APPLICATION OF MACHINE LEARNING TO STRUCTURAL INTERACTION FINGERPRINTS – INSIGHT INTO ACTIVITY AND SELECTIVITY OF LIGANDS

JAGNA WITEK, SABINA SMUSZ, KRZYSZTOF RATAJ, STEFAN MORDALSKI, ANDRZEJ J. BOJARSKI

DEPARTMENT OF MEDICINAL CHEMISTRY, INSTITUTE OF PHARMACOLOGY POLISH ACADEMY OF SCIENCES, 12 SMĘTNA STREET, 31-343 KRAKÓW, POLAND
E-MAIL: JAGNA.WITEK@GMAIL.COM

INTRODUCTION

A number of cheminformatic methods, such as Virtual Screening, constitute a vital part of modern drug design process. Those techniques enable not only viable prediction of physicochemical properties of the molecules, but also effective database mining, being particularly useful tool in the search for ligands of desired activity. Successful performance in case of single target experiments, implies consequent need to extend its capabilities to finding compounds bearing desired activity towards multiple targets.

In this research, we present application of Machine Learning (ML) algorithms to Structural Interaction Fingerprints (SIFts) as a basis for multitarget approach to Virtual Screening. Protein kinases were chosen as targets for method validation; three crystal structures of each kinase were retrieved from PDB repository. Collection of active (K<1000 nM) and inactive (K>1000 nM) compounds towards each target were acquired from ChEMBL database. Furthermore, a set of decoy compounds assumed as inactive, was obtained from ZINC database (**Table 1**). Afterwards, the compounds were docked to respective crystal structures, and SIFts were calculated for each successfully docked protein-ligand pair (**Fig. 2**).

FINGERPRINT PREPARATION

SIFts enable recognition of aminoacids involved in ligand binding and additionally, types of interactions between specific residues. In this research nine bits were used to describe those associations: any contact, backbone, side chain, polar, hydrophobic, hydrogen bond donor/acceptor, aromatic and charged (**Fig. 1**).

SIFts generated for each ligand docked into at least one of three kinase structures, were subsequently utilized to create SIFt profile. It was performed by averaging three fingerprint strings into single profile, describing ligand's interaction pattern in simplified manner (**Fig. 4**). In order to validate SIFt performance, method was compared with standard hashed fingerprints (Extended Fingerprints) generated by means of PaDEL Descriptor.

SIFT ANALYSIS

A crucial stage of interaction examination, was application of machine learning algorithms to SIFt profiles. Two distinct approaches to test and training sets selection were adopted to perform viable analysis:
Approach I – profiles for active and inactive ligands were seperately clustered, centroids were extracted to create training sets, remaining profiles were collected to create test set
Approach II – cross validation – profiles for active and inactive compounds were mixed, and divided into a number of groups. During each stage one group was selected as a training set, whilst test set consisted of all remaining profiles

Machine learning itself was performed by means of Sequential Minimal Optimization (SMO), Naive Bayes and Random Forest algorithms. Their performance was evaluated by recall (fraction of positives selected from test set), precision (correctness of positive instances prediction; low values indicate a high rate of false positives), and MCC (gives balanced measure of ML methods performance (**Fig. 5**)).

RESULTS AND CONCLUSIONS

Application of machine learning to SIFt analysis enabled discrimination of ligand's preference to target protein, independently of the chosen algorithm (**Fig. 6**). Nevertheless, there is a significant discrepancy in MCC values between ABL/CDK and GSK3b/LCK/SRC, being a cause of inequality in number of active and inactive compounds (**Fig. 6a**). To reduce this effect, analysis was performed additionally on set of decoys, assumed as inactive, retrieved from ZINC database. Such approach distinctly enhanced analysis reliability (**Fig. 6b**).
What is more, independent study was performed on Extended Fingerprints generated for all compounds used in SIFt preparation. Application of such approach resulted in significant recall's decrease, consequently affecting overall performance. The outcomes clearly showed prevalence of SIFt profiles application in estimating compound's activity.

Presented method may be useful in assessment of ligand's affinity towards target receptor structure, in case of paucity of experimental data. However, the most beneficial way to exploit this procedure would be determination of multitarget profile of ligand's interaction. Further evaluation may allow to investigate its capabilities and limitations.

