SMILES-based Molecular Similarity Functions for Drug-Target Interaction Prediction

Hakime Öztürk1, Elif Ozkirimli2 and Arzucan Özgür1

1 Department of Computer Engineering, Bogazici University, Bebek, 34342 Istanbul, Turkey
2 Department of Chemical Engineering, Bogazici University, Bebek, 34342 Istanbul, Turkey

Abstract

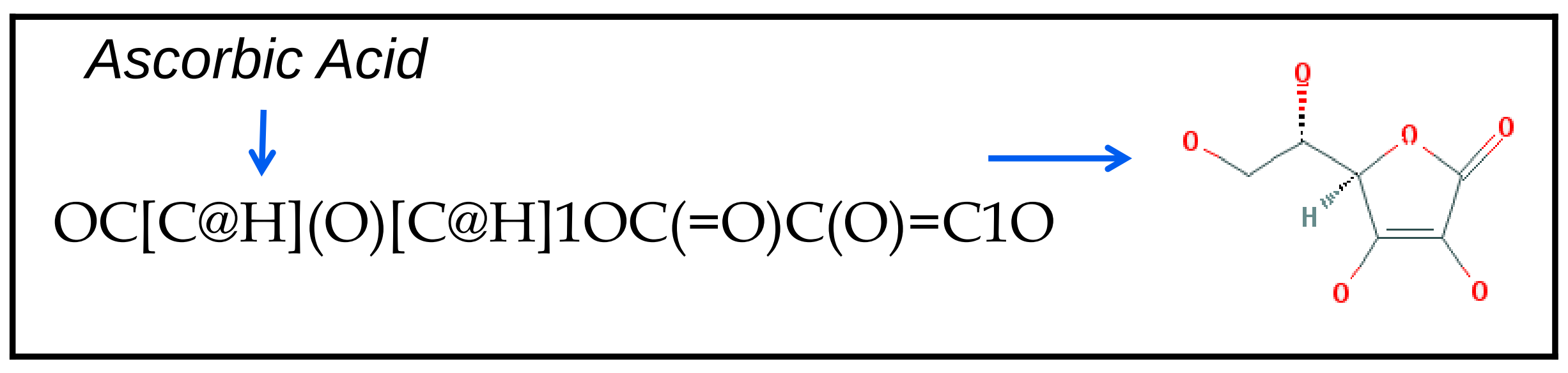
SMILES is a way of describing molecular structures in the form of strings. Considering that each molecule is represented as a string, the similarity between compounds can be computed using SMILES-based string similarity functions. In this study, several SMILES-based functions and popular string similarity functions are adapted and evaluated for drug-target interaction prediction.

Background

❖ The drug-target interaction methods utilize similarity information of drugs as well as targets.

❖ SMILES representation of the compound allows adaptation of different string similarity functions, which are easy to develop and respond fast.

Figure 1: SMILES and 2D representation of a sample compound



❖Weighted Nearest Neighbor-Gaussian Interaction Profile (WNN-GIP) model is utilized in [1] to predict drug-target interacton, and 2D-based similarity kernel **SIMCOMP** is used for comparison.

❖The benchmark data sets: GPCRs, enzymes, nuclear receptors, ion channels, and their interacting ligands, are utilized for performance evaluation [2]

Table 1: The benchmark data set [2].

