Exploring data-driven chemical SMILES tokenization approaches to identify key protein–ligand binding moieties

Asu Busra Temizer¹,² | Gökçe Uludoğan³ | Rıza Özçelik³ | Taha Koulani¹,² | Elif Ozkirimli⁴ | Kutlu O. Ulgen⁵ | Nilgun Karali¹ | Arzucan Özgür³

¹Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Istanbul University, Istanbul, Turkey
²Department of Pharmaceutical Chemistry, Institute of Health Sciences, Istanbul University, Istanbul, Turkey
³Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
⁴Science and Research Informatics, F. Hoffmann-La Roche Ltd, Basel Switzerland
⁵Department of Chemical Engineering, Boğaziçi University, Istanbul, Turkey

Correspondence
Asu Busra Temizer, Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Istanbul University, Istanbul, Turkey.
Email: asubusra.temizer@ogr.iu.edu.tr
Arzucan Özgür, Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey.
Email: arzucan.ozgur@boun.edu.tr

Abstract
Machine learning models have found numerous successful applications in computational drug discovery. A large body of these models represents molecules as sequences since molecular sequences are easily available, simple, and informative. The sequence-based models often segment molecular sequences into pieces called chemical words, analogous to the words that make up sentences in human languages, and then apply advanced natural language processing techniques for tasks such as de novo drug design, property prediction, and binding affinity prediction. However, the chemical characteristics and significance of these building blocks, chemical words, remain unexplored. To address this gap, we employ data-driven SMILES tokenization techniques such as Byte Pair Encoding, WordPiece, and Unigram to identify chemical words and compare the resulting vocabularies. To understand the chemical significance of these words, we build a language-inspired pipeline that treats high affinity ligands of protein targets as documents and selects key chemical words making up those ligands based on tf–idf weighting. The experiments on multiple protein–ligand affinity datasets show that despite differences in words, lengths, and validity among the vocabularies generated by different subword tokenization algorithms, the identified key chemical words exhibit similarity. Further, we conduct case studies on a number of targets to analyze the impact of key chemical words on binding. We find that these key chemical words are specific to protein targets and correspond to known pharmacophores and functional groups. Our approach elucidates chemical properties of the words identified by machine learning models and can be used in drug discovery studies to determine significant chemical moieties.
1 | INTRODUCTION

The last decade has witnessed the rise of machine learning. The models learned to sing [1], write [2], and paint [3] through large unlabeled datasets, despite lacking well-defined rules for the tasks. Drug discovery is also an excellent application domain for machine learning models. However, unlike images and audios, chemicals are non-numeric and need intermediate representations on which machines can learn.

Text-based representations of chemicals [4], can be used as intermediate representations: they are easily available, simple, and as powerful as more complex representations such as 2D or 3D representations [5]. The power of the text-based representations in machine learning is partially due to the chemical language perspective. The chemical language perspective views text-based representations of chemicals as documents written in a chemical language and borrows advanced approaches from the natural language processing domain to build computational drug discovery models [6]. Successful applications of the chemical language perspective in drug discovery include de novo drug design [7], binding site detection [8], and drug–target interaction prediction [9].

At the core of the chemical language perspective are chemical words, which correspond to the smaller building blocks of chemicals, similar to the words in natural languages. On the other hand, defining the chemical words poses another research problem since the chemical language, unlike natural (human) languages, has no pre-defined collection of words, i.e., a vocabulary. Several approaches are available to build a chemical vocabulary for state-of-the-art models for downstream tasks such as drug-target affinity prediction, similar compound identification, and de novo drug design [10]. However, to the best of our knowledge, the chemical vocabularies obtained from text-based representations are not studied from a chemical perspective so far, and therefore it is unknown whether these chemical vocabularies capture chemical information or not. This raises the question: do models utilizing chemical words rely on chemically meaningful building blocks, or arbitrary chemical subsequences? Here we seek an answer to this question in order to define the role of these models in new drug design and development from a medicinal chemist’s perspective.

Chemical word is an integral concept for the studies that rely on the chemical language hypothesis [6]. The chemical language hypothesis interprets the text-based representation (i.e. molecular strings) of chemicals as sentences and borrows language processing methods from natural languages. Chemical words are the building blocks of these “sentences”, and, unlike the words in natural languages, need to be discovered.

Subword tokenization algorithms allow learning the chemical words from large unlabeled corpora of molecular strings. In addition to discovering chemical words, they can segment the chemical sentences into chemical words in the discovered vocabulary. One commonly used algorithm in this context is Byte Pair Encoding (BPE) [11], which is widely adopted in various applications such as de novo drug design [10c], molecular property prediction [12], and protein–ligand binding affinity prediction [10a]. BPE starts with an initial vocabulary of individual characters and iteratively merges the most frequently occurring pairs until a desired vocabulary size is reached. Another subword tokenization method used in molecular design [13] is WordPiece [14], which is similar to BPE in that it also begins with an initial vocabulary and continues merging pairs of tokens. However, unlike BPE, WordPiece selects the pair that maximizes the likelihood of the training data, rather than simply the most frequent one. In contrast to both BPE and WordPiece, Unigram [15] is a top–down tokenization method that starts with an initial vocabulary of all possible words and reduces it based on the likelihood of the unigram model, and it has recently been used in molecular fingerprinting [16].

