Interpreting Chemical Words of a Data-driven Segmentation Method as Protein Family Pharmacophore and Functional Groups

Asu Büşra Temizer[1], Rıza Özçelik[2], Taha Koulani[1], Gökçe Uludoğan[2], Elif Ozkirimli[3], Kutlu O. Ulgen[4], Nilgün Karalı[1] and Arzucan Özgür[2]

1 Faculty of Pharmacy, Department of Pharmaceutical Chemistry, İstanbul University, İstanbul, Turkey
2 Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey
3 Data and Analytics Chapter, Pharma International Informatics, F. Hoffmann-La Roche AG, Switzerland
4 Department of Chemical Engineering, Boğaziçi University, İstanbul, Turkey

Scan for preprint

INTRODUCTION

- Representations of compounds as text, such as SMILES, which enables the application of NLP techniques.
- Chemical compounds can be thought of as composed of smaller building blocks, similar to sentences in natural languages ⟶ Chemical words
- Subword tokenization algorithms have been commonly used to identify the chemical words.
- A novel pipeline was proposed to highlight chemical words as pharmacophore and functional group candidates.

METHODS

- Database: BDB[1], LitPCBA[2] and ProtBENCH[3].
- Chemical vocabulary identified with a data-driven segmentation method called Byte-Pair Encoding (BPE).
- Rank the chemical words based on TF-IDF scores (Figure 1).
- Compare the importance scores of the chemical words in strong binders with those in weak binders.

CONCLUSION

- Chemical-word-based models rely on chemically meaningful building blocks.
- The highlighted chemical words can designate pharmacophores and/or functional groups for all studied protein families.
- We have obtained results that can improve the ADME (absorption, distribution, metabolism, and excretion) profiles of compounds.
- These findings may also guide the design and development of new target compounds.

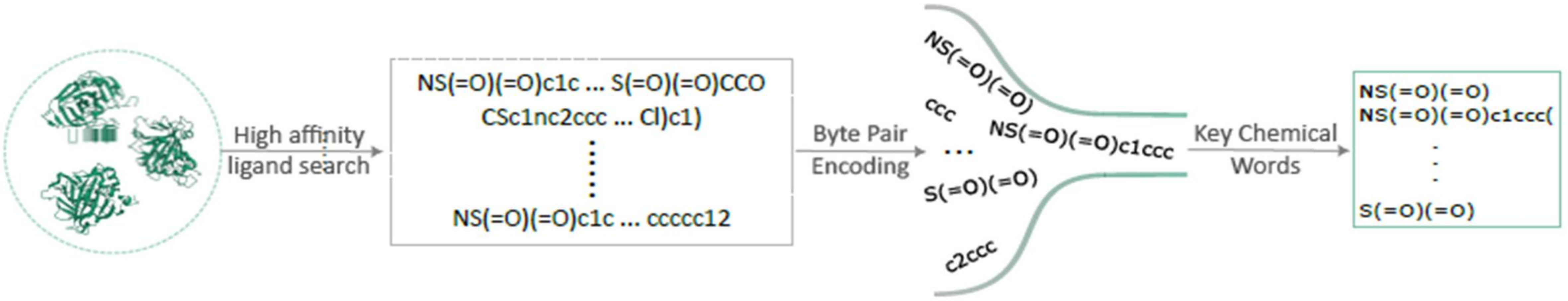


Figure 1. The proposed pipeline to highlight key chemical words for strong binding to a protein or protein family.