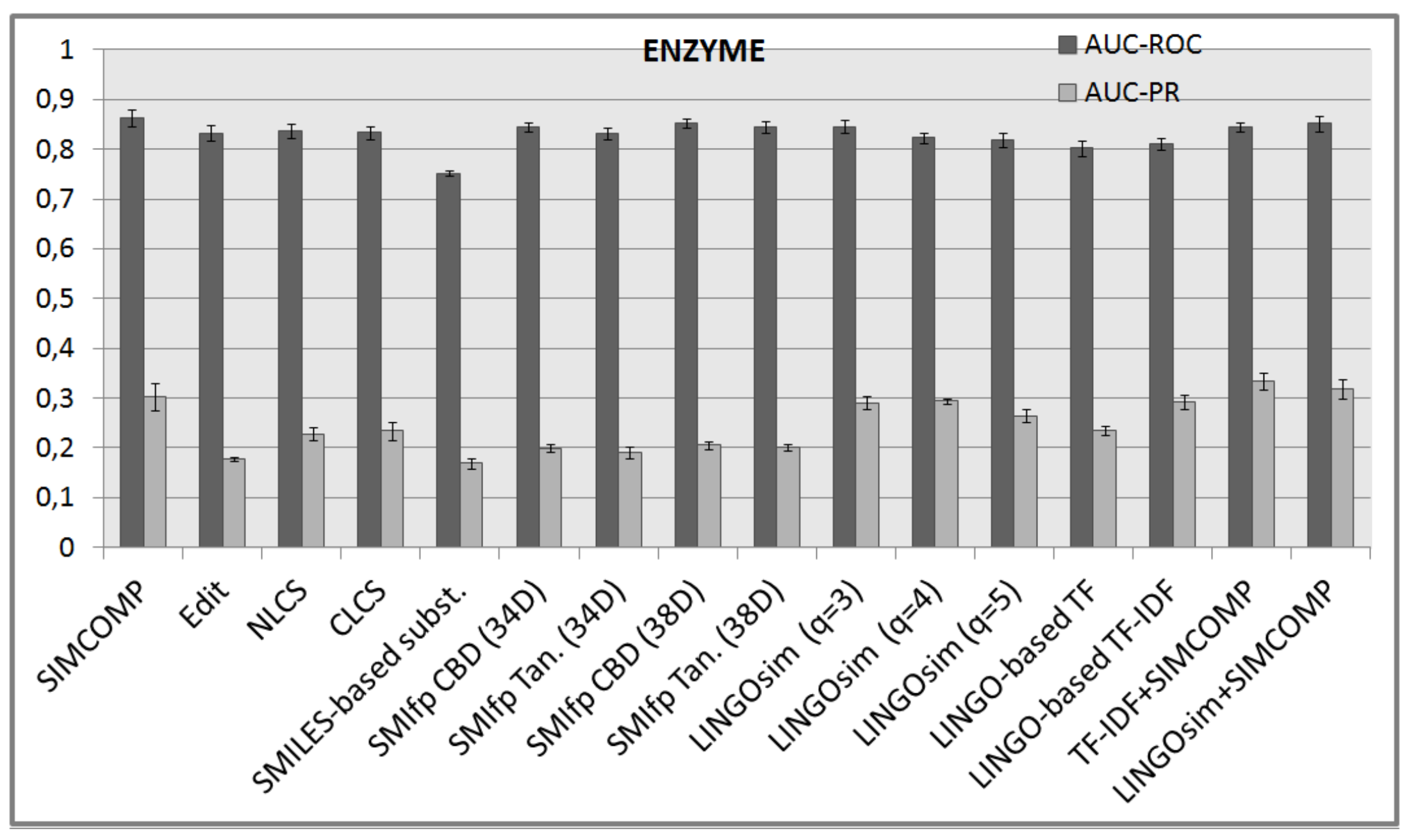
Dataset	Drugs	Targets	Interct.
Enzyme	445	664	2926
GPCR	210	204	1476
Ion Ch.	223	95	635
Nuc. Recp.	54	26	90