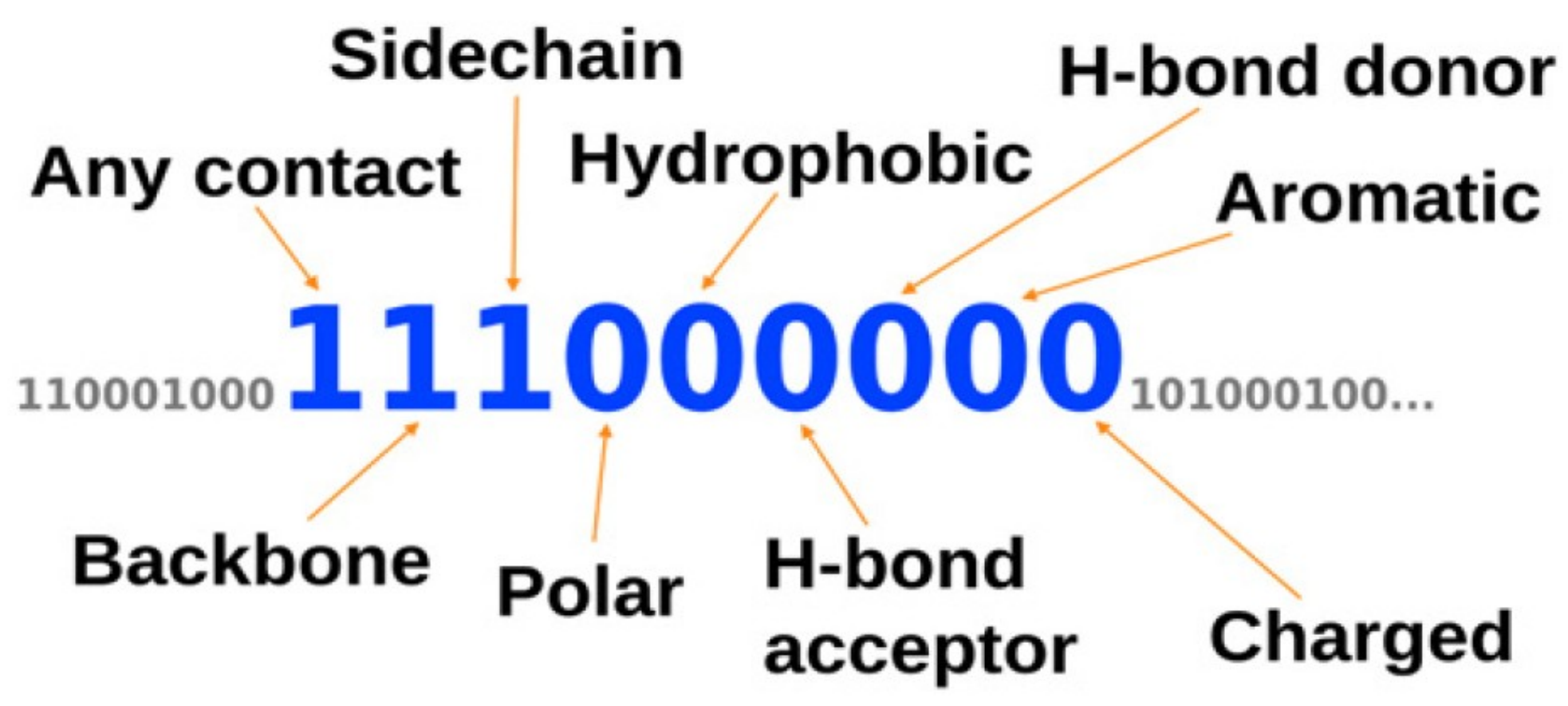


Figure 1. Fragment of SIFt describing bit positions for individual ligand-residue interactions.

