Beyond the scope of Free-Wilson analysis: Building interpretable QSAR models with machine learning algorithms

Hongming Chen
Discovery Sciences, AstraZeneca R&D Gothenburg

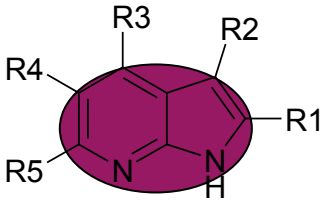
AstraZeneca

Free-Wilson analysis

- **FW method is one of the oldest QSAR method appeared in 1960's (*J. Med. Chem. 1964, 7, 395-9*)**
- **The basic idea in Free-Wilson approach is that the biological activity of a molecule can be described as the sum of the activity contributions of its R-groups**

Molecule ID	*x1*	*x2*	.....	*y1*	*y2*	.....
1	1	0	.....	1	0	.....
2	0	1	.....	1	0	.....
3	1	0	.....	0	1	.....
4	0	1	.....	0	1	.....
.....	.....	.....	.....	.....	.....	.....

R1 = x1, x2,....
R2 = y1, y2, ...



Scaffold and R-groups

$$Activity = C + a_1x_1 + a_2x_2 + a_3x_3 + \cdots + b_1y_1 + b_2y_2 + b_3y_3 + \cdots$$

$$= C + \sum_{i=1}^{n} a_i x_i + \sum_{j=1}^{m} b_j y_j$$

a_i and b_j are coefficients that represent the contribution made by each R-group to the activity of a compound;

Renaissance of Free-Wilson Method

- **Recently FW analysis has been reemerged as an useful QSAR method for lead optimization.**

- **The advantages of FW methods:**

  - **No R-group parameters needed**
  - **Interpretable model with clear contribution of R-groups.**

- **The disadvantage of FW methods:**

  - **Can't predict for compound whose R-group is outside the training set R-group list.**
  - **Contribution of R-group may not be additive**

J Comput Aided Mol Des (2012) 26:1143–1157
DOI 10.1007/s10822-012-9605-7

Composite multi-parameter ranking of real and virtual compounds for design of MC4R agonists: Renaissance of the Free-Wilson methodology

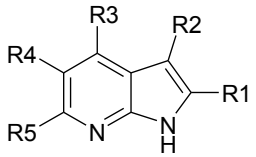
Ingemar Nilsson · Magnus O. Polla

Received: 22 May 2012 / Accepted: 7 September 2012 / Published online: 2 October 2012

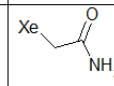
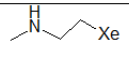
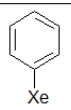
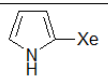
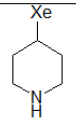
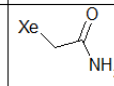
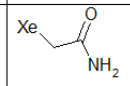
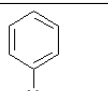
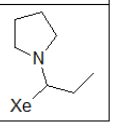
Abstract Drug design is a multi-parameter task present in the analysis of experimental data for synthesized compounds and in the prediction of new compounds with desired properties. This article describes the implementation of a binned scoring and composite ranking scheme for

Introduction

The pharmaceutical industry is currently facing a number of challenges to the ultimate goal of delivering clinically beneficial and profitable drugs to the market. To be suc-

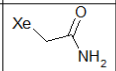
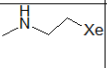
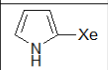
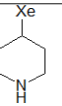
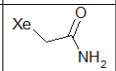
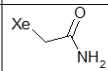
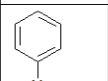
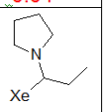


R-group structures as descriptor

	R1	R2	R3	R4	R5
Cpd1					
Cpd2					

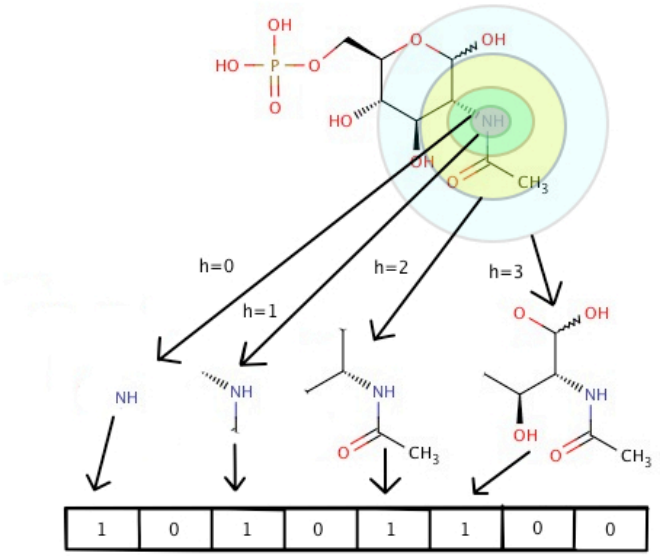
Linear regression

Building QSAR model comprising the contribution from R-groups

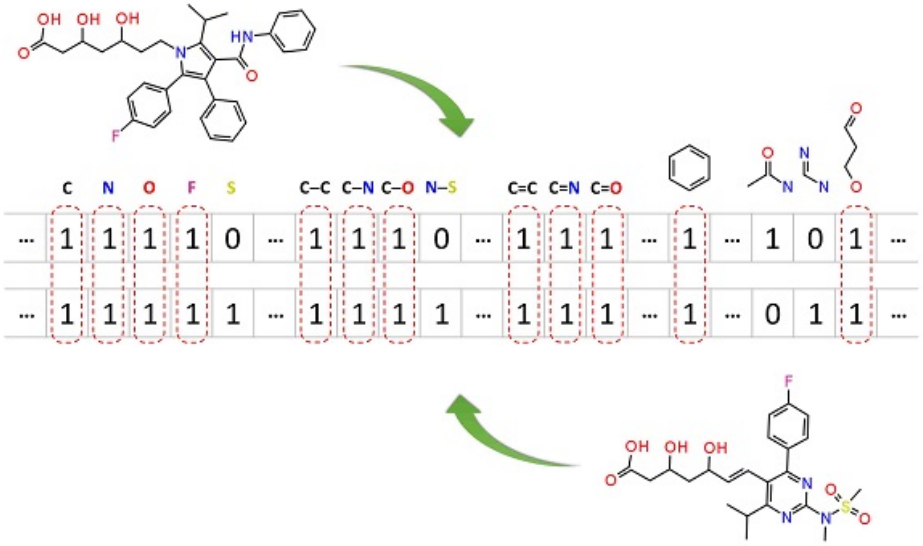
	R1 Contribution	R2 Contribution	R3 Contribution	R4 Contribution	R5 Contribution
Cpd1	 0.2	 -0.01	 0.02	 0.2	 -0.04
Cpd2	 0.01	 0.3	 -0.05	 0.05	 0.3

Molecular Fingerprint descriptors

- As an analogue to biometric fingerprint, molecular fingerprint is a set of values which represent the characteristics of a compound.
- Molecular fingerprint is normally expressed by an array of bits.