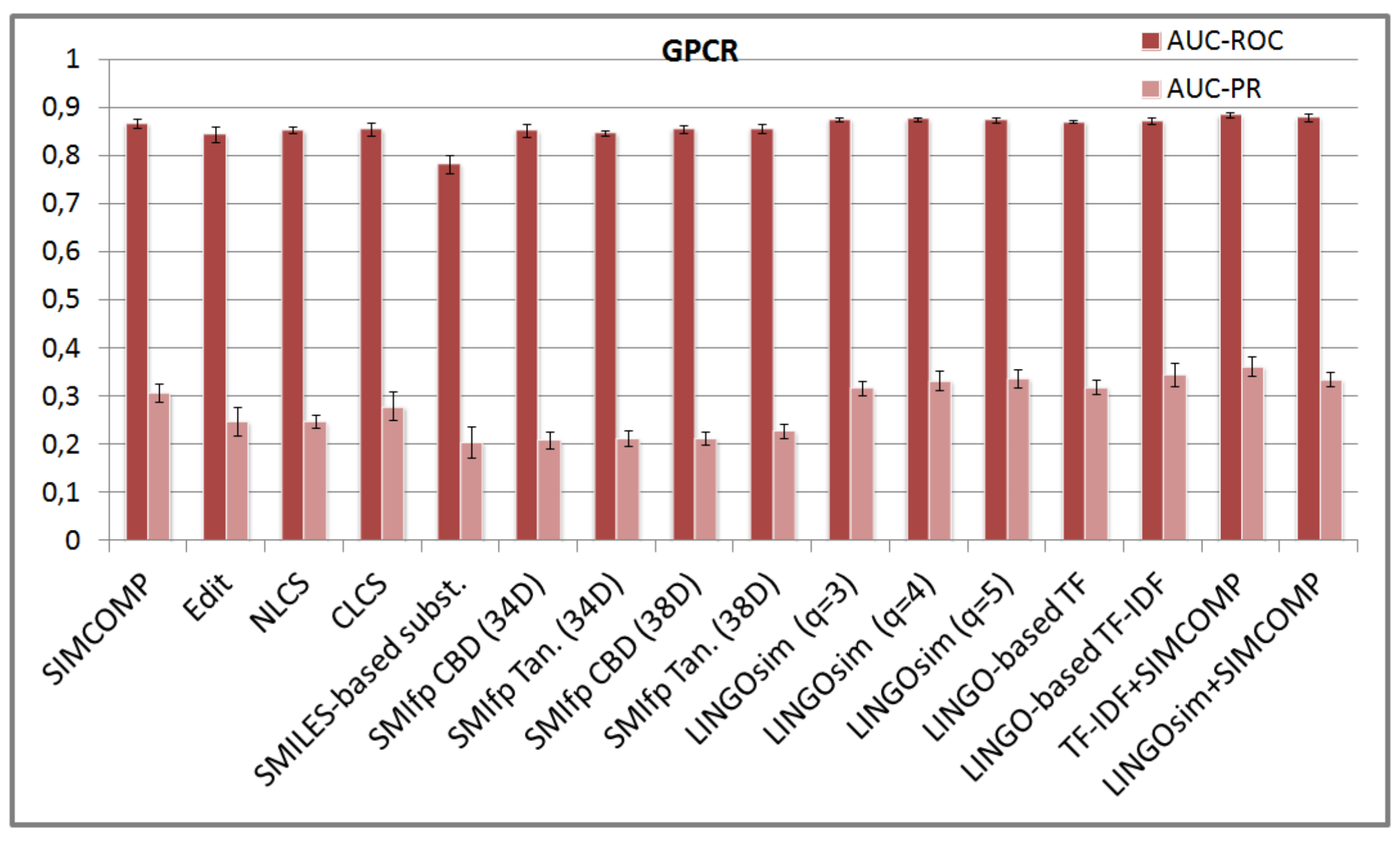
METHODS

	Features
String Similarity Functions	
Edit Distance	$Sim(SMI_1, SMI_2) = 1 - \frac{edit(SMI_1, SMI_2)}{MAX(length(SMI_1), length(SMI_2))}$
Normalized Longest Common Subsequence(NLCS)	$NLCS(S_1, S_2) = \frac{length(LCS(S_1, S_2))^2}{length(S_1) \times length(S_2)}$
Combination of LCS Models (CLCS)	Three different LCS methods are combined.
SMILES-based Similarity Functions	
SMILES representation-based String Kernel	$K(S_1, S_2) = \langle \theta(S_1), \theta(S_2) \rangle$ The function represents the frequencies of all possible strings with length at least *q*=2.
SMILES Fingerprint (SMIFp) Kernel	SMILES strings are represented as 34D/38D character frequency vectors. The distance between the vectors is calculated with City Block/ Tanimoto Distance.
LINGO (*q*=3,4,5)	LINGO represents *q*-character substrings created from SMILES string. $LINGOsim = \frac{\sum_{i=1}^{m} 1 - \frac{\lvert N_{S_1,i} - N_{S_2,i} \rvert}{\lvert N_{S_1,i} + N_{S_2,i} \rvert}}{m}$
Our Methods	
LINGO-based Term- Frequency (TF) Cosine Similarity	Each SMILES string is treated as a ***document*** and four character LINGOs, which are created from these strings, are denoted as ***terms***. SMILES strings are converted into feature vectors (*dimensionality equal to the num. of the unique LINGOs in the compound data set)* where each feature is equal to the TF of that LINGO in the SMILES string.
LINGO-based Term Frequency-Inverted Term Frequency (TF-IDF) Cosine Similarity	Feature vectors are created with TF-IDF values, and the similarity is determined according to the cosine angle between them.