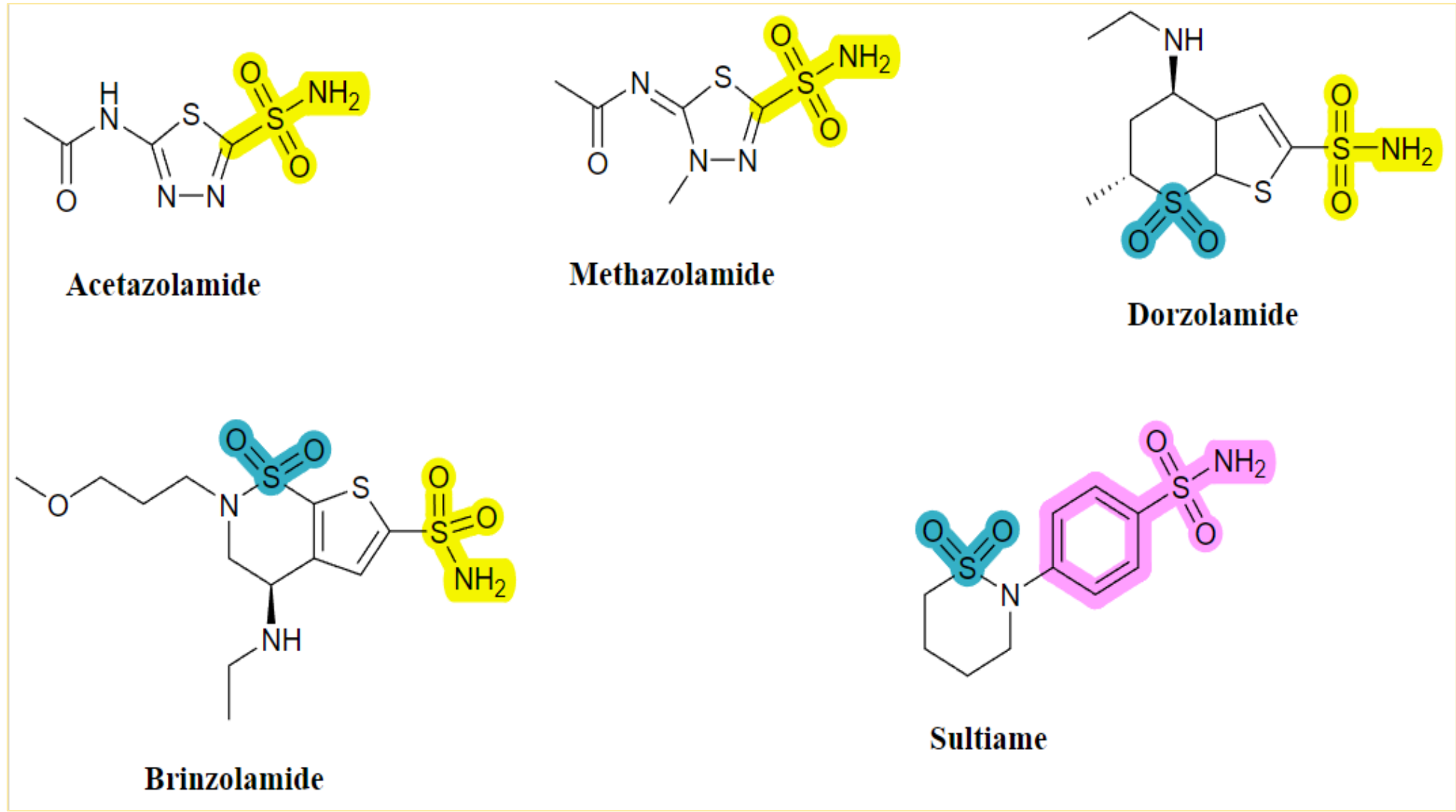


Figure 3. The chemical words, NS(=O)(=O), NS(=O)(=O)c1ccc, S(=O)(=O), proposed by the algorithm as the key chemical words (sulfonamide, aryl-substituted sulfonamide, and sulfone groups), are marked in yellow, pink, and green, on the drug molecules, respectively. A detailed study of the literature reveals that the highlighted chemical words are pharmacophore groups for the CA enzyme family.