Circular fingerprint



Fingerprint based on fragments

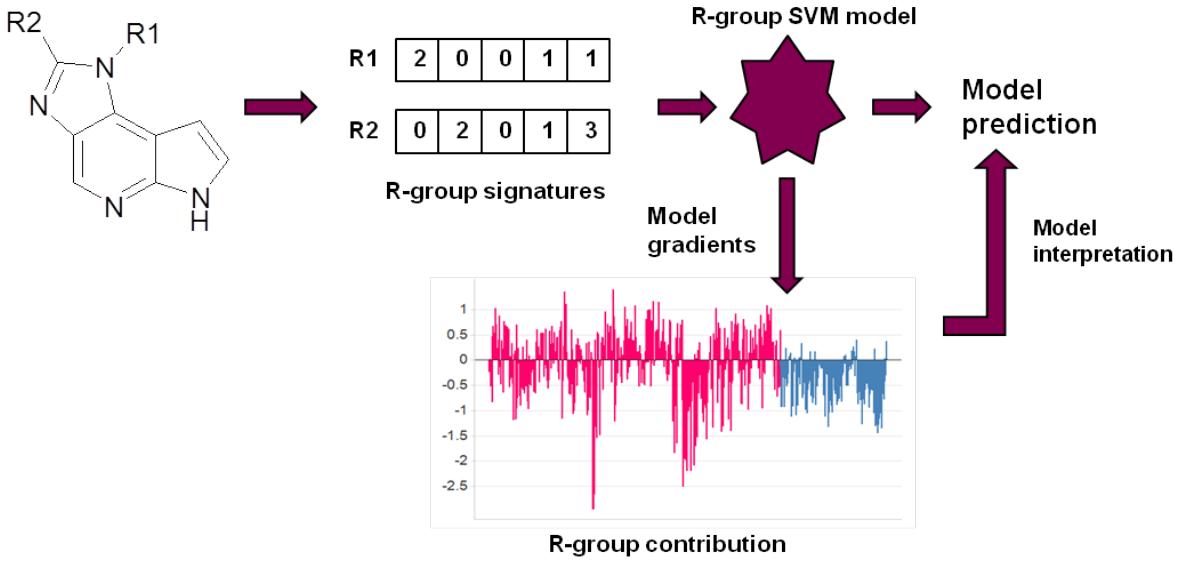
Beyond the scope of FW method

- **A new method is proposed to overcome the drawbacks of FW method.**

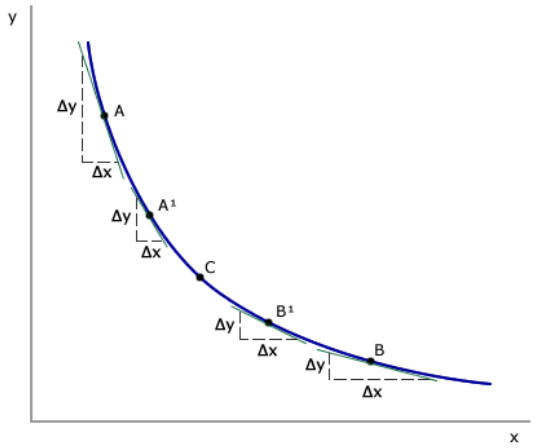
- **Using R-group signatures (a circular fingerprint by nature) to replace the R-group structures as descriptor. (Faulon, J. L., *et al J. Chem. Inf. Comput. Sci.* 2003, *43*, 707–20.)**

- **Using SVM (LIBSVM) as modeling method instead of linear regression**

- **Deriving signature gradient for SVM model to measure the relative contribution of R-groups**



R-group contribution in SVM model

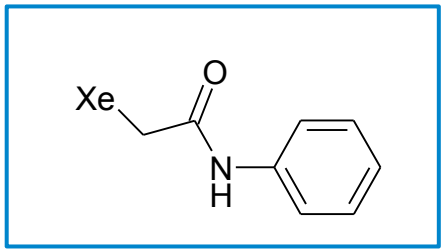


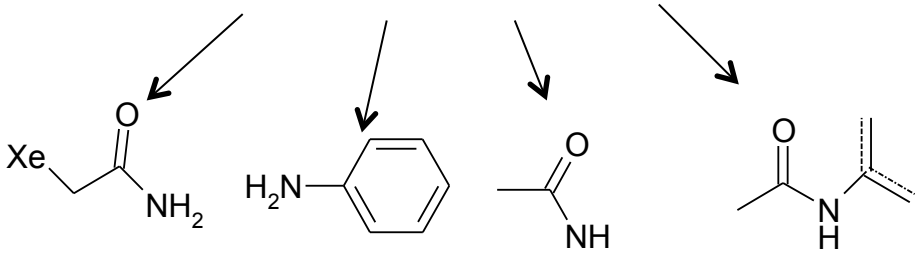
$$\frac{Df}{Dx_j} = \frac{f(x + h_j) - f(x)}{h_j}$$

Model gradient for signature

Carlsson, L. et al. J. Compt. Info. Model. 2009, 49, 2551-8.

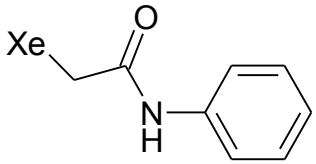
- **SVM gradient represents how large a variable impact the QSAR equation**
- **Model discrete gradient for individual signature can be calculated.**
- **The model gradient for each individual R-group can also be calculated.**





$\frac{Df}{Dx_j}$ Gradient for R-group signatures

$$C_n = \sum_{j=1}^{k} t_j \times \frac{Df}{Dx_j}$$ Gradient for whole R-group



R-group contribution

Benefits of the R-group Signature SVM method

- **Overcome the limited prediction scope problem for FW method due to the fuzziness of signature.**

- **Signatures (fingerprints) can capture subtle chemical functionalities in R-group and therefore might improve prediction accuracy.**

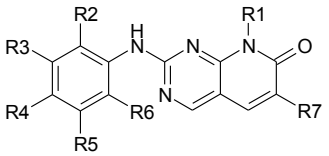
- **SVM method can efficiently handle the non-linear QSAR relationship.**

- **QSAR model is as interpretable as FW method via calculating the R-group contribution based model gradient for signatures.**

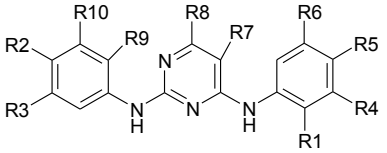
Test cases for the methodology

- **Eleven focused dataset were chosen from GVKBio chemical patents**

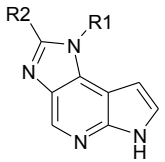
Dataset	Compound description	Data type	Nr. Training set	Nr. Test set	Data Source
CDK5	CDK5 inhibitor	IC50	184	46	US 20040224958 A1
IL4	IL-4 inhibitor	IC50	532	133	WO 2006/133426 A2
JAK1	JAK1 inhibitor	Ki	736	185	WO 2011/086053 A1
MAPK14	P38 alpha inhibitor	IC50	488	122	EP 1500657 A1
PIK3CA	PI3K alpha inhibitor	IC50	243	61	WO 2010/139731 A1
TYK2	TYK2 inhibitor	Ki	736	184	WO 2011/086053 A1
F7	Factor VIIa inhibitor	IC50	292	73	US20050043313
GNRHR	Gonadotropin-releasing hormone receptor antagonist	IC50	159	39	WO20020358
MGLL	Monoacylglycerol Lipase inhibitor	IC50	982	246	WO2010124082
MMP2	Matrix metalloprotease 2 inhibitor	IC50	439	110	WO2005042521
PRSS2	Trypsin II inhibitor	IC50	271	68	US7119094