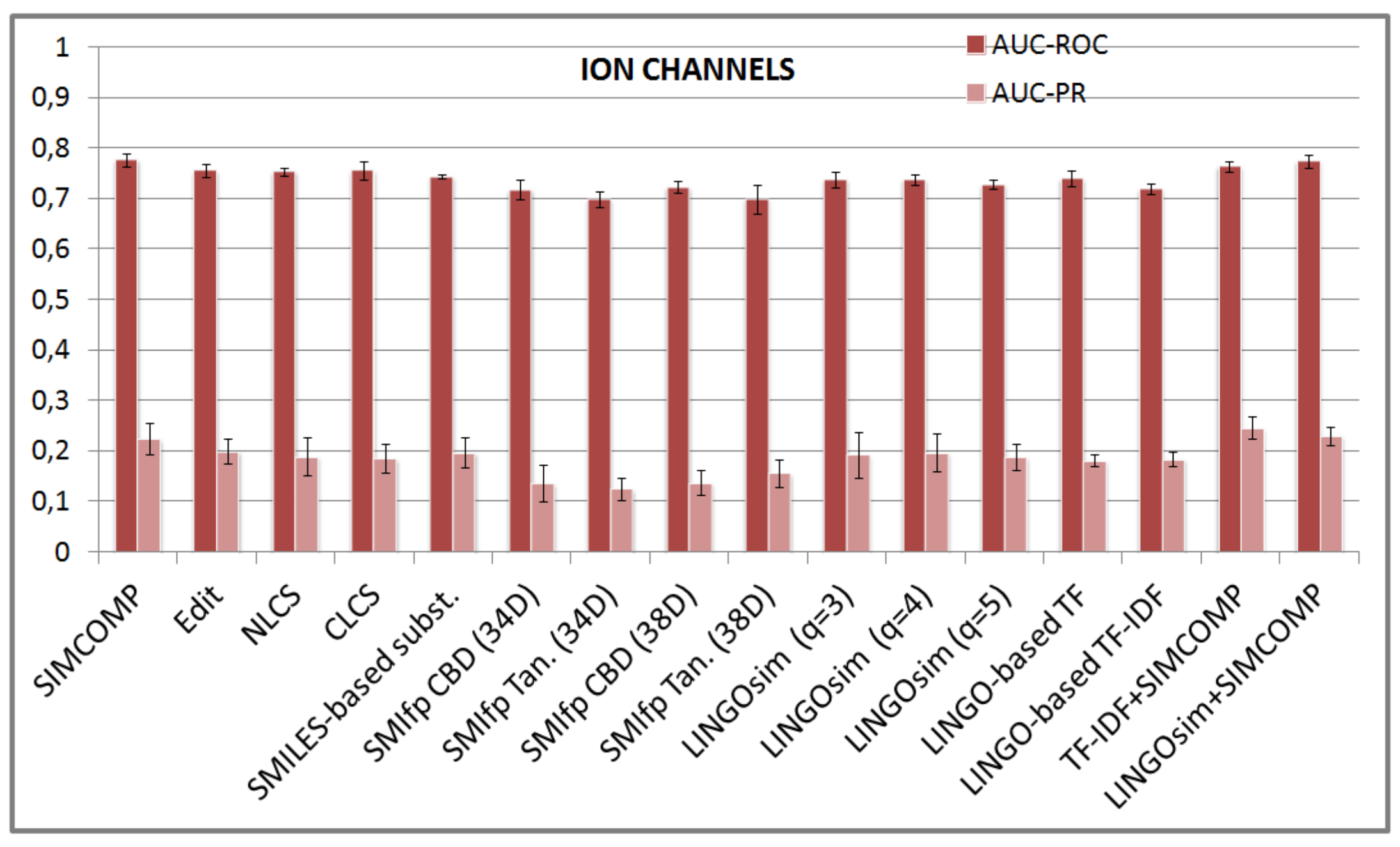
RESULTS

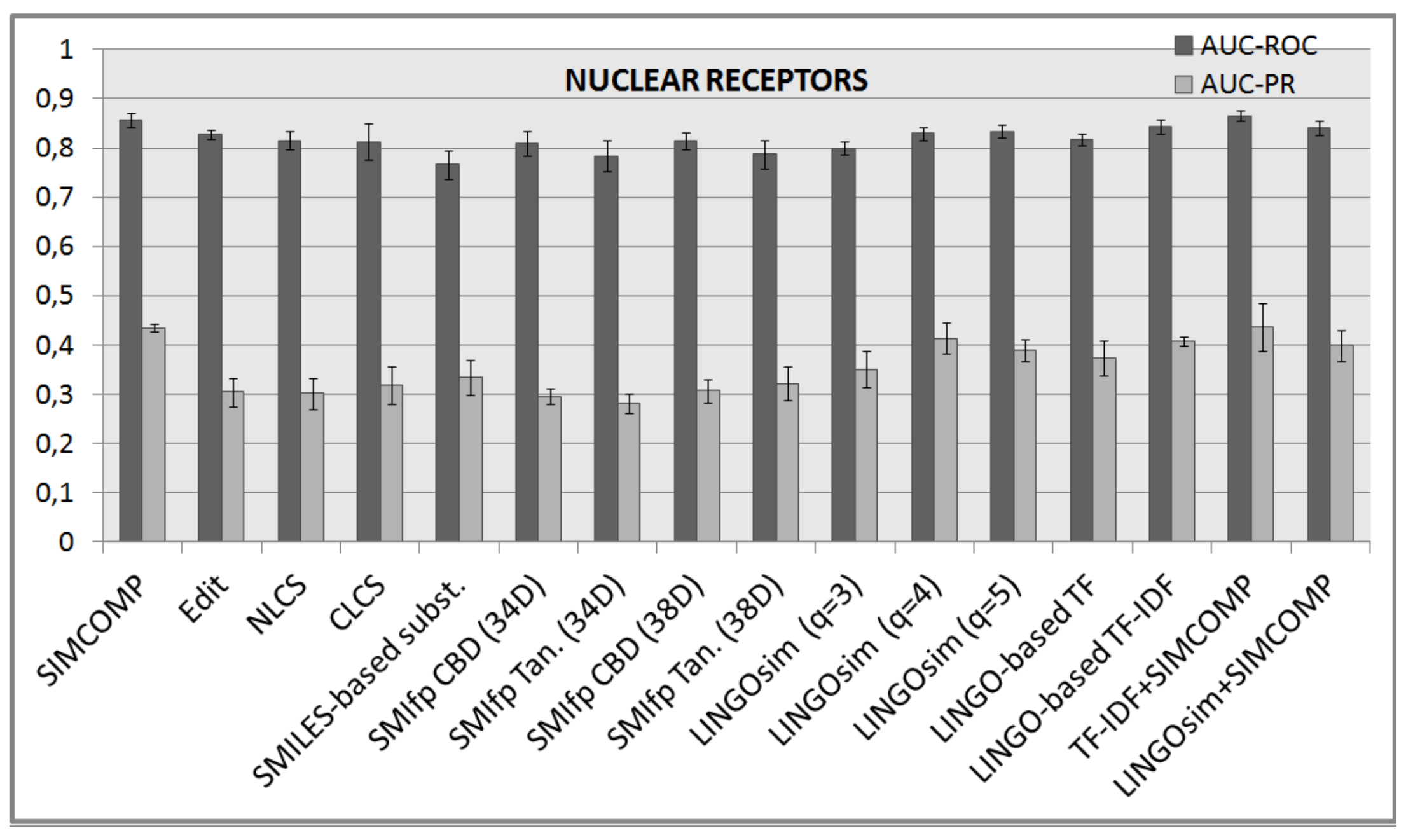
The performances of the kernels are compared using the Area Under the ROC Curve (AUC-ROC) and Area Under the Precision-Recall curve (AUC-PR) metrics. AUC-ROC presents the relation of True-Positive rate to the False-Positive rate, whereas AUC-PR shows the proportion of precision to recall.





LINGOsim (*q* = 4) and LINGO-based TF-IDF cosine similarity performs significantly better AUC-ROC scores than SIMCOMP on the GPCR data set.
The composition of TF-IDF cosine similarity with SIMCOMP produces the best AUC-ROC scores on the GPCR and Nuclear Receptors data sets.





Max and min run times for SIMCOMP are 35 mins and 30 secs respectively, while the max. and min. run times for the fastest SMILES-based string similarity function are 1 secs and 0,1 secs.

Conclusion

❖This work provides a comparison of the 13 different methods that utilize SMILES representation of molecules to measure their chemical similarity for protein-drug interaction prediction.

❖ 1D based methods of molecular similarity perform almost as well as he 2D based methods in the protein-drug interaction prediction task.

❖ The experiments indicate that SMILES based kernels are significantly faster than the 2D-based SIMCOMP.

❖ The application of TF and TF-IDF weighting to the SMILES similarity calculation domain gives promising results.

References

[1] Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W.,Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics **24**(13), 232–240 (2008).

[2] van Laarhoven, T., Marchiori, E.: Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS ONE **8**(6), 66952 (2013).

[Figure 1] https:// pubchem. ncbi.nlm.nih.gov/ substance/ 7847086

CONTACT:
{hakime.ozturk, arzucan.ozgur, elif.ozkirimli}
@
boun.edu.tr