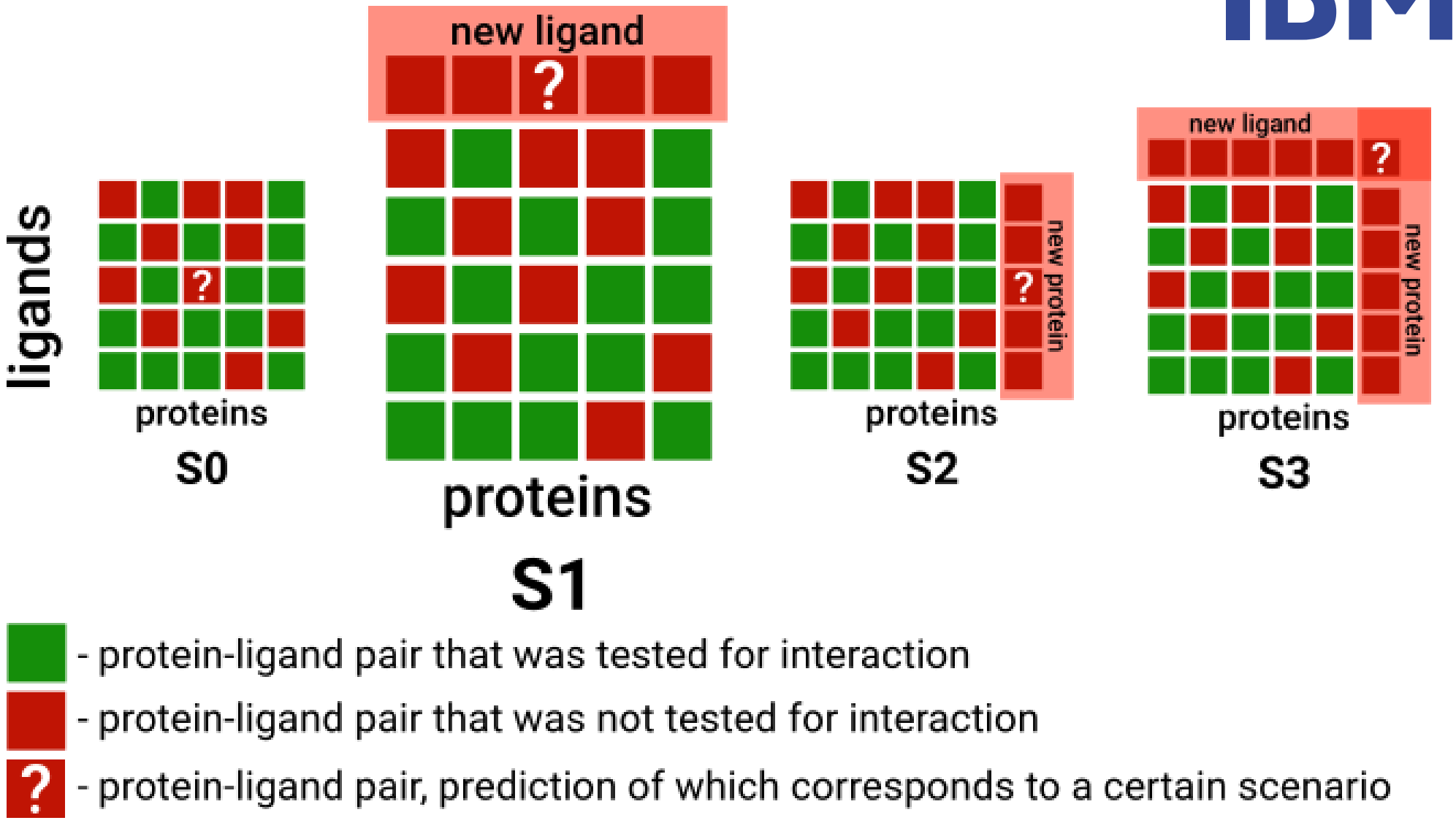
Comparative Efficacy of Structure Activity Relationship and Proteochemometric Modelling

Georgii Malakhov
Dmitry Karasev
Boris Sobolev

Institute of Biomedical Chemistry, Moscow, Russia

Scenarios of virtual screening



What's the difference?