We adopt three widely utilized subword tokenization algorithms, namely BPE, WordPiece, and Unigram, which have been successfully applied in computational drug discovery tasks [10a,c, 12,13,16] and compare the vocabularies generated by these methods. However, due to the large number of chemical words in the resulting vocabularies, it is impractical to interpret them all manually. Therefore, we propose a novel language-inspired pipeline to identify key chemical words that are strongly associated with protein–ligand binding. In natural languages, each genre or period has a certain distribution of words. Similarly, each protein family can be expected to have a certain distribution of substructures (or chemical words) to which it would bind. Therefore, we focused on the high affinity ligands of protein families to build protein family specific vocabularies and analyzed the
chemical significance of subwords for the protein family. Protein–ligand binding is selected as the chemical investigation perspective since strong binding relationships are fundamental in drug discovery pipelines, are widely available in the literature, and naturally group chemicals by their binding targets.

Identifying key substructures within a molecular set is not a new concept. There is a significant body of work that utilizes data mining techniques to extract meaningful substructures for a specific property or objective [17]. Some examples include frequency-based methods that rank these substructures using metrics like enrichment factor [18], likelihood ratio [19], and statistical tests [20]. Yamanishi et al. [21] have used sparse canonical correspondence analysis (SCCA) to identify a number of chemical substructures common among drugs with the ability to bind to a specific set of protein domains. More recently, through a combination of fragment-based analysis, frequency-based analysis and traditional machine learning techniques, important structures for MMP-2 inhibitors were investigated by Banerjee et al. [22]. A notable difference in our study is the focus on vocabularies obtained through subword tokenization. Moreover, unlike the study by Banerjee et al., our methodology is able to discover family-specific chemical words, as these words are evaluated based on their inverse frequency across families.

We apply the proposed pipeline to protein–ligand binding affinity datasets, identifying ten chemical words for each protein or protein family in the datasets. We find that while the vocabularies generated by different subword tokenization algorithms differ in words, lengths, and validity, the identified key chemical words are similar, as measured by the mean edit distance between protein or family-specific top ranking subwords with different vocabularies. We also observe that the selected words are protein or family-specific, often associated with only one protein/family, and significantly different from the words identified for weak binders. As a case study, we examine the selected chemical words for a number of important drug target families and find that the top-ranking chemical words are associated with the known pharmacophores and functional groups of the protein families. Notably, for the aldehyde dehydrogenase 1 enzyme family, the chemical words selected by the proposed algorithm were found to improve the drug-likeness of the molecules. Our results corroborate the chemical word-based models in computational drug discovery and are a step toward interpreting the computational drug discovery models from a pharmaceutical chemistry perspective.

2 | MATERIALS AND METHODS

In this work, we build a pipeline to analyze the chemical significance of the chemical words of high affinity ligands of proteins or protein families. The vocabularies are created for each protein family based on the hypothesis that the important subwords of each protein family will be specific to the protein family. The top ranking words for the ligands of carbonic anhydrases (EC. 4.2.1.1) and casein kinase 1 gamma (EC. 2.7.11.1) in BindingDB (BDB), as well as pyruvate kinase M2 (EC. 2.7.1.40) and aldehyde dehydrogenase 1 (EC.1.2.1.3) from Lit-PCBA are examined for chemical relevance.

2.1 | Segmenting chemicals into chemical words

In this study, we adopt three commonly used subword tokenization algorithms: BPE [11], WordPiece [14], and Unigram [15], and learn vocabularies with sizes of 8 K, 16 K, and 32 K. The vocabularies are identified by applying the tokenization algorithms on the SMILES representations of ~2.3 M compounds in ChEMBLv27 [23], and then, the vocabularies are used to segment the compounds into their chemical words.

2.2 | Characterizing chemical words

Subword tokenization algorithms, adopted from the field of natural language processing, have recently been widely used to identify chemical words and represent compounds in computational drug discovery studies. However, the identified vocabularies have not been characterized and a comparison of the different methods has not been performed yet. To investigate the chemical words learned by different tokenization methods, we analyze various statistics, such as word length, validity, vocabulary overlaps across the algorithms, and word similarity to the most similar extended functional group, which is a generalized version of traditional functional group and introduced by Lu et al. [24].

2.3 | Term Frequency–Inverse Document Frequency

Term Frequency–Inverse Document Frequency (TF–IDF) [25] is a document vectorization algorithm originally introduced in the field of information retrieval. TF–IDF represents documents based on the importance of each word in the document’s vocabulary. TF–IDF
postulates that the importance of a word for a document is proportional to the word’s frequency in the document and the inverse number of documents the word appears in the corpus. The TF–IDF vector for a document \( D \) is formulated as:

\[
\tilde{D} = [tf_{w_1,D} \times idf_{w_1}, \cdots, tf_{w_n,D} \times idf_{w_n}]_{1 \times V}
\]  

(1)

where \( tf_{w_i,D} \) is the count of the \( i \)th word in the vocabulary, \( w_i \) in \( D \); \( idf_{w_i} \) is the natural logarithm of the number of documents in the corpus divided by the number of documents in which \( w_i \) is present; and \( V \) is the number of different words in the corpus, i.e., the vocabulary size. However, the naive TF–IDF weighting may consider twenty occurrences of a word in a document twenty times more significant than a single occurrence, which may not be accurate. To address this issue, a sublinear term frequency scaling variant of the algorithm has been introduced \([25]\), which uses the logarithm of the term frequency and assigns a weight given by:

\[
tf_s(w, D) = \begin{cases} 
1 + \log tf_{w,D} & \text{if } tf_{w,D} > 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(2)

where \( tf_{w,D} \) is the count of word \( w \) in document \( D \) and \( tf_s(w, D) \) is its scaled counterpart used in the importance computation instead.

