature Fi

Adding virtual samples for each feature:

$$P_{corr}(C \mid F_i) = \frac{A_{Fi} + P(C)K}{T_{Fi} + K}$$

Estimating active virtual samples using baseline probability

Sample frequency 1/P(C) or T/A (Laplacien correction):

$$P_{final}(C \mid F_i) = \frac{A_{Fi} + 1}{T_{Fi} + T/A}$$

Pipeline Pilot implementation:

$$\log(P(C \mid F_1, \ldots, F_n)) = K + \sum_{i=1}^{n} \log(P_{final}(C \mid F_i))$$

Classification Model Evaluation

Receiver operating characteristic:

$$ROC = \frac{S}{1 - SP}$$

Sensitivity (true positive rate):

$$S = \frac{TP}{N_{act}}$$

Specificity (true negative rate):

$$SP = \frac{TN}{N_{inact}}$$

Enrichment: $$EF = \frac{\dfrac{TP}{TP + FP}}{\dfrac{N_{act}}{N}}$$

TP: True positives

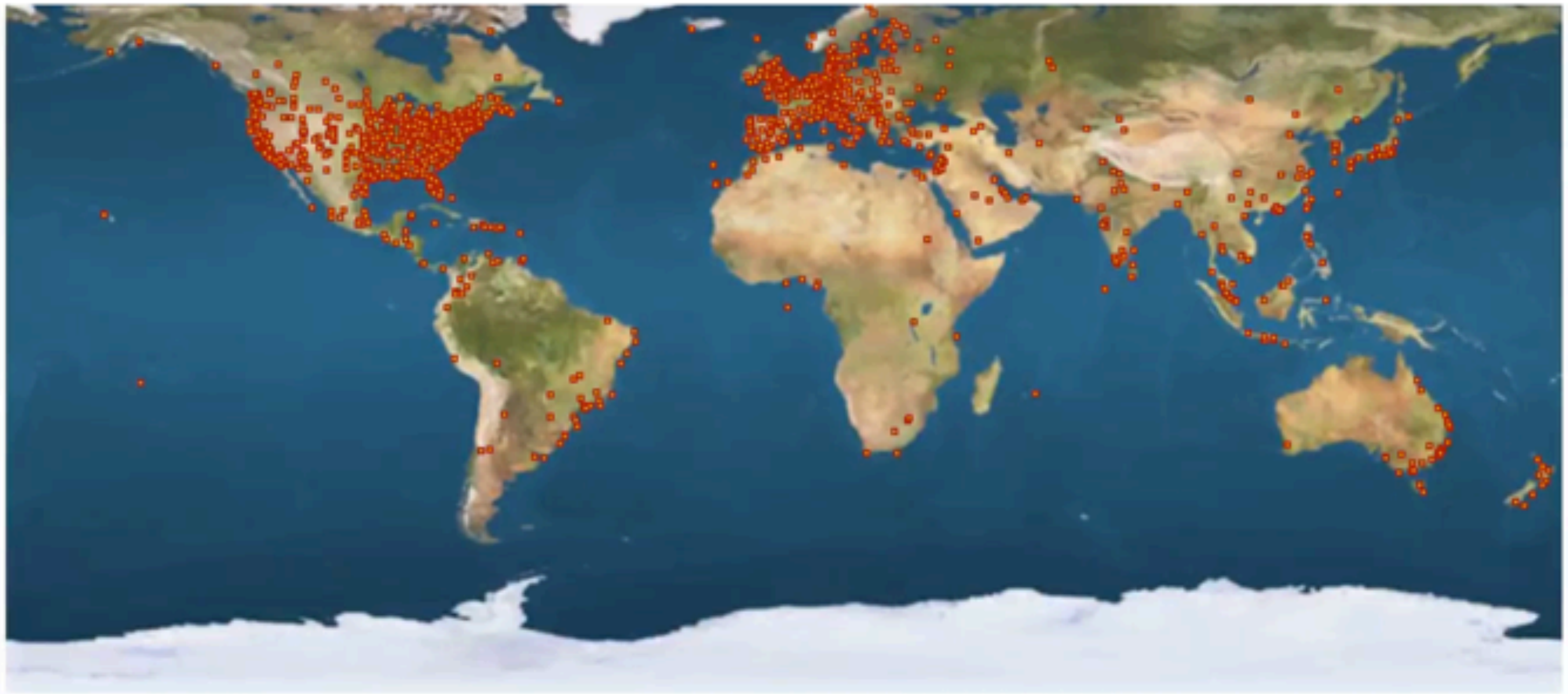
FP: False positives

N_{act}: Number of active samples

N_{inact}: Number of inactive samples

Report enrichment at 0.1 % 0.5 %, 1%, etc.

MobileApps: Worldwide Marketing Vehicles!



iKinase

iKinasePro

iOncology

iProtein

Mobile Reagents

Reaction101

Yield101

SPRESImobile

~ 30,000 People Use Eidogen Mobile Apps