	SAR (Structure-activity relationship)	PCM (proteochemometrics)
Purpose	To predict if a protein interacts with a ligand	
Object of prediction	A ligand	A protein-ligand pair
Number of models	One for each protein	One for a whole dataset
Feature vector	Feature vector for a ligand	Feature vector for a ligand + Feature vector for a protein
Applicability domain	S1 only	All four scenarios

Relevance of the Study

Researchers often use PCM in S1 instead of SAR without compelling evidence of advantages of PCM

- Sorgenfrei FA, Fulle S, Merget B. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem*. 2018;13(6):495-499. doi:10.1002/cmdc.201700180
- Cortés-Ciriano I, Bender A, Malliavin T. Prediction of PARP Inhibition with Proteochemometric Modelling and Conformal Prediction. *Mol Inform*. 2015;34(6-7):357-366. doi:10.1002/minf.201400165
- Paricharak S, Cortés-Ciriano I, IJzerman AP, Malliavin TE, Bender A. Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules. *J Cheminformatics*. 2015;7(1):15. doi:10.1186/s13321-015-0063-9

Aims

- To verify the advantage of PCM over SAR in the scenario S1
- To develop a suitable validation strategy and prove its fairness

Materials and Methods

Data

- Datasets on four most represented families of drug target with their ligands and interaction values, retrieved from Papyrus database. Interactions with K_i less than 6.5 log-units were considered “active”.

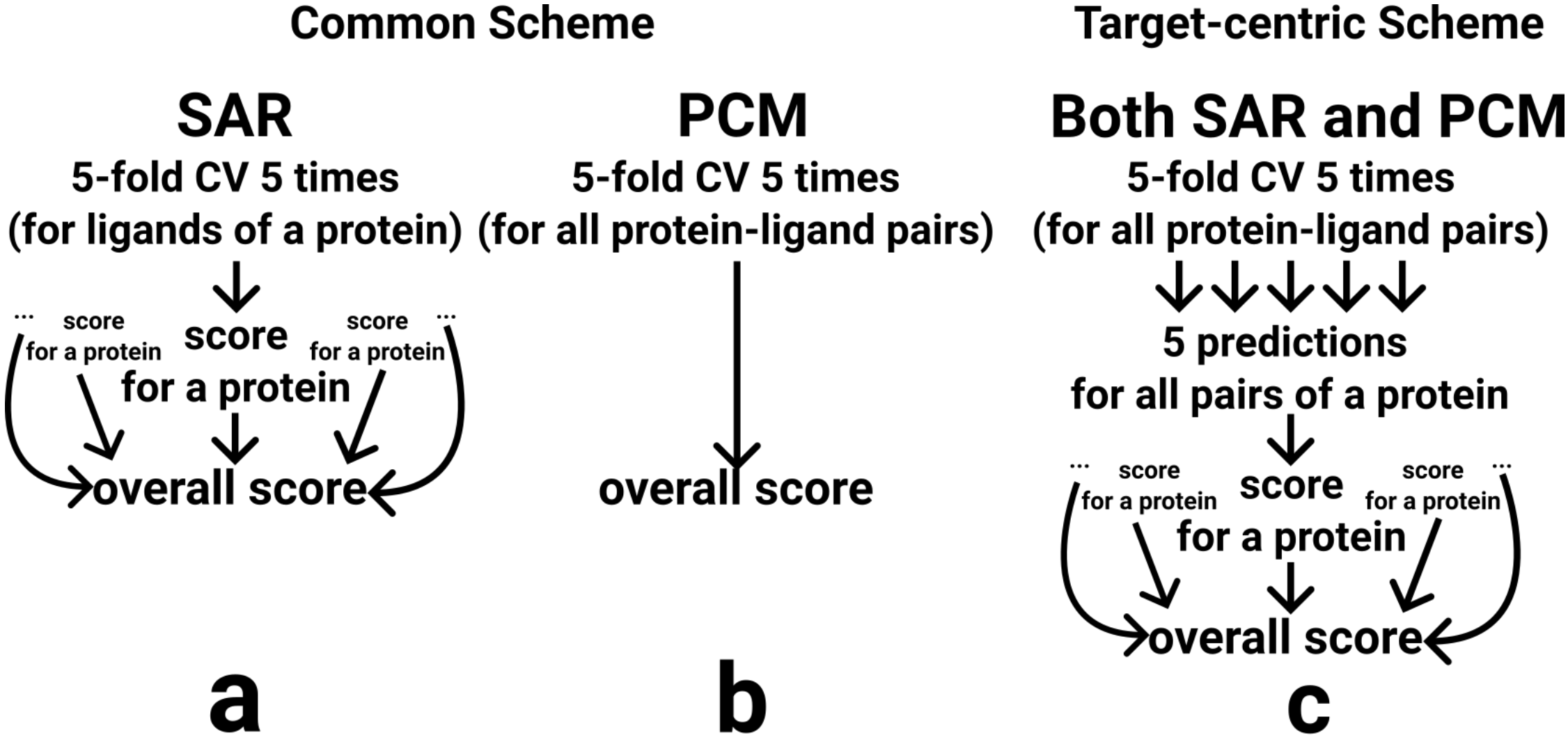
Descriptors

- ECFP6 (Extended Connectivity Fingerprints of radius 3) as feature vectors of ligands (structure-based)
- UniRep (Unified Representations) as feature vectors of proteins (sequence-based)