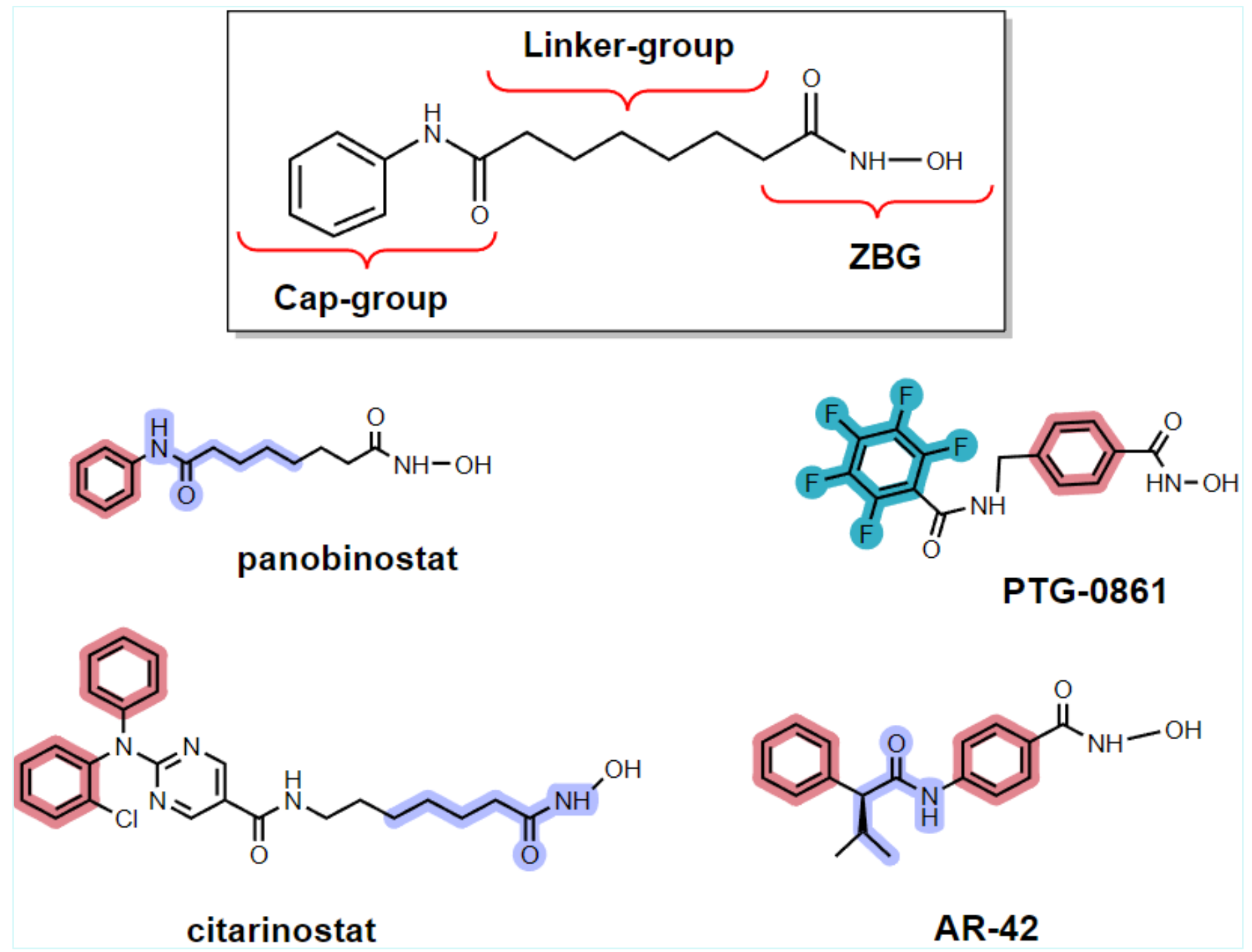


Figure 4. The chemical words, NC(=O)CCCC, CCC(=O)N, c1c(F)c(F)c(F)c(F)c1F, proposed by the algorithm as the key chemical, are marked on the molecules. These words represent pentanamide, propanamide, and pentafluorobenzene moieties, respectively.