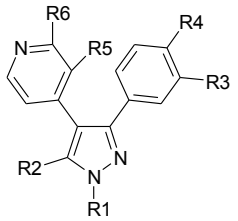
CDK5



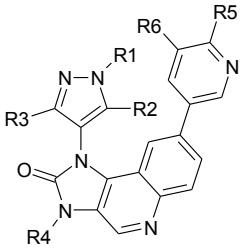
IL4



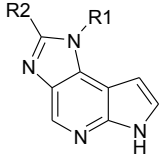
JAK1



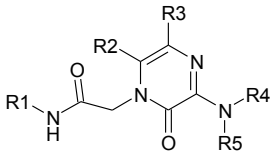
MAPK14



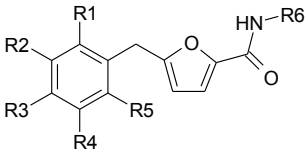
PIK3CA



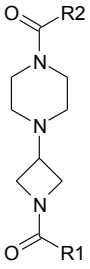
TYK2



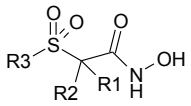
F7



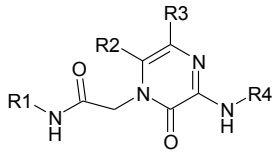
GNRHR



MGLL

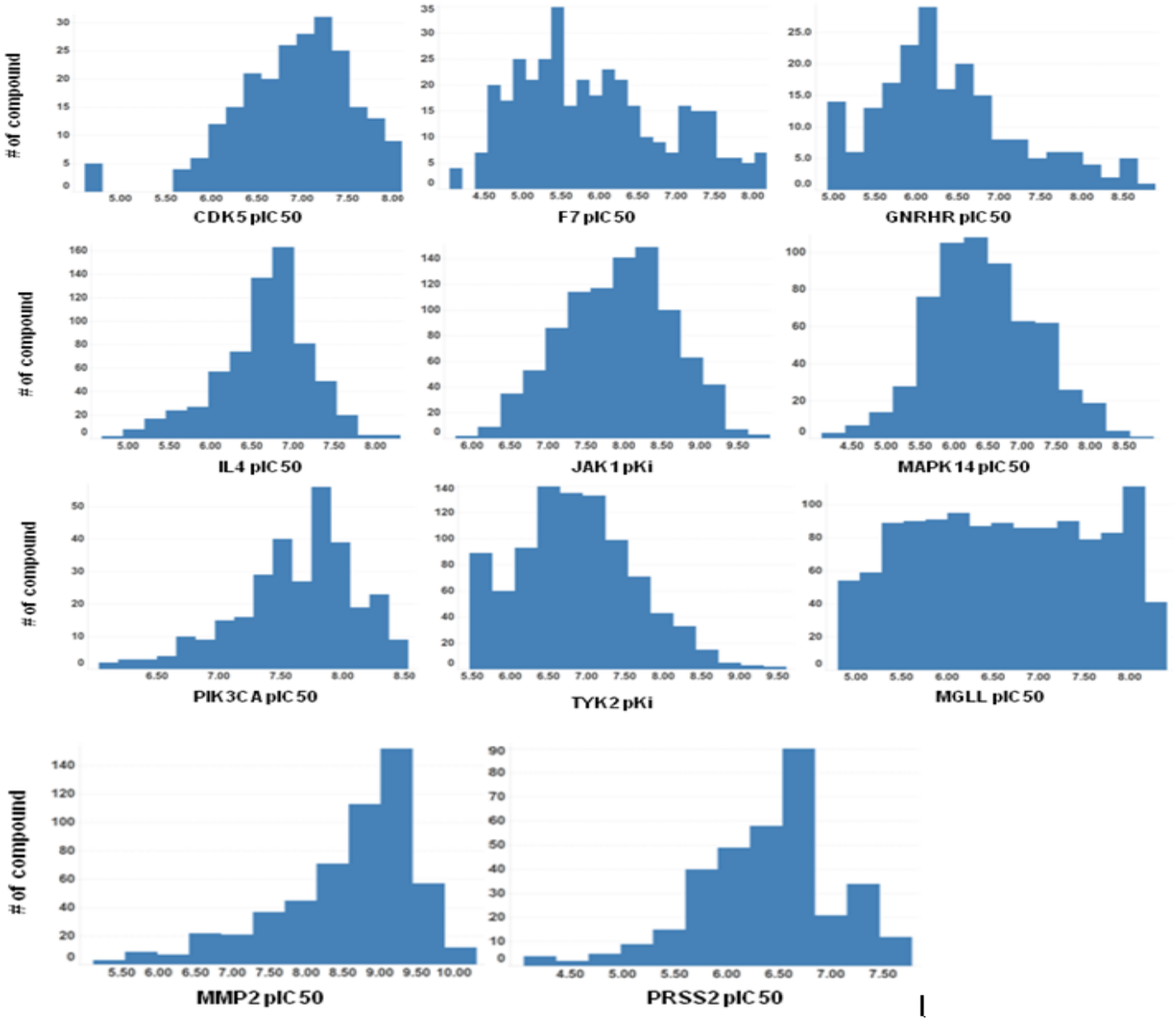


MMP2



PRSS2

Distribution of bioactivity data



Building QSAR models

- **Various model building strategies were used:**

  - **Mol. Signature + SVM (Mol_Sign_SVM)**
  - **Mol. ECFP_6 + SVM (Mol_ECFP6_SVM)**
  - **Rgroup Signature + SVM (Rgp_Sign_SVM)**
  - **Azdescriptor + SVM (Azdesc_AZOSVM)**
  - **Free-Wilson method (FW)**

- **γ and C value in LIBSVM were optimised through grid search**

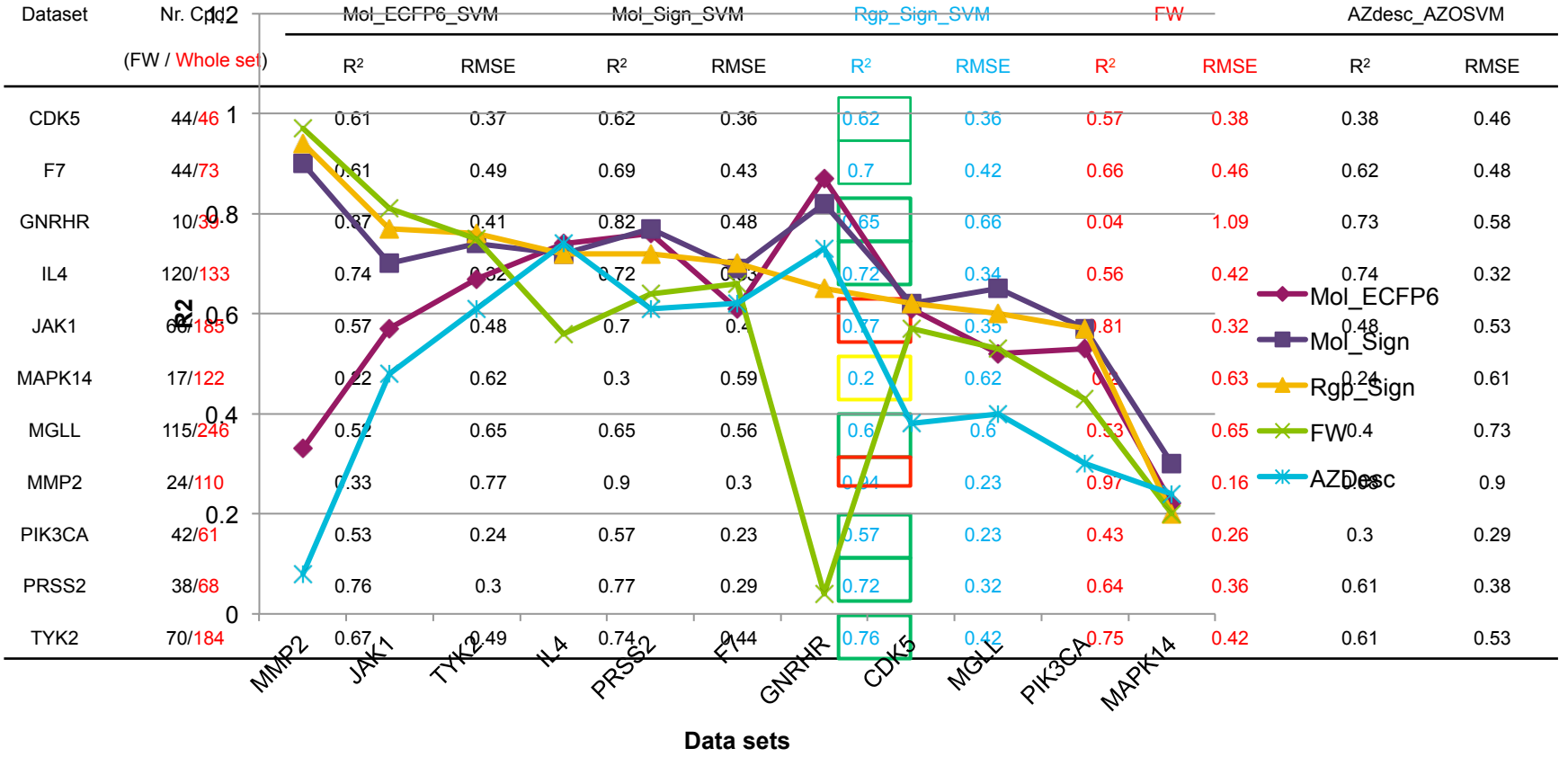
- **Each dataset was split into training/test set with a ratio of 4:1 for validation**

- **10-fold cross validation to check the model robustness.**

Model performance

Comparison of performance on FW test set

Dataset	Nr. Cpd (FW / Whole set)	Mol_ECFP6_SVM		Mol_Sign_SVM		Rgp_Sign_SVM		FW		AZdesc_AZOSVM	
		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
CDK5	44/46	0.61	0.37	0.62	0.36	0.62	0.36	0.57	0.38	0.38	0.46
F7	44/73	0.61	0.49	0.69	0.43	0.7	0.42	0.66	0.46	0.62	0.48
GNRHR	10/39	0.37	0.41	0.82	0.48	0.65	0.66	0.04	1.09	0.73	0.58
IL4	120/133	0.74	0.32	0.72	0.3	0.72	0.34	0.56	0.42	0.74	0.32
JAK1	66/185	0.57	0.48	0.7	0.4	0.77	0.35	0.81	0.32	0.48	0.53
MAPK14	17/122	0.22	0.62	0.3	0.59	0.2	0.62	0.4	0.63	0.24	0.61
MGLL	115/246	0.52	0.65	0.65	0.56	0.6	0.6	0.53	0.65	0.4	0.73
MMP2	24/110	0.33	0.77	0.9	0.3	0.24	0.23	0.97	0.16	0.08	0.9
PIK3CA	42/61	0.53	0.24	0.57	0.23	0.57	0.23	0.43	0.26	0.3	0.29
PRSS2	38/68	0.76	0.3	0.77	0.29	0.72	0.32	0.64	0.36	0.61	0.38
TYK2	70/184	0.67	0.49	0.74	0.44	0.76	0.42	0.75	0.42	0.61	0.53