### 2.4 Datasets

We used three different datasets of protein–ligand affinity. The BDB dataset contains affinity information for 31 K interactions of 924 compounds and 490 proteins from 81 protein families. The dataset provides protein sequences, compound SMILES strings, and affinity values represented as \( pK_d \) and has been used in previous protein-compound interaction prediction studies \([9c, 26]\). To identify protein family-specific words, we map each protein in this dataset to its corresponding PFAM families \([27]\) using the InterPro API \([28]\).

LIT-PCBA \([29]\) includes 15 target proteins and 7844 active and 407,381 inactive compounds along with their corresponding SMILES strings. The dataset is specifically curated for virtual screening and machine learning with efforts made to ensure it is unbiased and realistic.

ProtBENCH \([30]\) contains protein family-specific bioactivity data, spanning multiple protein superfamilies, including membrane receptors, ion channels, transporters, transcription factors, epigenetic regulators, and enzymes with five subgroups (i.e., transferases, proteases, hydrolases, oxidoreductases, and other enzymes). The family subsets vary in number of interactions (19 K—220 K), number of proteins (100—1 K), and number of compounds (10 K—120 K). As this dataset provides ChEMBL ids for compounds, we retrieved the corresponding SMILES strings of the compounds from ChEMBL.

### 2.5 Identifying the key chemical words for strong protein–ligand binding

Here we propose a novel pipeline to identify key chemical words for strong binding to both individual proteins and to protein families. The proposed algorithm is language-inspired and postulates, similar to TF–IDF, that if a chemical word is common and unique to the strong binders of a protein or protein family, then it signifies a key chemical substructure for binding. Accordingly, the algorithm first identifies strong binders for each protein or protein family. While LIT-PCBA already distinguishes the interactions as strong and weak binders, strong binders in other datasets are identified using predefined thresholds such as a \( pK_d \) value higher than 7 for BDB and a bioactivity score \((pChEMBL)\) greater than the median score for ProtBENCH. Next, strong binders of each protein or protein family are represented as a document, in which each strong binding compound is a sentence composed of chemical words identified via the algorithms described in the Segmenting Chemicals into the Chemical Words section.

Finally, documents containing the SMILES representations of the high affinity ligands of each protein or protein family are vectorized with TF-IDF and the chemical words are ranked based on their TF-IDF scores. The proposed pipeline is illustrated in Figure 1.

### 2.6 Comparing key chemical words

While the proposed pipeline aims at identifying key chemical words for strong binders of individual proteins or protein families, it can also be used to compute word importance scores for weak binders. To compare the importance scores of weak binders with those of strong binders, we apply the pipeline on weak binders by considering weak interactions as documents associated with proteins or protein families. Next, we conduct Wilcoxon rank-sum tests \((p < 0.05)\) on a target level to compare the importance scores of chemical words associated with strong binding to each protein or family with those associated with weak binding to that particular target. To investigate whether the chemical words identified by using
different vocabularies are similar, we compute mean edit distance scores between selected chemical words using the following equation:

\[
\frac{1}{|A_p|} \sum_{h_{A_p} \in A_p} \min_{h_{B_p} \in B_p} d(h_{A_p}, h_{B_p})
\]

where \(A\) and \(B\) denote chemical word vocabularies, with \(A_p\) and \(B_p\) being the subsets identified for a specific protein \(p\). The individual words in these subsets are represented by \(h_{A_p}\) and \(h_{B_p}\). The Levenstein edit distance function, \(d(h_{A_p}, h_{B_p})\), is used to determine the distance between two words. This score allows us to compare the similarity of chemical words identified for the same protein or protein family across different vocabularies.

3 | RESULTS

In this study, we investigated the characteristics of the chemical vocabularies generated by different subword tokenization algorithms (Unigram, BPE, and WordPiece). We also conducted case studies on protein families to analyze the impact of key chemical words on binding.

3.1 | Characterizing vocabularies

The mean length of chemical words varied from 9 to 13, depending on the subword tokenization algorithm and vocabulary size. Words identified using Unigram are longer, with the longest word having a length of 19. Extended functional groups (EFG), on average, are even longer with a mean length of approximately 19. There is a trend of increase in mean length illustrated in Figure 2, where the distribution of word lengths across the tokenization methods shows a slight shift towards longer words with increasing vocabulary size.

3.1.1 | BPE and WordPiece yield vocabularies with a larger number of valid words compared to Unigram

Although subword tokenization algorithms can identify chemical words, it is important to note that not all identified words may be chemically valid. However, having valid words can be desirable as it may enhance the interpretability of models and facilitate approaches such as scaffold hopping. We used RDKit [31] to assess validity, which considers a molecule valid if it complies with the principles of chemical bonding and doesn’t exhibit any unusual or unfeasible structures. The valid word ratio of the compared vocabularies ranges from 11% to 17%, with Unigram vocabularies having the lowest ratio, and WordPiece vocabularies achieving the highest ratio as shown in Figure 3. While BPE and WordPiece vocabularies achieve similar valid ratios across vocabulary sizes, the valid ratio of Unigram vocabularies decreases with increasing vocabulary size.

3.1.2 | Do the subword tokenization algorithms yield the same words?

Subword tokenization algorithms, despite having similar approaches, can result in different vocabularies. Figure 4 shows the number of common words between 8 K-sized vocabularies produced by different tokenization methods. BPE and WordPiece vocabularies share approximately 80% of their words, while Unigram vocabularies share at most about 47% of their words with other
tokenization algorithms. The ratio of common words is 38% in 8 K-sized vocabularies, but it decreases to 33% in 16 K and 29% in 32 K sized vocabularies.