Model and Implementation

- Scikit-learn Python package for machine learning implementation
- Random Forest as a model

Validation strategy

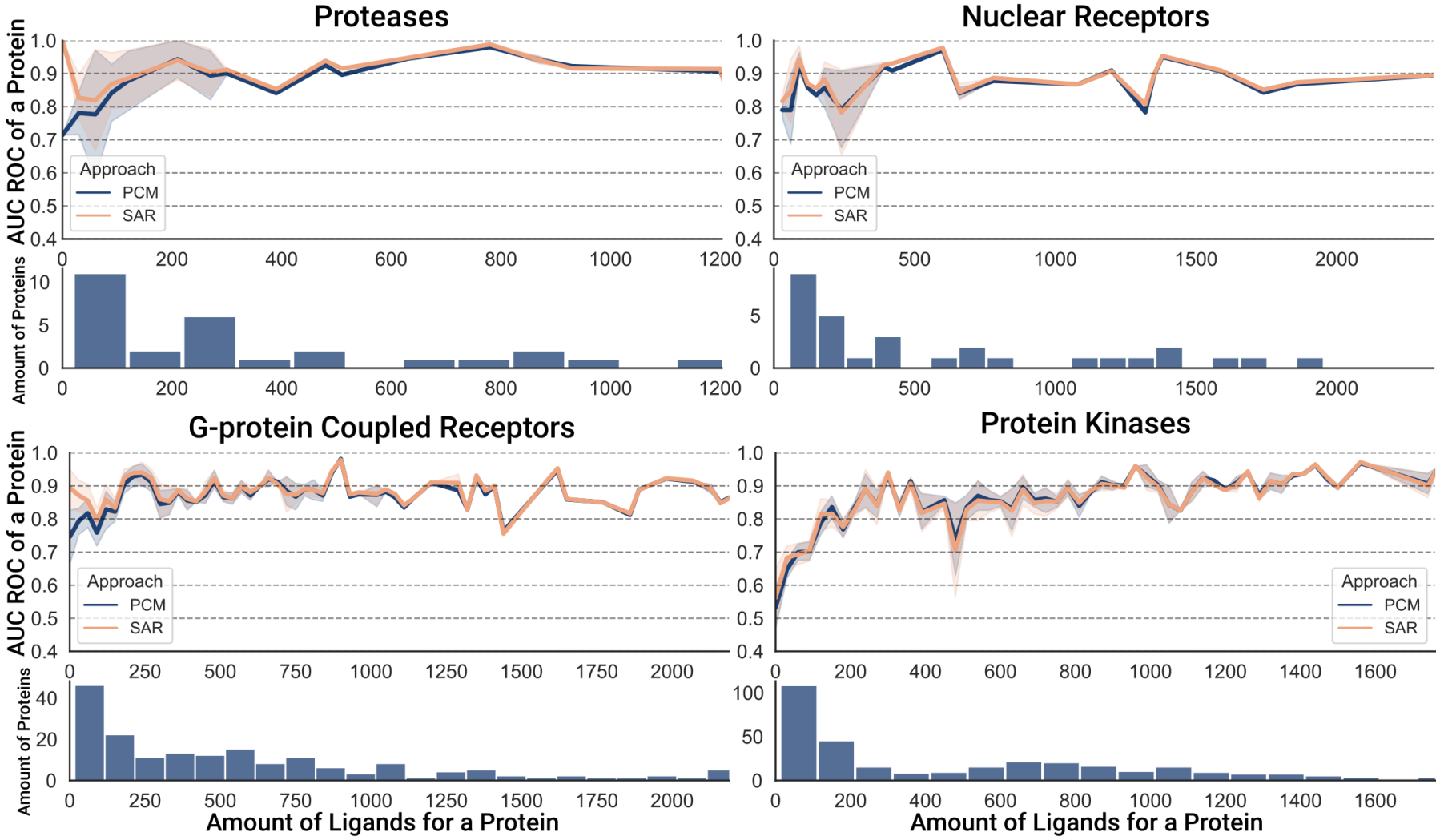


CV – Cross-validation

Results. AUC ROC of predictions

Protein family	Approach	Common scheme	Target-centric scheme
Proteases	SAR	0.875	0.875
	PCM	0.943	0.866
Nuclear Receptors	SAR	0.875	0.875
	PCM	0.913	0.871
GPCR	SAR	0.863	0.863
	PCM	0.947	0.854
Protein Kinases	SAR	0.799	0.799
	PCM	0.928	0.798

Results. Dependence from representation



Conclusions

- There is no evidence that PCM has an advantage over SAR in S1, so there is no need to increase computational complexity by involving protein feature vector.

- Our validation strategy fairly compares efficacy of SAR and PCM, so we suggest using it in further analyses.

Thank you for your attention!

Now I will answer your questions, if any.