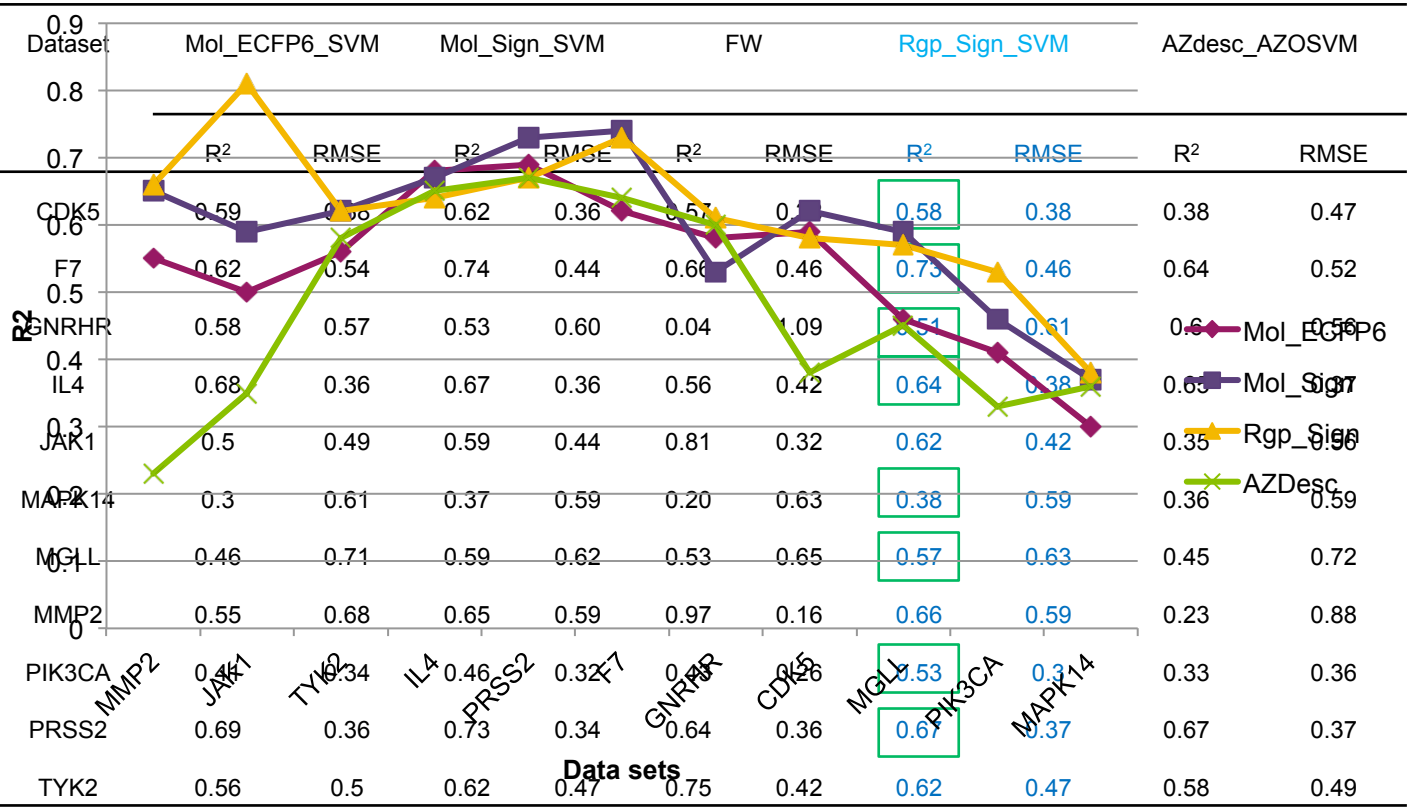


- **FW can only predict part of the test set due to its limitation on predicting domain.**
- **R-group signature models perform better than FW models in most of the cases.**

Model performance

Comparison of performance on full test set (FW model only predict on part of the test set.)

Dataset	Mol_ECFP6_SVM		Mol_Sign_SVM		FW		Rgp_Sign_SVM		AZdesc_AZOSVM	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
CDK5	0.59	0.[illegible]	0.62	0.36	0.[illegible]	0.[illegible]	0.58	0.38	0.38	0.47
F7	0.62	0.54	0.74	0.44	0.66	0.46	0.73	0.46	0.64	0.52
GNRHR	0.58	0.57	0.53	0.60	0.04	1.09	0.[illegible]	0.61	0.6[illegible]	0.56
IL4	0.68	0.36	0.67	0.36	0.56	0.42	0.64	0.38	0.63	0.37
JAK1	0.5	0.49	0.59	0.44	0.81	0.32	0.62	0.42	0.35	0.56
MAPK14	0.3	0.61	0.37	0.59	0.20	0.63	0.38	0.59	0.36	0.59
MGLL	0.46	0.71	0.59	0.62	0.53	0.65	0.57	0.63	0.45	0.72
MMP2	0.55	0.68	0.65	0.59	0.97	0.16	0.66	0.59	0.23	0.88
PIK3CA	0.45	0.34	0.46	0.32	0.[illegible]	0.[illegible]	0.53	0.3[illegible]	0.33	0.36
PRSS2	0.69	0.36	0.73	0.34	0.64	0.36	0.67	0.37	0.67	0.37
TYK2	0.56	0.5	0.62	0.47	0.75	0.42	0.62	0.47	0.58	0.49



- **Full molecular signature/ECFP6 model have comparable performance with R-group signature model. Azdesc model perform worst.**

Model interpretation

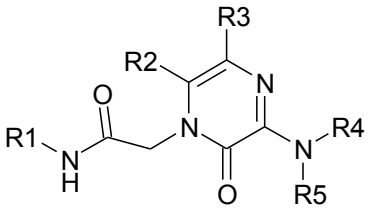
- **Machine learning model normally were regarded as “black-box” model due to lack of interpretability.**
- **Signature gradients and R-group gradients for SVM model were calculated for model interpretation.**
- **The validity of interpretability for R-group gradients was examined by correlating with R-group contribution coefficient in the FW model.**
- **Significant correlation was observed in most of the cases. This promising result suggests that the R-group SVM model could be as interpretable as the FW model.**

Correlation of SVM R-group gradients with R-group contr. Of FW model

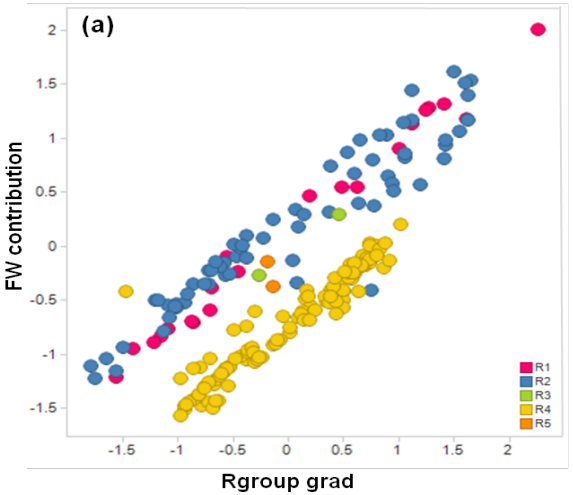
Dataset	Nr. Compound	Nr. R-groups	R^2	RMSE
F7	292	204	0.62	0.47
JAK1	736	573	0.50	0.47
TYK2	736	576	0.71	0.32
CDK5	184	53	0.33	0.29
GNRHR	159	157	0.005	0.44
IL4	532	241 (11[a])	0.42	0.38
MAPK14	488	476	0.0007	0.94
MGLL	982	681 (3[a])	0.35	0.79
MMP2	439	473	0.42	0.66
PIK3CA	243	122	0.52	0.32
PRSS2	271	166	0.64	0.37

Note: a) Outliers which were excluded in the regression analysis

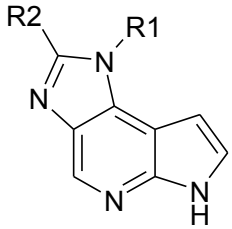
Model interpretation



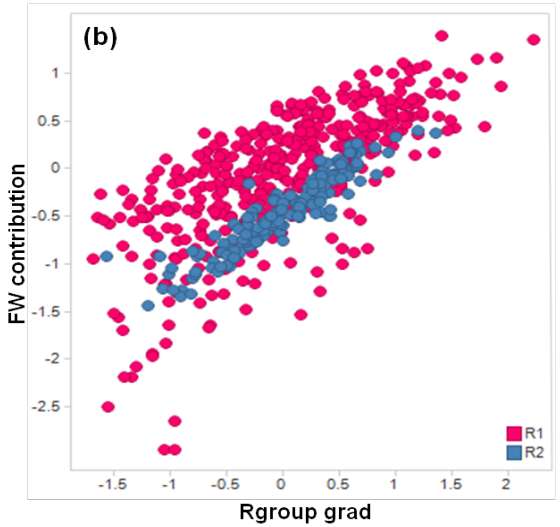
F7 data set



R1: R^2=0.99, n=21
R2: R^2=0.89, n=67
R4: R^2=0.86, n=112
All R-groups: R^2=0.62, n=204



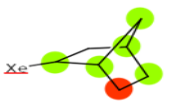
JAK1 data set



R1: R^2=0.49, n=419
R2: R^2=0.86, n=154
All R-groups: R^2=0.5, n=573

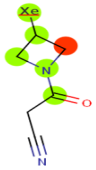
- **Splitting the R-groups into sub-groups according to substituent positions can further improve the correlation.**