3.2 | Identifying chemical key words

Our proposed pipeline was applied to identify chemical key words for protein families in the BDB dataset, as well as proteins in the LitPCBA and ProtBENCH datasets. To identify chemical key words, we ran the pipeline on each dataset using the three different tokenization algorithms with three distinct vocabulary sizes. We then evaluated the similarity of the identified key chemical words across different vocabularies by calculating the mean edit distances between these words for each target protein or protein family. Figure 5 shows the distribution of the mean edit distance between 8 K-sized vocabularies. Although the words selected by different vocabularies were not identical, they were quite similar, as indicated by the mean edit distance scores primarily ranging between 0 and 6, compared to a mean edit distance score of 6 for sets of words randomly selected from distinct vocabularies. For the majority of the targets, the mean edit distance was quite low, ranging from 2 to 4, depending on the compared vocabularies. This finding indicates that the three different tokenization algorithms using different sized vocabularies yielded similar chemical key words. Consequently, we used the BPE vocabulary with 8 K words for the remainder of our analysis.

The proposed method identified key words ranging from ~150 to ~3000, depending on the number of proteins or protein families in the dataset (see Table 1). The mean number of associations per identified chemical key word varies between ~1 and ~3, indicating that these words are indeed protein or family specific. However, due to the large vocabulary sizes, it was not feasible for us to manually assess all the chemical key words. Therefore, we compared the importance scores of these words
between strong and weak binders for each protein or family. Our analysis revealed that these scores were significantly different for at least half of the families or proteins across datasets, as shown in Figure 6.

In summary, our proposed pipeline successfully identified chemical keywords for protein families in multiple datasets. Despite variations in tokenization algorithms and vocabulary sizes, the identified words showed significant similarity. Additionally, the specificity of these chemical keywords to specific proteins or families was evident, highlighting the potential of our approach in understanding protein–ligand binding moieties.

3.3 | Case studies

We selected two protein families (carbonic anhydrases and casein kinase 1 gamma) from BDB dataset and two protein families (pyruvate kinase M2 and aldehyde dehydrogenase 1) from Lit-PCBA dataset as case studies and analyzed the chemical properties of the key chemical words and studied their impact on binding.

3.3.1 | Carbonic anhydrases

Carbonic anhydrase (CA) (EC. 4.2.1.1) is a metalloenzyme found in both eukaryotes and prokaryotes and is involved in many metabolic activities such as ion transport, pH balance, and gluconeogenesis [32]. Since the discovery that CA enzymes are involved in important
processes in the organism, they have become therapeutic targets for drug development. CA inhibitors are used clinically for diuretic, antiglaucoma, antiepileptic and antiobesity effects. New compounds with many anti-cancer effects that selectively inhibit CA isoenzymes have been designed and developed [33].

The algorithm proposed the following chemical words as CA pharmaphore or functional groups, in decreasing order of TF-IDF score: c1c(F)c(F), c(F)c1F, S(N)(−)=O, CC(=O)c1ccc, c(S, NS(=O)(−)=O)c1ccc, NS(−)=O(−)=O, c(Cl)c1, c1nc2ccc2cn1, and CSc1n. Here, the words S(N)(−)=O, NS(−)=O(−)=O, NS(−)=O(−)=O)c1ccc, and c1nc2ccc2cn1 are either components of chemical moieties or are already complete chemical moieties. These chemical words are recognized as sulfonamide, sulfonamide, aryl substituted sulfonamide, and 1H-benzo[d]-imidazole. We noted that the chemical words NS(−)=O(−)=O and S(N)(−)=O are both sulfonamides, although their SMILES representations are different.

The sulfonamide (RSO2NH2) group interacts with the Zn (II) ion in the active site of the CA enzyme and constitutes the main pharmaphoric group for CA inhibition. Sulfonamide derivatives bearing aromatic or heteroaromatic rings are potent inhibitors of the CA enzyme [33–34]. Systemic inhibitors such as acetazolamide (a) and dichlorphenamide (b) have long been used clinically as antiglaucoma agents [32]. Sulthiame (c) is a clinically used antiepileptic agent that shows inhibition of many CA isoenzymes in the brain [35]. The 2D structures of the clinically used CA enzyme inhibitors acetazolamide (a), dichlorphenamide (b) and sulthiame (c) are shown in Figure 7. Isoform selectivity of CA inhibitors is important to minimize side effects and increase therapeutic efficacy. For this purpose, researchers have designed a new generation of drugs that provide isoform selectivity. For example, CA II isozyme is an important target in the treatment of glaucoma, while CA IX and XII isoenzymes are mostly found in tumor cells. Second generation antiglaucoma drugs such as dorzolamide (d), and brinzolamide (e) are potent selective CA II inhibitors [36]. In addition, studies have shown that benzzenesulfonamide derivatives show strong inhibition of highly tumor-specific CA IX and XII isoenzymes [37]. These compounds contain sulfonamide and aryl-substituted sulfonamide structures, which were also identified by the algorithm as key chemical words for the CA enzyme family.

The key chemical word c1nc2ccc2cn1 found by the algorithm is notable among CA inhibitors with promising selectivity profiles. This is because CA enzyme inhibitors containing the benzimidazole structure represented by the word have been shown to be derivatives with high potential in the development of novel anticancer agents [38]. Figure 7 shows the 2D structure of a series of newly synthesized 2-substituted-benzimidazole-6-sulfonamides (f) with high potential against hCA IX and/or XII isoforms [38a].