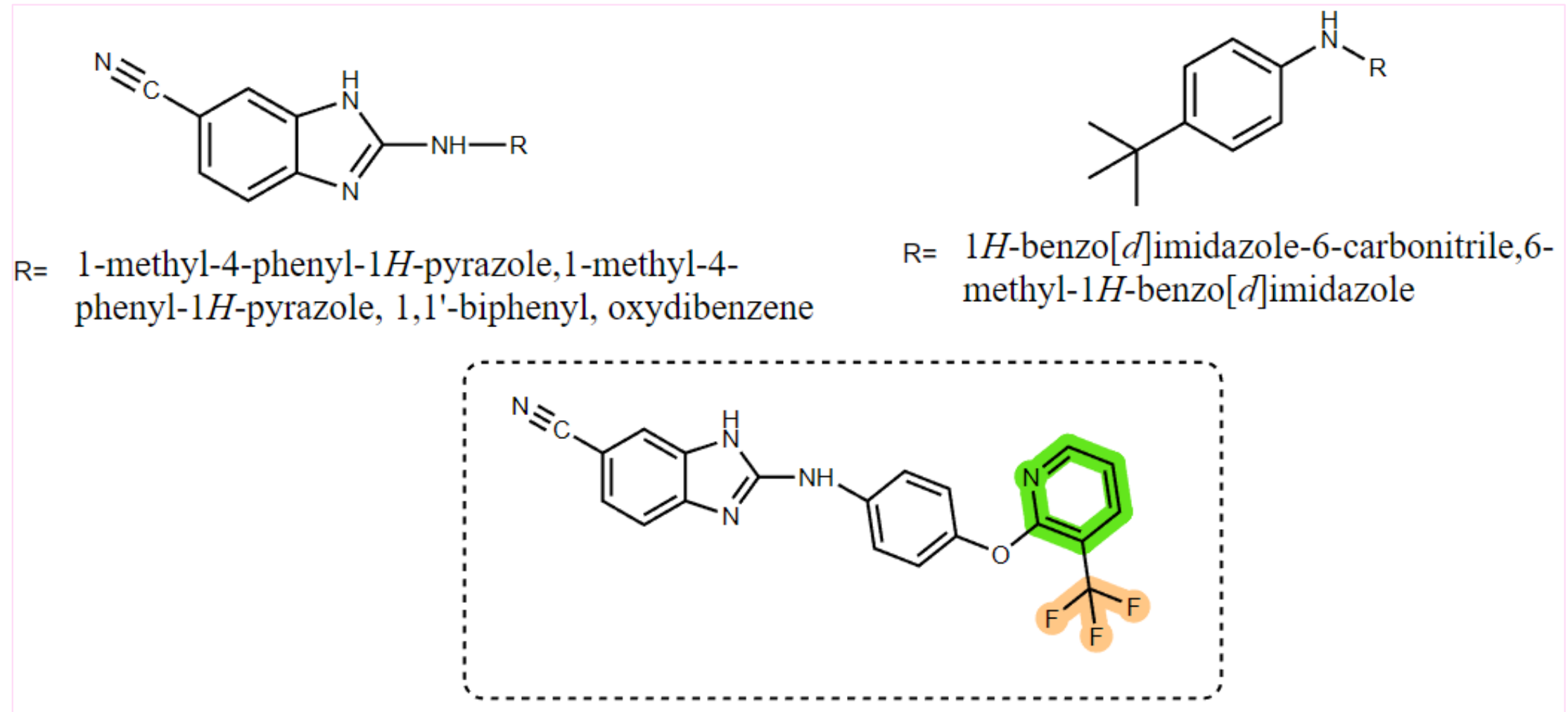

R= 1-methyl-4-phenyl-1H-pyrazole,1-methyl-4-phenyl-1H-pyrazole, 1,1'-biphenyl, oxydibenzene
R= 1H-benzo[d]imidazole-6-carbonitrile,6-methyl-1H-benzo[d]imidazole

Figure 5. Structure-activity relationship (SAR) studies of 2-phenylamino-6-cyano-1*H*-benzimidazole derivatives showed that the compound bearing trifluoromethyl at the pyridine ring is a potent CK1 gamma inhibitör and exhibits unprecedented selectivity for the CK1 isoform. The chemical words c1ccncc1and C(F)(F)F)CC1proposed by the algorithm represent pyridine and the trifluoromethyl groups. The chemical words c1ccncc1,c1ccncc1),C(F)(F)F)CC1 proposed by the algorithm include pyridine, pyridine and the trifluoromethyl groups, which are compatible with the literature data.