Signature examples

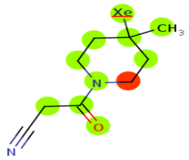


690 (R1), 0.14

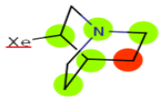
622 (R1), 0.13



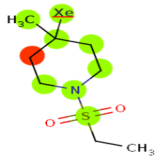
404 (R1), -0.13



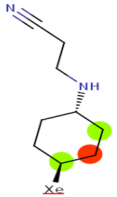
625 (R1), -0.11



894 (R1), -0.11



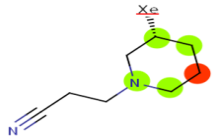
850 (R1), 0.11



1061 (R1), 0.11

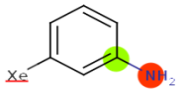
900 (R1), 0.10

3109 (R2), -0.10



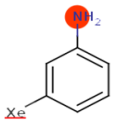
708 (R1), -0.10

JAK1 set

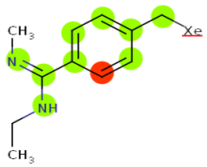


966 (R2), 0.19

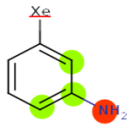
761 (R2),-0.17



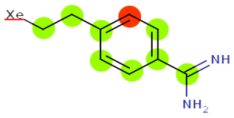
947 (R2), 0.15



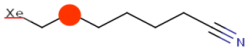
169 (R1), -0.13



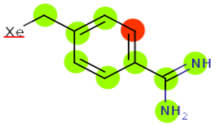
962 (R2), 0.13



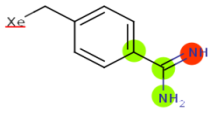
163 (R1), -0.13



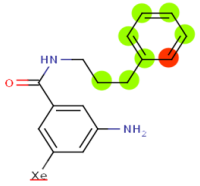
1112 (R4), -0.12



170 (R1), 0.11



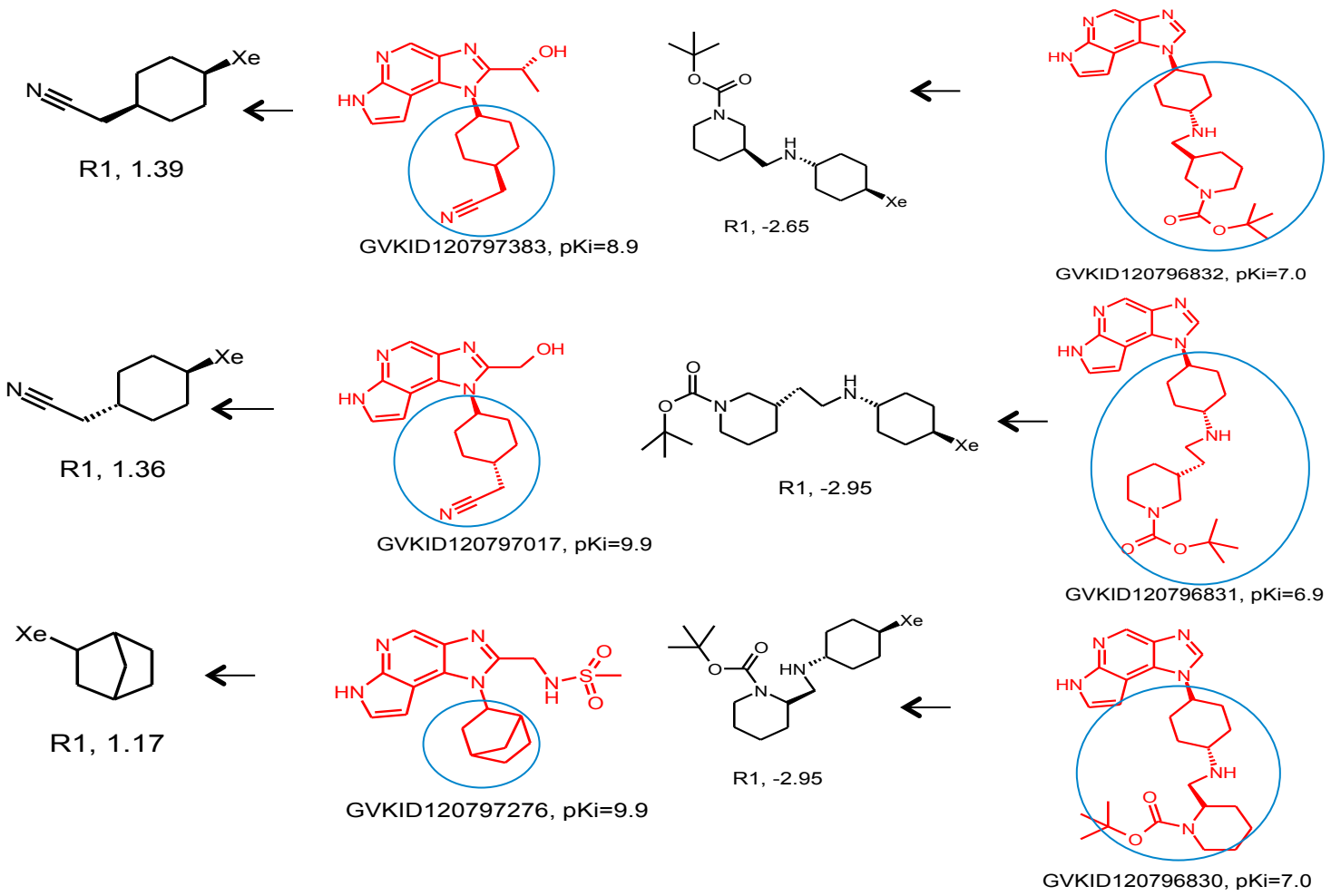
320 (R1), 0.11



623 (R2), -0.11

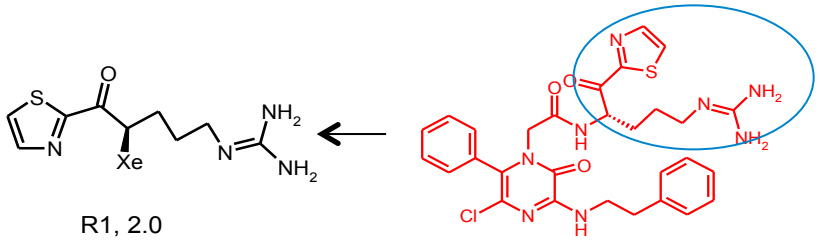
F7 set

JAK1 R-group examples

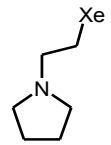


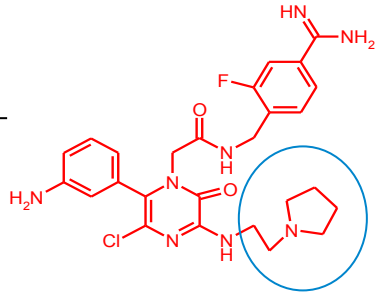
Top three most influential R-groups

F7 R-group examples

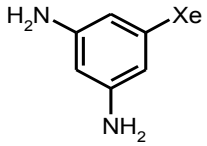


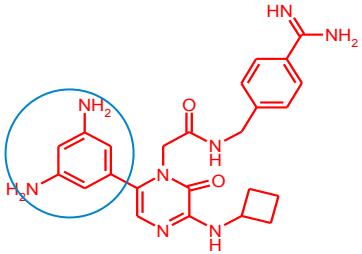
GVKID3690119, pIC50=6.7



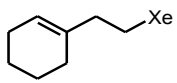


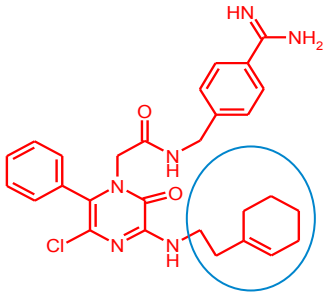
GVKID3690559, pIC50=5.9



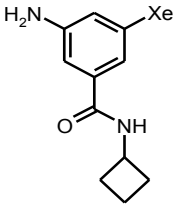


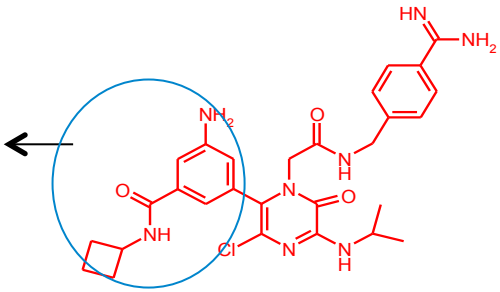
GVKID3690222, pIC50=8.0



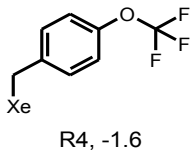


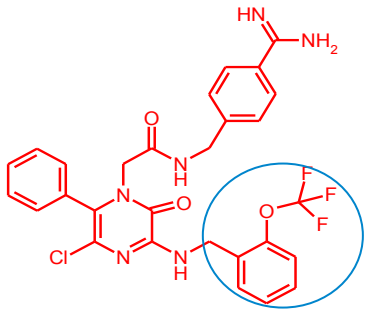
GVKID3690356, pIC50=4.6





GVKID3690243, pIC50=8.2





GVKID3690407, pIC50=4.6

AstraZeneca “SARPlatform”

- **This new method has been implemented into the proprietary web based tool (“SARPlatform”) for chemist to build and visualize R-group QSAR models**

Overview of all R-group gradients

Individual R-group gradient info.