The words c1c(F)c(F), c(F)c1F, c(Cl)c1 do not correspond to a full chemical moiety since ring closures are missing, but they are part of fluoride and chlorine substituted aromatic ring structures. In the literature, studies have investigated the thermodynamics, CA inhibitory effects, and binding affinities of benzenesulfonamide structures bearing fluoride or chloride [39]. A study by Dudutiene et al. showed that all fluorinated benzenesulfonamides had nanomolar binding potency against the tested CA enzymes. Furthermore, fluorinated benzenesulfonamides were found to have higher binding potency than non-fluorinated compounds [40]. Figure 7 shows the 2D structure of 2,3,5,6-tetrafluoro-4-phenoxypybenzenesulfonamide (g), a polyfluorinated benzene-sulfonamide derivative.

Similarly, CC(=O)c1ccc, which is not fully determined due to missing ring closures and is not a meaningful word for the CA family. Last c(S and CSc1n do not designate any chemical substructure.

### 3.3.2 Casein kinase 1 gamma

Casein kinase 1 (CK1) (EC. 2.7.11.1) isoforms are important therapeutic targets for diseases such as Alzheimer’s disease (AD), amyotrophic lateral sclerosis (ALS), and cancer [41]. Casein kinase 1 gamma (CK1y), one of the isoforms of CK1, plays a role in biological processes such as signal transduction, DNA repair, cell
division and circadian rhythm in cells [42]. CK1γ domain family refers to a domain found in the structure of proteins and associated with CK1γ. The C-terminal region may play a role in the interaction of the protein with target molecules. Drug studies focus on developing compounds or inhibitors that can bind to this region by targeting CK1γ [43]. Designing isoform-specific small molecules for members of the CK1 family is important for therapeutic strategies in human cancer. For this purpose, 2-phenylamino-6-cyano-1H-benzimidazole derivatives were synthesized as isoform-selective CK1 inhibitors by Hua et al. [44]. Activity assays showed that the compound (h) carrying a trifluoromethyl group on the pyridine ring shown in Figure 8 is the most potent CK1γ inhibitor. More importantly, the compound demonstrated high selectivity over other CK1 isoforms, such as CK1 alpha and delta. The pyridyl group is involved in an “edge-to-face” hydrophobic interaction with Pro331 [44].

3.3.3 | Pyruvate kinase M2

Pyruvate Kinase M2 (PKM2) (EC. 2.7.1.40) is an enzyme that catalyses pyruvate production and is involved in energy metabolism and biosynthesis. It also functions as a protein kinase that contributes to tumour formation. It is overexpressed in cancerous cells, and pyruvate kinase activity has been proposed as a useful approach for disrupting the required metabolic pathways for cancer cell proliferation [45]. The algorithm indicated the following as key chemical words (in decreasing order of TF–IDF score): ccc12), S(=O)(=O)N2CCN(, C2=NN(N, S(=O)(=O)-N, C(=O)N(C), c2nnnn2, SC, c(SCC(=O)N, NS(=O)(=O), and c5c4. S(=O)(=O)N and NS(=O)(=O) correspond to the sulfonamide group, which is present in 159 molecules out of the 546 PKM2 inhibitors in the Lit-PCBA dataset. In addition, S(=O)(=O)N2CCN isl a heterocyclic molecule containing an N atom to which the sulfonamide group is linked and C(=O)N(C) represents the amid group. Figure 9 shows several compounds that have been documented in the literature to target the PKM2 enzyme. The PDB codes of these molecules are N-[4-(4-(pyrazin-2-yl)piperazin-1-yl)carbonylphenyl]-quinoline-8-sulfonamide (i) [46], 3-[4-(2,3-dihydro-1,4-benzodioxin-6-ylsulfonyl)-1,4-diazepan-1-yl]-sulfonylaniline (j) [47], 1-(2,3-dihydro-1,4-benzodioxin-6-ylsulfonyl)-4-[(4-methoxyphenyl)sulfonyl]piperazine (k) [48] respectively. One of the oxygen atoms of the sulfonamide group makes a H-bond with the backbone oxygen of Tyr390. Moreover, the oxygen of the amide moiety forms an H-bond with side-chain nitrogen of Lys311 [46,49].

3.3.4 | Aldehyde dehydrogenase 1

Human aldehyde dehydrogenases (ALDHs) (EC.1.2.1.3) are a family of 19 enzymes that oxidise endogenous and exogenous aldehydes to less reactive, more soluble carboxylic acids in a nicotinamide adenine dinucleotide phosphate (NAD(P))+- dependent reaction [50]. ALDH family is involved in various biological processes required for cell survival as well as cell protection such as detoxification of toxic aldehydes, protection from oxidative stress, and regulation of retinoic acids metabolism [51]. Overexpression of ALDH isoforms is associated with disease progression and the development of resistance in many cancers such as breast and lung cancer [52]. It has been confirmed that ALDHs overexpressed in cancer stem cells (CSCs) promote tumour growth,
metastasis, therapeutic resistance, and immune escape. Therefore, ALDHs stand out as biomarkers for CSCs in several cancers. Consequently, targeting these enzymes may represent a new potential strategy to overcome therapeutic resistance in cancer patients. The ALDH1 isoenzyme consists of 6 members, namely ALDH 1 A1, 1 A2, 1 A3, 1B1, 1 L1, and 1 L2. The isoenzymes in this family are used as potential diagnostic markers in various cancers as well as in non-tumour diseases such as alzheimer, parkinson, cataract formation, and diabetes [53].