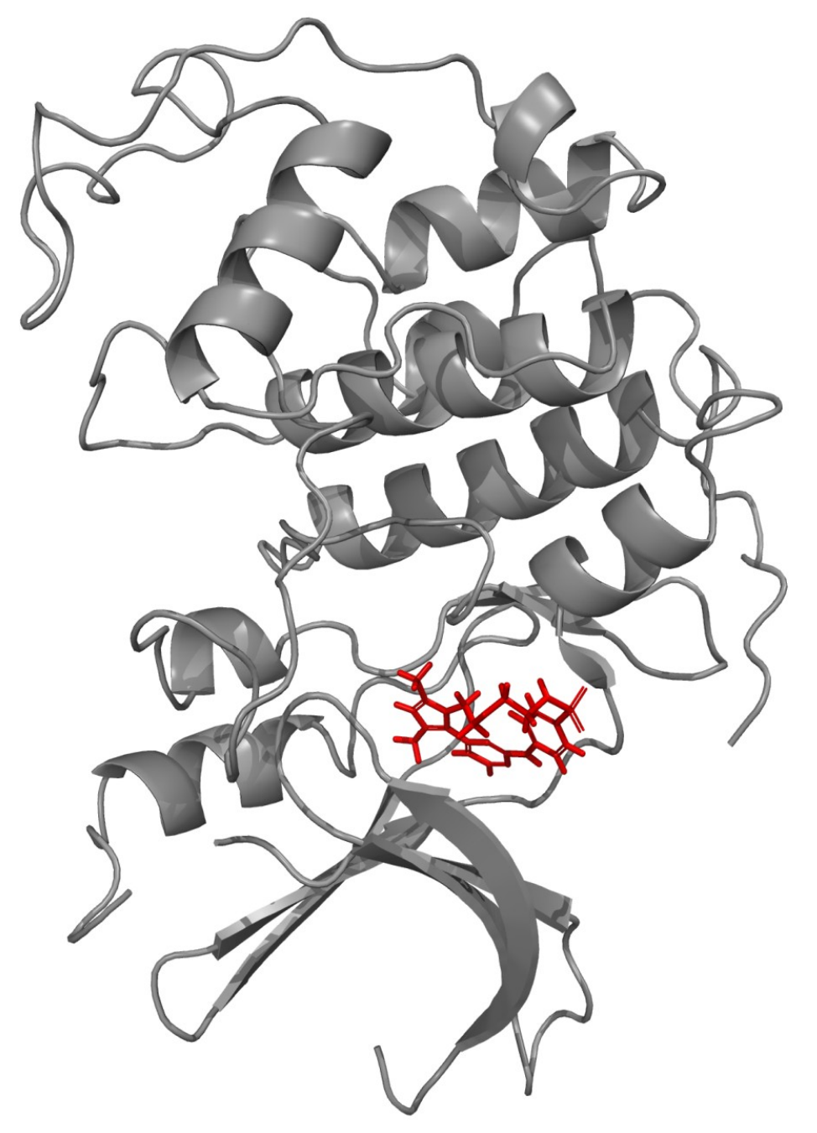


Figure 2. Crystal structure of cyclin-dependent kinase 2 with active ligand docked.