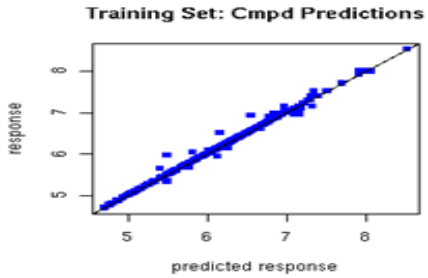
Info. about cpds containing the R-group

SARPlatform: Fragment-based QSAR analysis

Stored Model Results: PACT

Training set cmpds measured vs predicted:

R^2 = 1.00, ste = 0.04



Spotfire Visualizations:

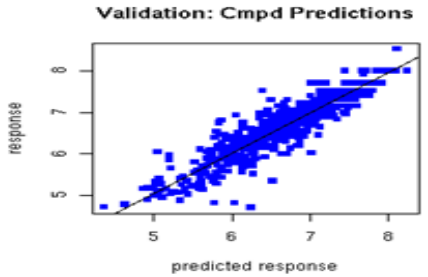
Open in Web Player (opens in a new tab)
Open in Spotfire client (Only for Windows users.
Enter "427" when prompted for a parameter)

Download Model Data:

Get Prediction Data
Get R-group Gradient Data
Get R-group Decomposition Data

External/Cross Validation Statistics:

R^2: 0.85 , Std Error: 0.27
Number of predictions: 616



Parameters:

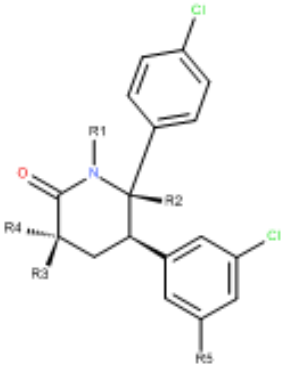
SVM Signature Gradient Visualizations:

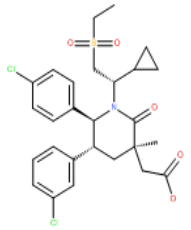
Training set R-groups:

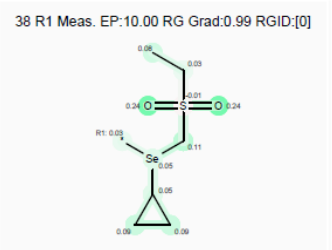
Download PDF (sorted by R1 gradient contributions)
Download PDF (sorted by R2 gradient contributions)
Download PDF (sorted by R3 gradient contributions)

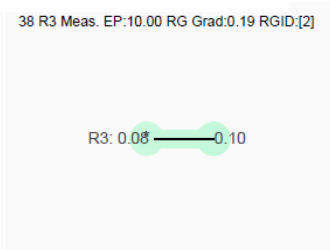
AstraZeneca “SARPlatform”

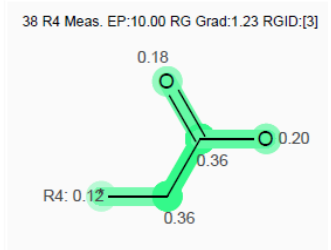
- **Signature gradient can be mapped and visualized at atom level to help to understand the SAR**

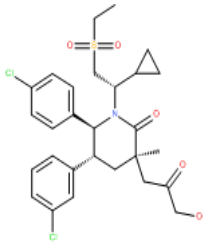


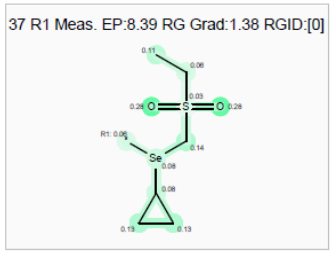


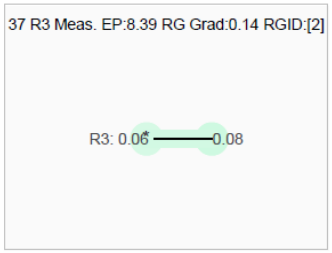


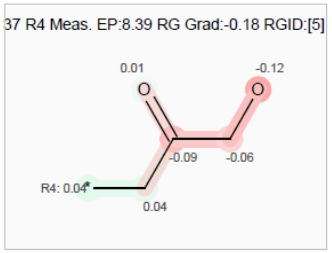


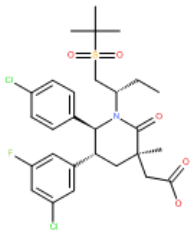


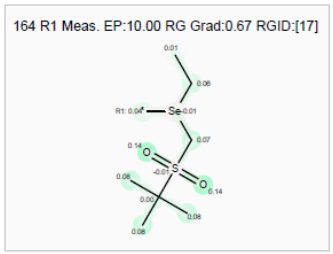


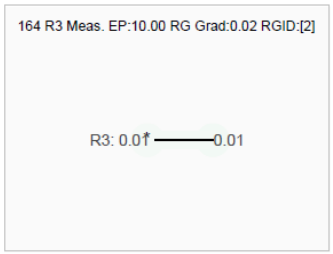


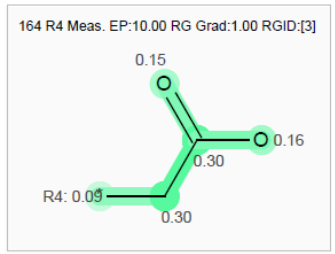


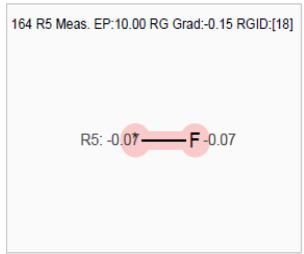












Future development

- **Including information of distance to attachment point into the signatures to reduce noise.**
- **Making inverse QSAR to design new compounds based on SVM model.**
- **Develop confidence metrics for model prediction**

Conclusions

- ***A novel methodology was developed to build Free-Wilson like local QSAR model by combining R-group signatures and the SVM algorithm.***
- ***The signature/R-group gradient concept was introduced to interpret SVM model (applicable to machine learning model in general).***
- ***The signature models overcome the FW's prediction domain problem and also have better accuracy than typical FW models.***
- ***Significant correlation between R-group gradient and FW R-group contribution highlight that the signature SVM model is as interpretable as FW model***