In our literature research on the ALDH1 enzyme family, NCT-501 molecule, which shows in vitro activity against cisplatin-resistant Cal-27 as well as head and neck squamous cell carcinoma cell lines and reduced tumor growth in vivo, attracted our attention [54]. However, its bioavailability is limited due to hepatic metabolism [55]. To improve the bioavailability of NCT-501 molecule, we analyzed the chemical words suggested by the algorithm. The NC(=S)N chemical word, which represents the thiourea group, caught our attention due to its known anti-cancer activity [56]. After replacing the urea group with the thiourea group, we found that the oral absorption percentage of the molecule increased from 72% to 82%. Additionally, the number of hydrogen-bond acceptors decreased from 10 to 9.5, which improved the physicochemical parameters of the molecule in terms of Lipinski’s rule of 5 (Table 2) [57]. CM026 is a competitive ALDH1 A1 inhibitor. Its ineffectiveness when used alone in GROV1 and A2780DK ovarian cancer cell lines is probably due to a low partition coefficient (logP) leading to low cell permeability [58]. The replacement of the urea group with the thiourea group in CM026 increased the oral absorption percentage from 79.551% to 90.170% and increased the logP value from 2.134 to 2.954 (Table 3). We used the QikProp v5.4 software package included in the Schrödinger package (Maestro11.9.011, Schrödinger, LLC, New York, NY, 2019–1) to calculate the physicochemical properties of the studied molecules in Tables 2 and 3.

4 | DISCUSSION

We developed a novel approach utilizing data-driven SMILES tokenization and tf-idf weighting to select key chemical words associated with protein-ligand binding. Our approach identified chemical words that are specific

### TABLE 2 Physicochemical properties of NCT501.

|                      | NCT-501  | NCT-501-NC(=S)N |
|----------------------|----------|-----------------|
| Structural formula   | ![Chemical Structure](image) | ![Chemical Structure](image) |
| Molecular weight (≤ 500 Da) | 416.522 | 432.583 |
| LogP octanol/water (≤ 5) | 1.363 | 2.159 |
| Number of hydrogen bond donor (≤ 5) | 0 | 0 |
| Number of hydrogen bond acceptor (≤ 10) | 10 | 9.5 |
| Human oral absorption % (>80 high, <25 weak activity) | 72.846 | 82.307 |

### TABLE 3 Physicochemical properties of CM026.

|                      | CM026  | CM026-NC(=S)N |
|----------------------|--------|---------------|
| Structural formula   | ![Chemical Structure](image) | ![Chemical Structure](image) |
| Molecular weight (≤ 500 Da) | 420.511 | 436.571 |
| LogP octanol/water (≤ 5) | 2.134 | 2.954 |
| Number of hydrogen bond donor (≤ 5) | 0 | 0 |
| Number of hydrogen bond acceptor (≤ 10) | 9.5 | 9 |
| Human oral absorption % (>80 high, <25 weak activity) | 79.551 | 90.170 |
to targets and match known pharmacophores and functional groups, as shown by several case studies. However, the proposed method also has certain limitations. Firstly, the method relies on the use of SMILES notation as the primary representation system for chemicals. While SMILES is widely used and offers readable character strings to represent chemical structures, it has inherent limitations. One significant drawback is its inability to capture and process extensive information regarding the spatial arrangement, bond orientation, and overall 3D structure of molecules. Furthermore, during the extraction of language units or tokens, the chemical word identification algorithm struggles to render complete SMILES strings for ring systems, resulting in incomplete words like c2ccc(=c1c1n. However, these incomplete SMILES strings for ring systems, resulting in incomplete words like c2ccc(=c1c1n. However, these incomplete words still contained valuable chemical information. For example, the word c1ccc2[nH]cnc2c1 represents the 1H-benzimidazole ring structure. If we were given the SMILES string c1ccc2[nH]cnc2c1, we could infer that it contains a 1H-benzimidazole ring based on the presence of c2. However, we would not be able to infer that it contains a 1H-benzimidazole ring structure because the ring starting with c1 is not completely closed. To interpret such words on a fragment-based level, we need to have some knowledge of chemistry, such as how rings can be formed and how many bonds an atom can make. Additionally, there are some words for which no interpretation can be made. For instance, the word CC[C@]34 C) seems to indicate isomeric structure, ring, and aliphatic structure, but it is not possible to draw any meaningful chemical inference based on only three carbons.

Another limitation of the proposed pipeline is that it relies only on strong binders to identify chemical structures important for binding. The presence of common chemical words among the strong binders does not necessarily imply a role in binding. The proposed approach has been shown to be effective, as significant differences in the importance scores of chemical words in strong binders compared to weak binders were observed for different proteins or families. However, a method that also considers the weak binders to identify key chemical words for binding would provide a more holistic approach.

5 | CONCLUSIONS

In this work, we investigated the chemical vocabularies generated by three subword tokenization algorithms (BPE, WordPiece, and Unigram) and proposed a novel language-inspired pipeline to select a subset of chemical key words that are strongly associated with protein-ligand binding. The chemical key words are found to be similar across different tokenization algorithms and are specific to each protein or family, often associated with only one protein/family and significantly different from the key words for weak binders. A detailed study of the literature shows that the chemical key words can designate pharmacophores and/or functional groups for all studied protein families and we also obtained important results that can improve the ADME profiles of the compounds. We view our study as a step towards interpreting the chemistry awareness of chemical words from the perspective of pharmaceutical chemistry and elucidating machine learning models for computational drug discovery, at large.