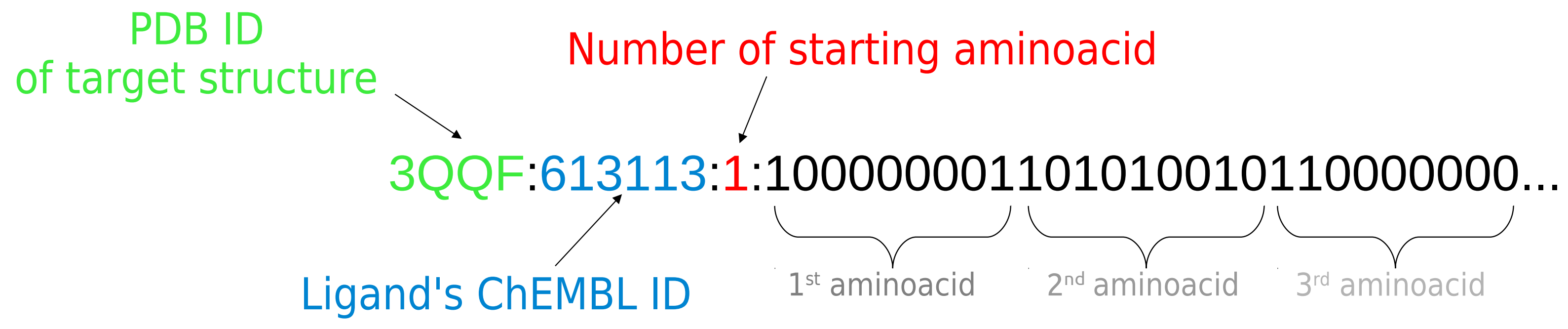
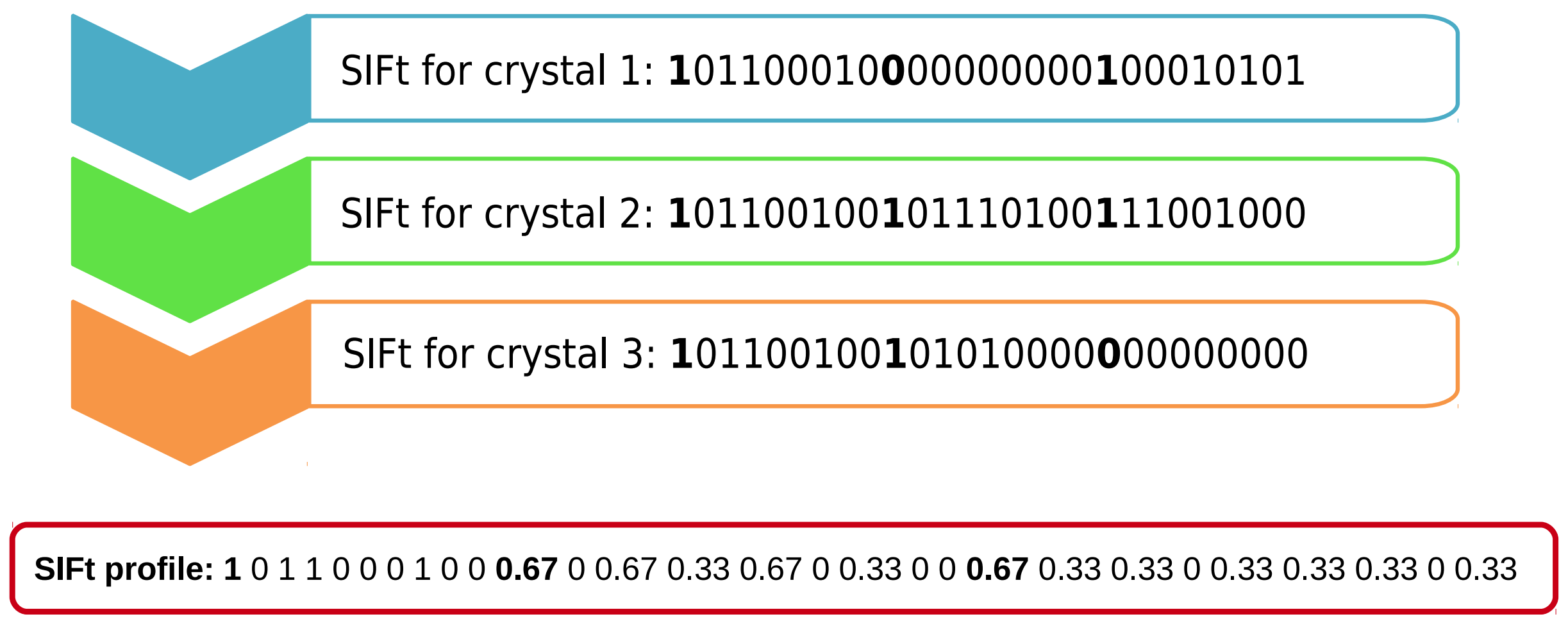


Figure 3. Scheme of fingerprint string.

Table 1. Averaged number of compounds docked to crystal structures.

Target	Actives		Inactives		Decoys	
	Input	Docked	Input	Docked	Input	Docked
ABL	1117	1110	20	16	2662	2607
CDK	3567	3396	233	214	2662	2640
GSK3b	2010	1985	65	64	2662	2658
LCK	2199	2000	75	74	2662	2617
SRC	3196	3081	77	73	2662	2655



SIFt for crystal 1: 101100010**0**00000000**1**00010101

SIFt for crystal 2: 101100100**1**01110100**1**11001000

SIFt for crystal 3: 101100100**1**01010000**0**00000000

SIFt profile: 1 0 1 1 0 0 0 1 0 0 **0.67** 0 0.67 0.33 0.67 0 0.33 0 0 **0.67** 0.33 0.33 0 0.33 0.33 0.33 0 0.33

Figure 4. Scheme of SIFt profile construction.