Acknowledgement

- Ola Engkvist
- John Cumming
- Willem Nissink

Backup Slides

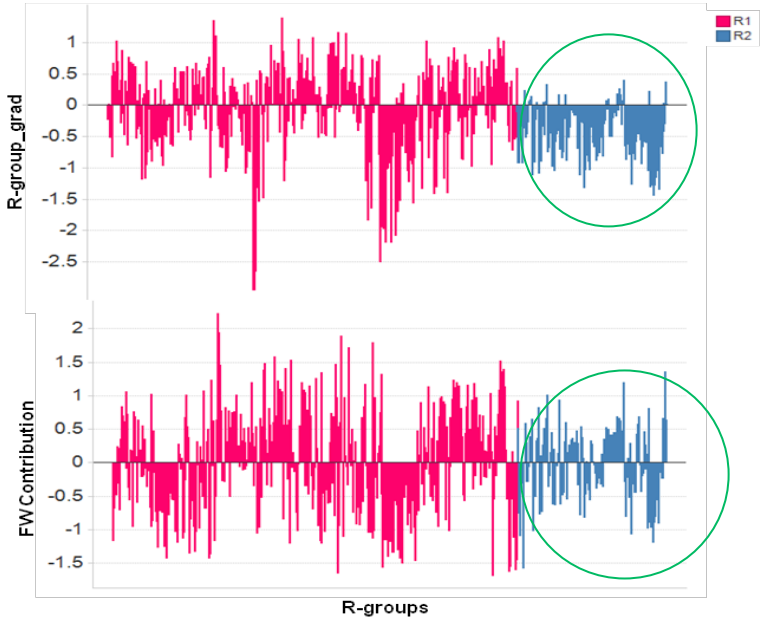
Model performance

Performance on 10-fold cross validation

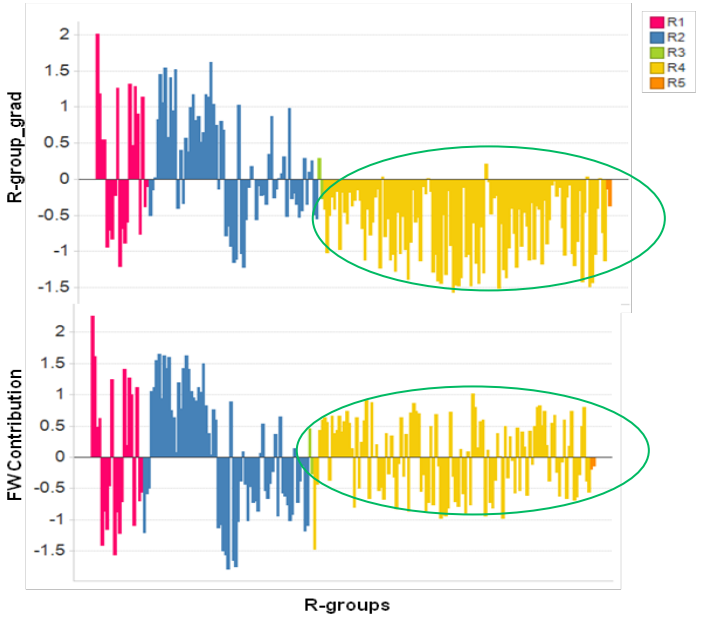
Dataset	Mol_ECFP6_SVM		Mol_Sign_SVM		Rgp_Sign_SVM		AZdesc_AZOSVM	
	Q^2	RMSE	Q^2	RMSE	Q^2	RMSE	Q^2	RMSE
CDK5	0.54	0.45	0.62	0.41	0.58	0.44	0.58	0.36
F7	0.73	0.50	0.75	0.48	0.74	0.50	0.70	0.53
GNRHR	0.49	0.62	0.47	0.64	0.44	0.65	0.46	0.64
IL4	0.64	0.34	0.63	0.34	0.63	0.34	0.58	0.36
JAK1	0.58	0.47	0.58	0.47	0.63	0.43	0.50	0.51
MAPK14	0.34	0.64	0.31	0.66	0.37	0.62	0.28	0.66
MGLL	0.51	0.69	0.60	0.62	0.54	0.66	0.49	0.70
MMP2	0.61	0.59	0.67	0.55	0.70	0.53	0.55	0.64
PIK3CA	0.37	0.38	0.40	0.38	0.47	0.35	0.27	0.41
PRSS2	0.60	0.42	0.65	0.39	0.60	0.42	0.61	0.42
TYK2	0.60	0.51	0.64	0.47	0.68	0.44	0.61	0.49

- **10-fold cross validation results are similar to the full test set results. No bias introduced in the test set.**

Model interpretation



JAK1 data set



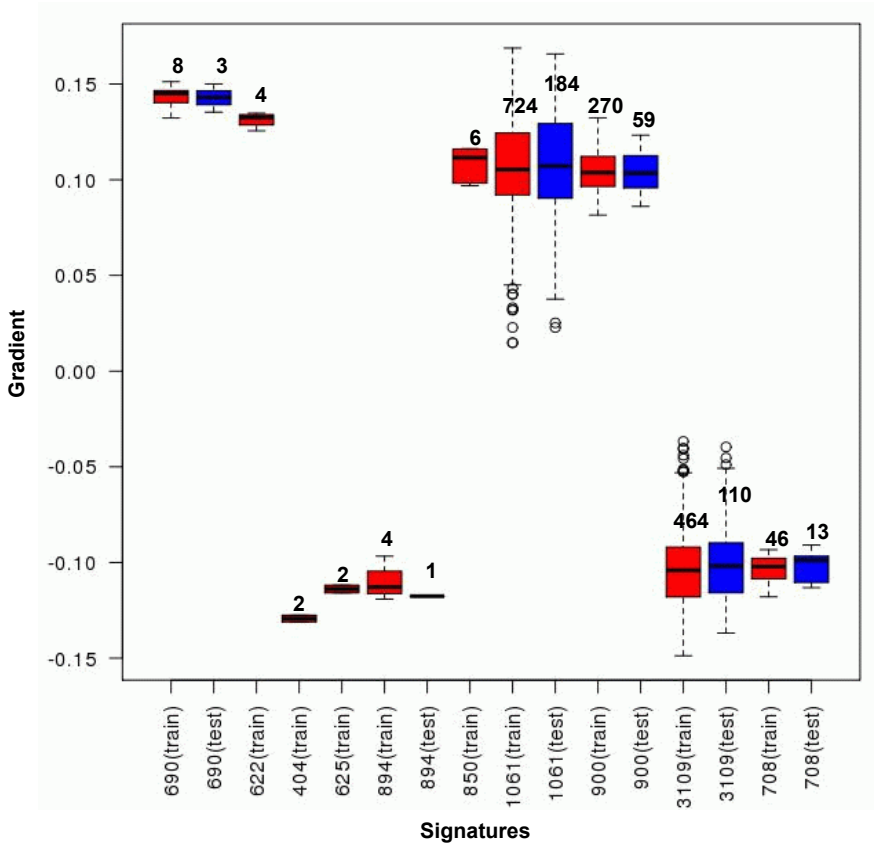
F7 data set

- **Distribution pattern for R2 in JAK1 set, R4 in F7 set are significantly different, while R-group contribution for other positions aligned well.**

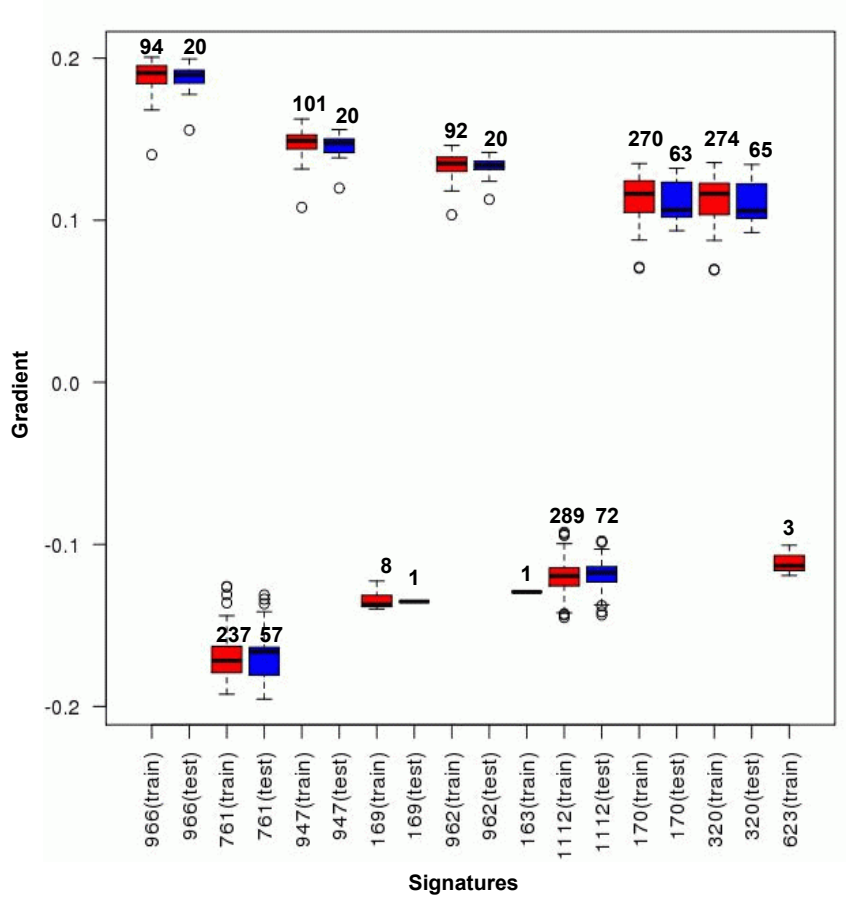
- **This results may imply that SVM R-group contribution value does not reflect the absolute contribution to bioactivity, but a relative ranking of R-group's contribution instead**

- **It is probably better to compare R-groups which belong to the same substituent position**

Most influential signatures in JAK1 and F7 set



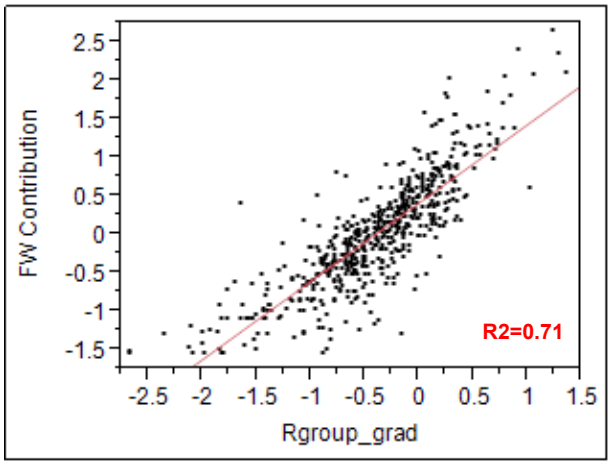
JAK1 data set



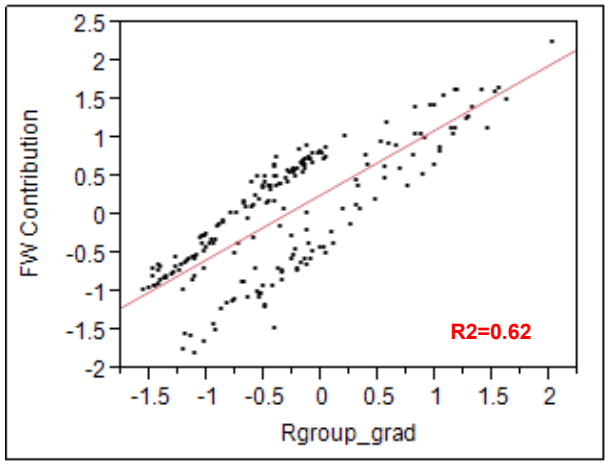
F7 data set

- The gradients for some highly influential signatures have pretty low deviation.
- Some small signatures (mostly having only 2 atoms in signature) have larger variation.