The success of the proposed pipeline in identifying key chemical substructures encourages prospective studies to discover new pharmacophore groups with machine learning. The proposed algorithm can leverage already available large protein–ligand interaction datasets to annotate more chemicals and propose novel pharmacophore and/or functional groups for protein targets. Such novel groups can guide the discovery of novel drugs and drug-like molecules.

ACKNOWLEDGMENTS

This work was supported by the Scientific and Technological Research Council of Turkey [Grant Number 119E133].

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

The LIT-PCBA, BDB, and ProtBENCH datasets are publicly available at the following URLs: LIT-PCBA dataset: https://drugdesign.unistra.fr/LIT-PCBA/, BDB dataset: https://github.com/boun-tabi/chemboost, ProtBENCH dataset: https://github.com/HUBioDataLab/ProtBENCH. We have made the source code publicly available at https://github.com/boun-tabi/exploring-chemical-words. The associated materials (i.e., data and outputs) are archived on Zenodo at https://doi.org/10.5281/zenodo.10245264.

ORCID

Asu Busra Temizer http://orcid.org/0000-0003-1811-093X
Gökçe Uludoğan http://orcid.org/0000-0002-8684-2457
Rıza Özçelik http://orcid.org/0000-0002-4052-9721
Taha Koulanı http://orcid.org/0000-0001-7159-6006
Elif Ozkırımı http://orcid.org/0000-0002-3206-8427
Kutlu O. Ulgen http://orcid.org/0000-0003-3668-3467
Nilgun Karali http://orcid.org/0000-0002-6916-122X
REFERENCES

1. P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, I. Sutskever, arXiv preprint arXiv:2005.00341, 2020.
2. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, arXiv preprint arXiv:2204.02311, 2022.
3. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, arXiv preprint arXiv:2204.06125, 2022, 1, 3.
4. a) D. Weininger, J. Chem. Inf. Comput. Sci. 1988, 28, 31–36; b) M. Krenn, F. Hase, A. K. Nigam, P. Friederich, A. Aspuru-Guzik, Mach. Learn.: Sci. Technol. 2020, 1(4), 045024; c) N. O’Boyle, A. Dalke, chemrxiv 10.26434/chemrxv.7097960.v1,2018.
5. D. Flam-Shepherd, K. Zhu, A. Aspuru-Guzik, Nat. Commun. 2022, 13(1), 3293.
6. H. Öztürk, A. Özgür, P. Schwaller, T. Laino, E. Ozkirimli, Drug Discov. Today 2020, 25, 689–705.
7. a) M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, ACS Cent. Sci. 2018, 4, 120–131; b) D. Grechishnikova, Sci. Rep. 2021, 11, 321; c) M. Moret, F. Grisoni, P. Katzberger, G. Schneider, J. Chem. Inf. Model. 2022, 62, 1199–1206; d) G. UludogAN, E. Ozkirimli, K. O. Ulgen, N. Karali, A. Ozgur, Bioinformatics 2022, 38(Supplement_2):ii155–ii161.
8. a) M. Tsubaki, K. Tomii, J. Sese, Bioinformatics 2019, 35, 309–318; b) M. Karimi, D. Wu, Z. Wang, Y. Shen, Bioinformatics 2019, 35, 3329–3338; c) I. Lee, H. Nam, J. Chem. Inform. Model. 2022, 14(1):1–15; d) V. Gligorijevic, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis, Nat. Commun. 2021, 12(1), 1–14.
9. a) B. Shin, S. Park, K. Kang, J. J. Cho, in Machine Learning for Healthcare Conference, PMLR, 2019, pp. 230–248; b) K. Abbasi, P. Razzaghi, A. Poso, M. Amanlou, J. B. Ghasemi, A. Masoudi-Nejad, Bioinformatics 2020, 36, 4633–4642; c) R. Özçelik, H. Öztürk, A. Özgür, E. Ozkirimli, Mol. Inf. 2021, 40, 2000212; d) Q. Zhao, G. Duan, M. Yang, Z. Cheng, Y. Li, J. Wang, IEEE/ACM Trans. Comput. Biol. Inform. 2022, 20, 852–863.
10. a) K. Huang, C. Xiao, L. M. Glass, J. Sun, Bioinformatics 2021, 37, 830–836; b) H. Öztürk, E. Ozkirimli, A. Özgür, Bioinformatics 2018, 34, 1295–1303; c) X. Li, D. Fourches, J. Chem. Inf. Model. 2021, 61(4), 1560–1569.
11. R. Sennrich, B. Haddow, A. Birch, arXiv preprint arXiv:1508.07909, 2015.
12. S. Chitrananda, G. Grand, B. Ramsundar, arXiv preprint arXiv:2010.09885, 2020.
13. A. E. Blanchard, D. Bhownik, Z. Fox, J. Gounley, J. Glaser, B. S. Akpa, S. Irle, J. Chem. 2023, 15, 1–12.
14. Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, arXiv preprint arXiv:1609.08144, 2016.
15. T. Kudo, arXiv preprint arXiv:1804.10959, 2018.
16. H. Abdel-Aty, I. R. Gould, J. Chem. Inf. Model. 2022, 62, 4852–4862.
17. E. J. Gardiner, V. J. Gillet, J. Chem. Inf. Model. 2015, 55, 1781–1803.
18. H. Du, Y. Cai, H. Yang, Z. Zhang, Y. Xue, G. Liu, Y. Tang, W. Li, Chem. Res. Toxicol. 2017, 30, 1209–1218.
42. L. Zhai, P. R. Graves, L. C. Robinson, M. Italiano, M. R. Culbertson, J. Rowles, M. H. Cobb, A. A. DePaoli-Roach, P. J. Roach, *J. Biol. Chem.* 1995, 270, 12717–12724.