RESULTS

We selected several protein or protein families as case studies then scan the chemical substructures highlighted by the key chemical words in the literature (Figure 3-5, Table 2).

Table 1. The number of highlighted words and the mean number of protein/family per highlighted word for each dataset.

Dataset	# Highlights	Mean Number of Protein/Family per Highlight
BDB	531	3.01
LitPCBA	147	1.02
ProtBENCH/Epigenetic Regulators	538	1.53
ProtBENCH/Hydrolases	2191	1.94
ProtBENCH/Ion Channels	1068	1.45
ProtBENCH/Membrane Receptors	2912	1.80
ProtBENCH/Other Enzymes	1430	1.59
ProtBENCH/Oxidoreductases	1809	1.59
ProtBENCH/Proteases	1545	1.82
ProtBENCH/Transcription Factors	664	1.41
ProtBENCH/Transferases	2969	2.25
ProtBENCH/Transporters	865	1.33

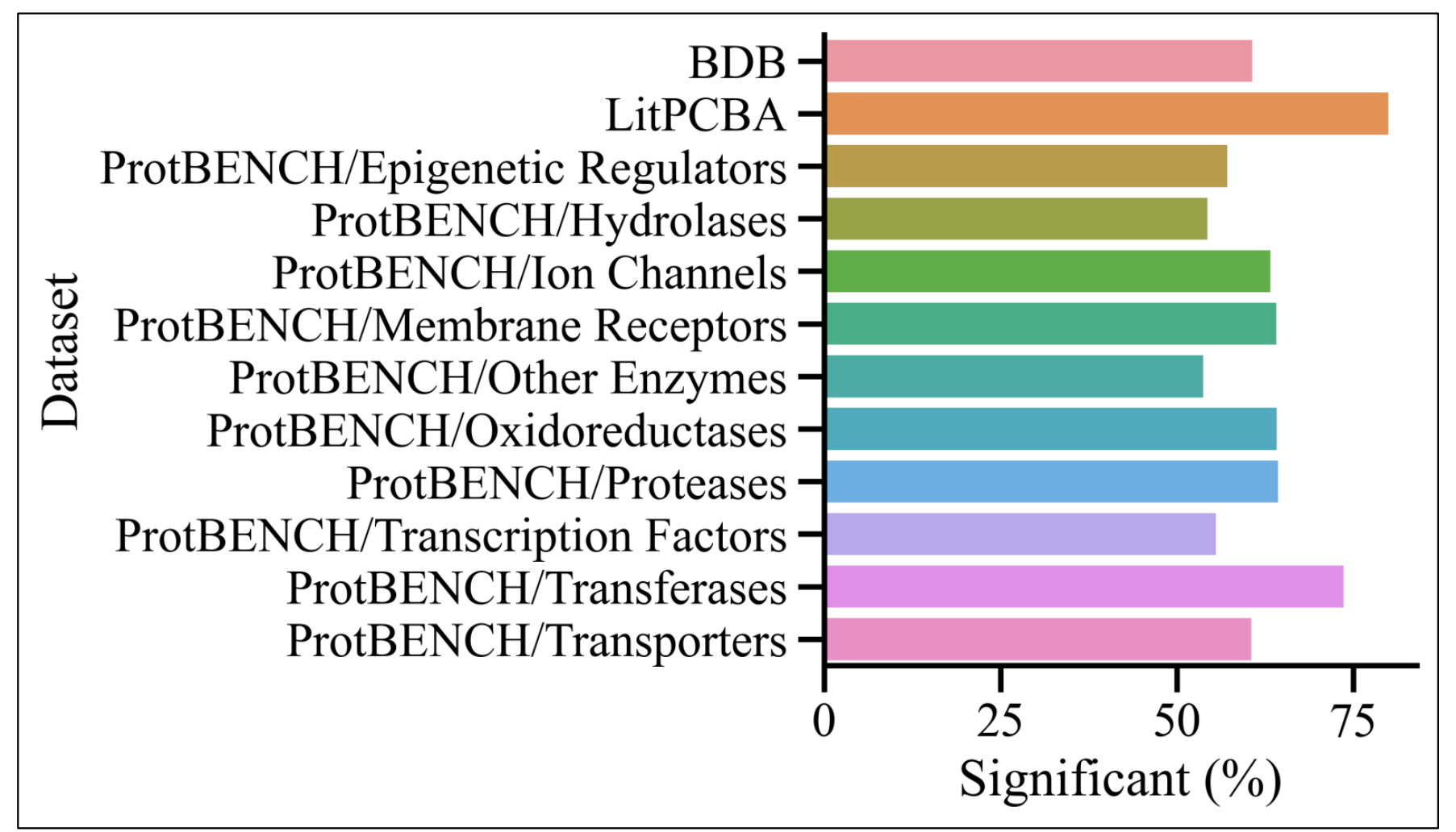
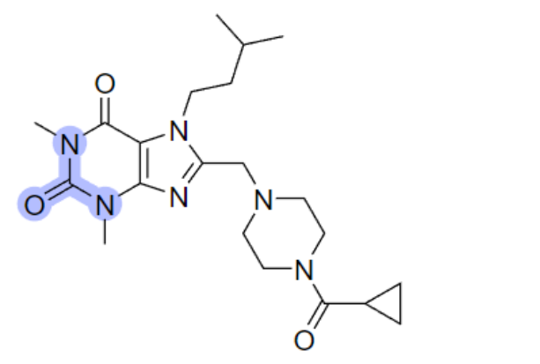
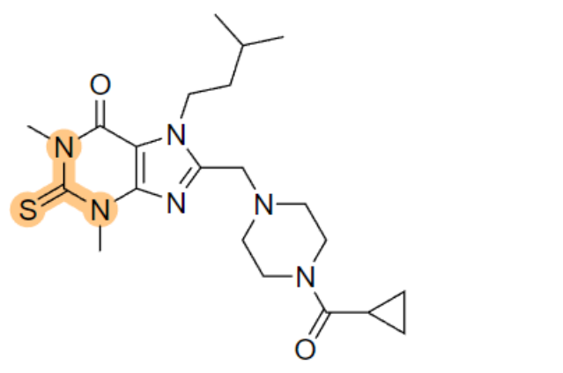