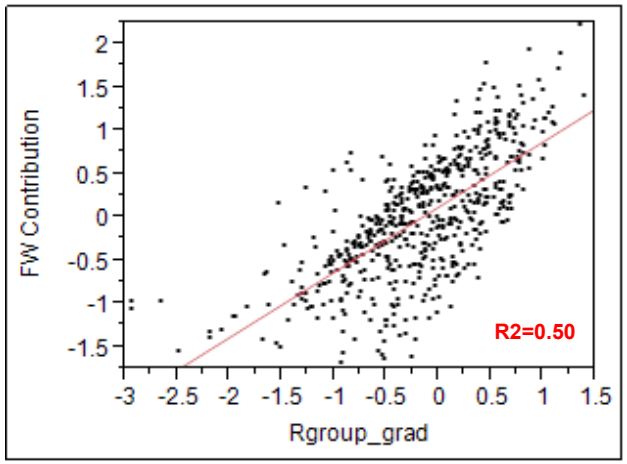
Their contribution may depend on their surrounding environment.



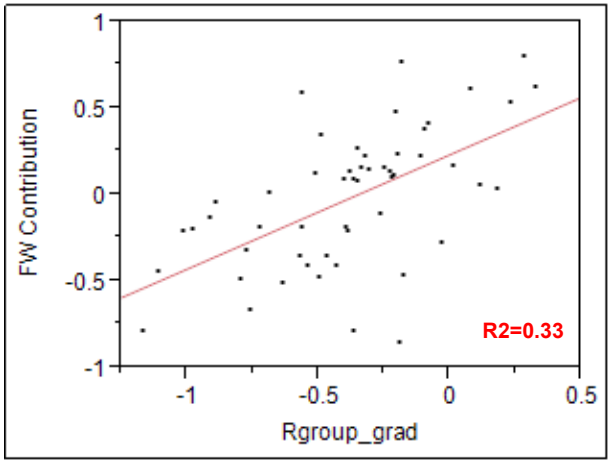
TYK2 dataset (n=576)



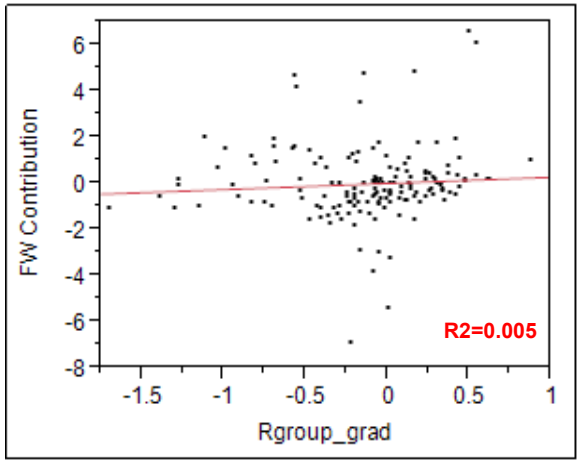
F7 dataset (n=204)



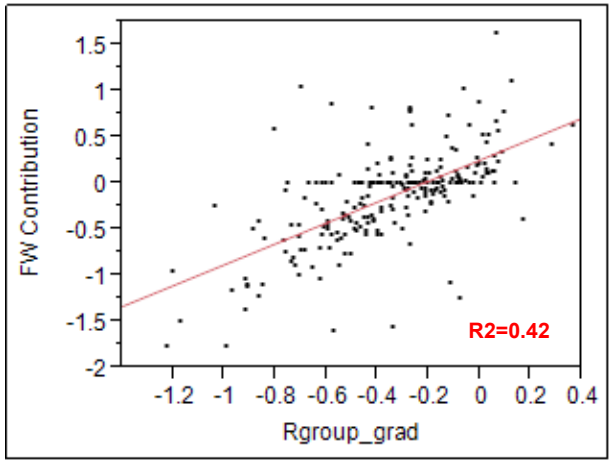
JAK1 dataset (n=573)



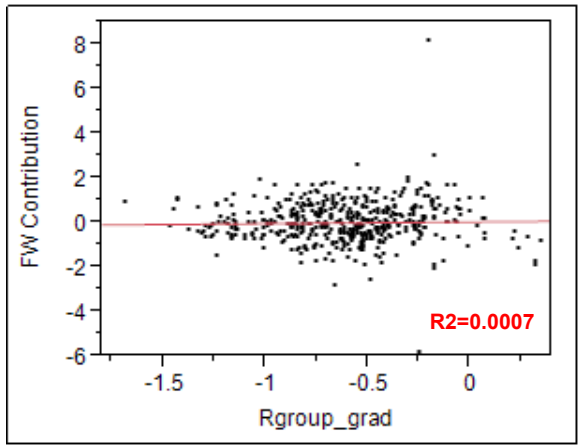
CDK5 dataset (n=53)



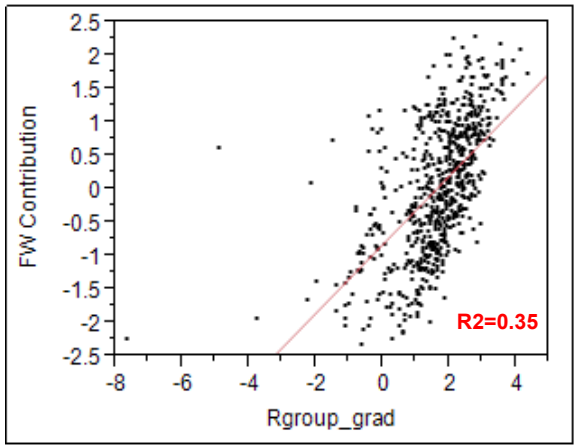
GNRHR dataset (n=157)



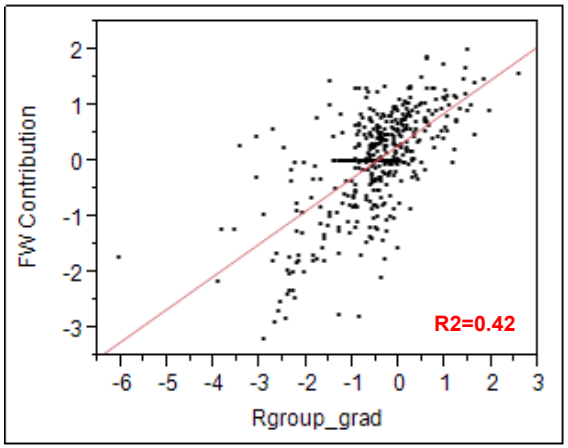
IL4 dataset (n=241, outliers=11)



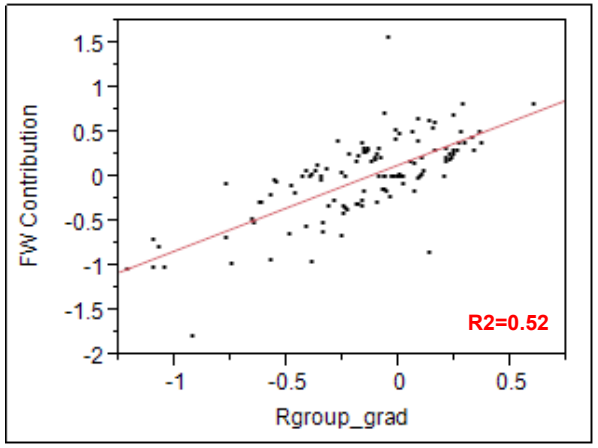
MAPK14 dataset (n=476)



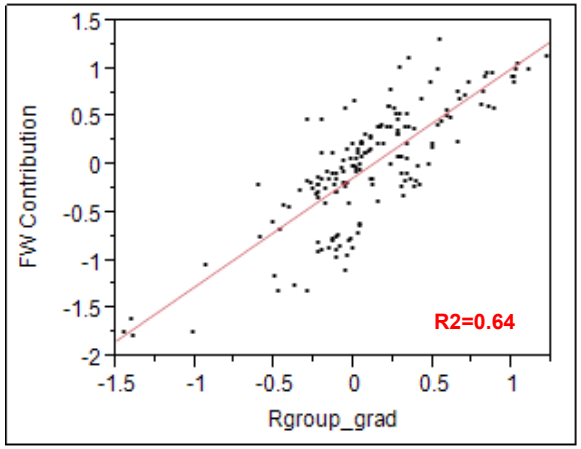
MGLL dataset (n=681, outliers=3)



MMP2 dataset (n=473)



PIK3CA dataset (n=122)



PRSS2 dataset (n=166)