43. A. Venerando, O. Marin, G. Cozza, V. H. Bustos, S. Sarno, L. A. Pinna, *Cell. Mol. Life Sci.* 2010, 67, 1105–1118.

44. Z. Hua, X. Huang, H. Bregman, N. Chakka, E. F. DiMauro, E. M. Doherty, J. Goldstein, H. Gunaydin, H. Huang, S. Mercede, J. Newcomb, V. F. Patel, S. M. Turci, J. Yan, C. Wilson, M. W. Martin, *Bioorg. Med. Chem. Lett.* 2012, 22, 5392–5395.

45. G. Dong, Q. Mao, W. Xia, Y. Xu, J. Wang, L. Xu, F. Jiang, *Oncol. Lett.* 2016, 11, 1980–1986.

46. C. Kung, J. Hixon, S. Choe, K. Marks, S. Gross, E. Murphy, B. DeLaBarre, G. Cianchetta, S. Sethumadhavan, X. Wang, S. Yan, Y. Gao, C. Fang, W. Wei, F. Jiang, S. Wang, K. Qian, J. Saunders, E. Driggers, H. K. Woo, K. Kunii, S. Murray, H. Yang, K. Yen, W. Liu, L. C. Cantley, M. G. Vander Heiden, S. M. Su, S. Jin, F. G. Salituro, L. Dang, *Chem. Biol.* 2012, 19, 1187–1198.

47. D. Anastasiou, Y. Yu, W. J. Israelsen, J. K. Jiang, M. B. Boxer, B. S. Hong, W. Tempel, S. Dimov, M. Shen, A. Jha, H. Yang, K. R. Mattaini, C. M. Metallo, B. P. Fiske, K. D. Courtney, S. Malstrom, T. M. Khan, C. Kung, A. P. Skoumbourdis, H. Veith, N. Southall, M. J. Walsh, K. R. Brimacombe, W. Leister, S. Y. Lunt, Z. R. Johnson, K. E. Yen, K. Kunii, S. M. Davidson, H. R. Christofk, C. P. Austin, J. Inglese, M. H. Harris, J. M. Asara, G. Stephanopoulos, F. G. Salituro, S. Jin, L. Dang, D. S. Auld, H. W. Park, L. C. Cantley, C. J. Thomas, M. G. Vander Heiden, *Nat. Chem. Biol.* 2012, 8, 839–847.

48. M. B. Boxer, J. K. Jiang, M. G. Vander Heiden, M. Shen, A. P. Skoumbourdis, N. Southall, H. Veith, W. Leister, C. P. Austin, H. W. Park, J. Inglese, L. C. Cantley, D. S. Auld, C. J. Thomas, *J. Med. Chem.* 2010, 53, 1048–1055.

49. B. DeLaBarre, J. Hurov, G. Cianchetta, S. Murray, L. Dang, *Chem. Biol.* 2014, 21, 1143–1161.

50. R. Januchowski, K. Wojtowicz, M. Zabel, *Biomed. Pharmacother.* 2013, 67, 669–680.

51. S. Singh, C. Brocker, V. Koppaka, Y. Chen, B. C. Jackson, A. Matsumoto, D. C. Thompson, V. Vasiliiou, *Free Radical Biol. Med.* 2013, 56, 89–101.

52. S. S. Dinavahi, C. G. Bazewicz, R. Gowda, G. P. Robertson, *Trends Pharmacol. Sci.* 2019, 40, 774–789.

53. J.-J. Duan, J. Cai, L. Gao, S.-C. Yu, *J. Enzyme Inhib. Med. Chem.* 2023, 38, 2166035.

54. S.-M. Yang, A. Yasgar, B. Miller, M. Lal-Nag, K. Brimacombe, X. Hu, H. Sun, A. Wang, X. Xu, K. Nguyen, *J. Med. Chem.* 2015, 58, 5967–5978.

55. S. Kulsum, H. V. Sudheendra, R. Pandian, D. R. Ravindra, G. Siddappa, R. Nisheena, P. Chevour, B. Ramachandran, M. Saggar, A. Jayaprakash, A. Mehta, V. Kekatpure, N. Hedne, M. A. Kurikase, A. Suresh, *Mol. Carcinog.* 2017, 56, 694–711.

56. a) V. Asati, A. Anant, P. Patel, K. Kaur, G. D. Gupta, *Eur. J. Med. Chem.* 2021, 225, 113781; b) I. H. Eissa, A. M. El-Naggar, M. A. El-Hashash, *Bioorg. Chem.* 2016, 67, 43–56; c) A. Shakeel, A. A. Altaf, A. M. Qureshi, A. Badshah, *J. Drug Des. Med. Chem.* 2016, 2, 10.

57. C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* 2012, 64, 4–17.

58. C. A. Morgan, T. D. Hurley, *J. Med. Chem.* 2015, 58, 1964–1975.

How to cite this article: A. B. Temizer, G. Uludogan, R. Ozcelik, T. Koulanli, E. Ozkirimli, K. O. Ulgen, N. Karali, A. Ozgur, *Molecular Informatics* 2024, 43, e202300249. https://doi.org/10.1002/minf.202300249