Figure 2. Percentage of protein/family where the importance scores of words between strong and weak binders are significantly different.

Table 2. The physicochemical parameters of NCT-501 and NCT-501- NC(=S)N

Structural Formula		
Mol. W. (≤ 500)	416.522	432.583
LogPo/w (≤ 5)	1.363	2.159
HBA (≤ 5)	0	0
HBD (≤ 10)	10	9.5
Oral Absorption (<80 high, <25 weak activity)	72.846	82.307

Acknowledgments
This work is supported by the Scientific and Technological Research Council of Turkey (Grant Number 119E133).

References
[1] Özçelik, R., Öztürk, H., Özgür, A., & Ozkirimli, E. (2021). Chemboost: A chemical language based approach for protein–ligand binding affinity prediction. Molecular Informatics, 40(5), 2000212.
[2] Tran-Nguyen, V. K., Jacquemard, C., & Rognan, D. (2020). LIT-PCBA: an unbiased data set for machine learning and virtual screening. Journal of chemical information and modeling, 60(9), 4263-4273.
[3] Atas Guvenilir, H., & Doğan, T. (2023). How to approach machine learning-based prediction of drug/compound–target interactions. Journal of Cheminformatics, 15(1), 1-36.

Asu Büşra Temizer
✉ asubusra.temizer@ogr.iu.edu.tr