$$Recall=\frac{TP}{TP+FN}$$

$$Precision=\frac{TP}{TP+FP}$$

$$MCC=\frac{TP\cdot FN-FP\cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

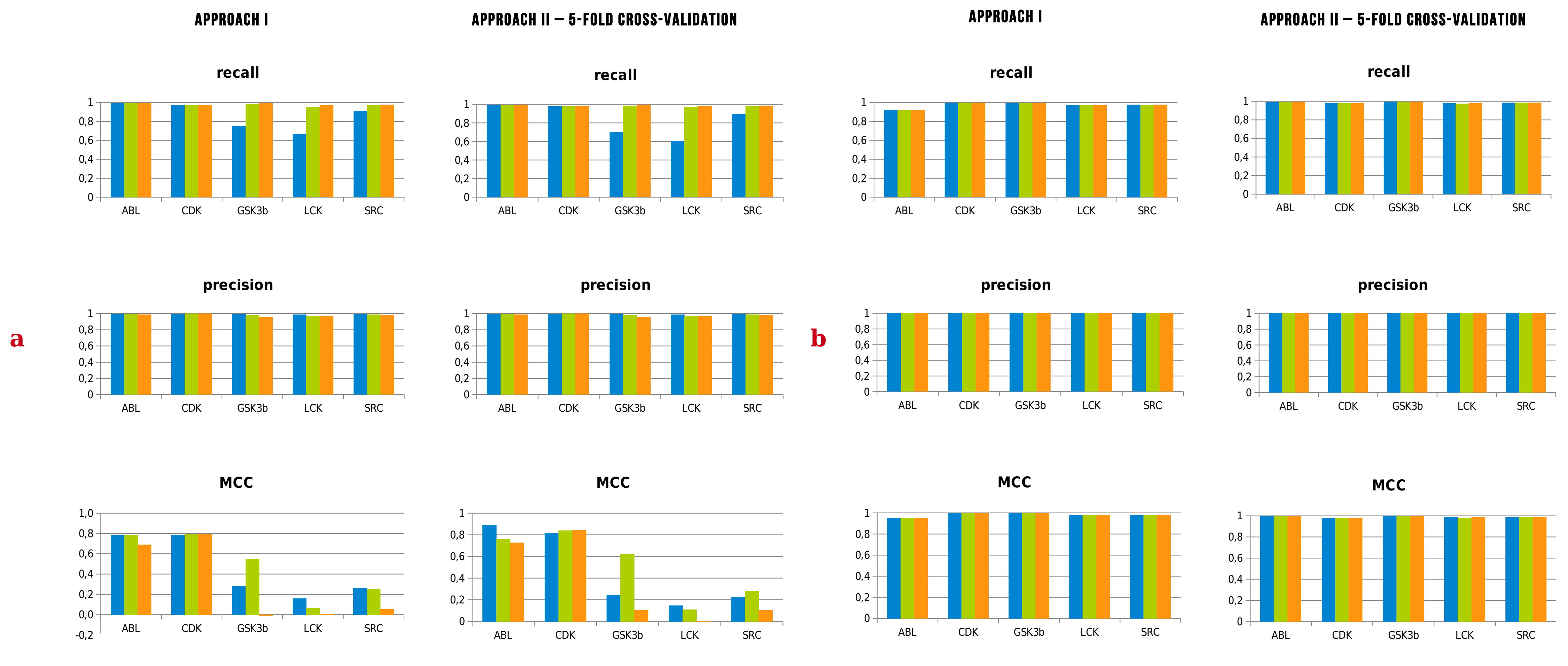
TP – number of true positives
(correctly classified actives)

FP – number of false positives
(inactives wrongly classified as actives)

TN – number of true negatives
(correctly classified inactives)

FN – number of false negatives
(actives wrongly classified as inactives)

Figure 5. Measures of machine learning performance.



🟦 - Naive Bayes
🟩 - SMO
🟧 - Random Forest

ABL - Tyrosine protein-kinase ABL
CDK – Cyclin dependent kinase 2
GSK3b – Glycogen synthase kinase 3-beta
LCK - Tyrosine protein-kinase LCK
SRC - Tyrosine protein-kinase SRC

Figure 6. Evalutation of machine learning methods performance in predicting compounds activity.
(a) Actives vs inactives **(b)** Actives vs decoys.

Literature

(1) Deng, Z.; Chuaqui, C.; Singh, J.; *Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions,* J Med Chem 2004, 47, 337 -344
(2) Mordalski, S.; Kosciolek, T.; Kristiansen, K.; Sylte, I.; Bojarski, A. J.; *Protein binding site analysis by means of Structural Interaction Fingerprint patterns* Bioorg Med Chem Lett, 2011, 21, 6816-6819
(3) Liew, C.Y.; Ma, X.H.; Yap, C.W.; *Consensus model for identification of novel PI3K inhibitors in large chemical library* Comput Aided Des 2010, 4, 131-141

Acknowledgments

This study is supported by project UDA-POIG.01.03.01-12-100/08-00 co-financed by European Union from the European Fund of Regional Development (EFRD); http://www.prokog.pl

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN REGIONAL
DEVELOPMENT FUND