Substructure search optimizations

June 4, 2023

1 Introduction

Substructure search in chemical compound databases is a vital task in cheminformatics, underpinning broad applications in drug discovery, materials science, and toxicology. The objective is to identify all molecules in a database that contain a given query substructure, which typically corresponds to a specific chemical motif or functional group. This search has been a cornerstone in understanding the influence of specific substructures on a compound's biological activity, physicochemical properties, and reactivity, a recognized concept for decades [Barnard, 1993].

Historically, computer-based substructure search started with pioneers like Ledley and colleagues who developed the Chemical Substructure Search (CSS) algorithm in the 1960s and 1970s [Ledley et al., 1964]. This algorithm employed a graph-based approach to identify specific substructures in chemical compounds. Further advancements in the field came with the development of the Simplified Molecular Input Line Entry System (SMILES) notation by Weininger in the 1980s [Weininger, 1988], which provided a simple, linear representation of molecular structures as strings.

Substructure search fundamentally rests on the solution of the subgraph isomorphism problem, a problem known to be NP-complete [Ullmann, 1976]. Due to its high computational complexity, numerous heuristics and algorithms have been devised to accelerate the search process. Among these, the Filter-and-Verification paradigm has been a prevalent approach, involving an initial filtering step to quickly eliminate unsuitable candidate graphs, and a more computationally intensive verification step to confirm the presence of the query substructure in the remaining candidates [Cordella et al., 2004, Shasha et al., 2002]. Over time, graph-based subgraph isomorphism algorithms, such as the Ullmann algorithm [Ullmann, 1976] and the VF2 algorithm [Cordella et al., 2004], have emerged as more efficient and scalable solutions for substructure search in large chemical databases.

In addition to these, frequent subgraph mining algorithms like gSpan [Yan and Han, 2002], FFSM [Kuramochi and Karypis, 2001], and Gaston [Nijssen and Kok, 2004] have proven valuable in identifying frequently occurring substructures in large sets of chemical compounds. These approaches are particularly beneficial for ap-

plications such as structure-activity relationship (SAR) analysis and molecular classification.

Efficient filtering techniques often involve the use of binary and quantitative features, or fingerprints, to represent molecular structures. These fingerprints facilitate the rapid elimination of graphs that do not contain specific features required by the query subgraph, thereby speeding up the substructure search process [Bonchi et al., 2011, Klein et al., 2014].

However, as the number of known molecules and the size of chemical databases have grown significantly, the traditional approaches, which often require full or nearly full enumeration of candidates, have become increasingly challenging to implement efficiently. This development underlines the need for more scalable solutions.

Информация про экспоненциальный рост взята из обсуждений с коллегами проекта. Однако, я не могу сослаться на авторитетный источник, подтверждающий экспоненциальный рост. Поэтому предлагаю написать более аккуратно. Например, сказать, что у компании есть датасет размера 4 миллиарда, и что текущие подходы не позволяют эффективно поддерживать такую большую базу данных. Поэтому и есть потребность в поиске нового алгоритма. (А где-то в итогах можно упомянуть, что наш подход в текущей версии всё ещё не решает задачу для 4 миллиардов)

Я заменил exponentially на significantly. В принципе, можно найти как менялись размеры PubChem, Chembl, ZINC, ChemSpider в зависимости от версий. Писать ли про 4 млрд - я не знаю.

Our work introduces a unique approach aimed at mitigating these challenges. While in certain cases the algorithm may resort to exhaustive enumeration, in most scenarios it employs a more sophisticated strategy, transcending the conventional full enumeration paradigm. в худшем случае наше решение выполняет полный перебор, поэтому не уверен, можно ли тут сказать, что мы вышли за его рамки переформулировал Instead, we introduce a unique index structure: a tree that segments the molecular dataset into clusters based on the presence or absence of features. Inspired by the binary Ball-Tree concept [Omohundro, 1989, Clarkson, 1994], this structure demonstrates superior performance over exhaustive search on average, leading to a significant acceleration in the filtering process.

We provide a comparative analysis of our method with Sachem [Kratochvíl et al., 2018], a known fast search method, highlighting key differences. While Sachem uses advanced filtering effectively, it relies on exhaustive search in a chemical space that continually expands. In contrast, our approach departs from exhaustive search and places an existing molecular fingerprint into a tree structure, rather than a conventional relational database.While our current version does not impact the verification stage, it holds promise for reducing the false positive rate. Но ведь текущая версия абсолютно не влияет на verification stage. Поэтому у нас абсолютно нет улучшений части verification в предыдущем абзаце мне было бы непонятно, что уже реализовано, а что только появиться быть в нашей будущей версии. я немного переформулировал... By introducing this innovative structure, we aim to cater to the growing scale of chemical

databases and the escalating demand for efficient and scalable search solutions. Our approach offers potential for future research and application in the quest for more efficient and accurate substructure search techniques.

2 Algorithm description

2.1 Notation and main idea

The objective of our research is to facilitate the identification of specific substructures within the molecules from a database $\mathcal{M}$. For this, we utilize the concept of a "fingerprint", a binary string of constant length fl, corresponding to each molecule. To perform this mapping, we define a function $\mathsf{fp} : \mathcal{M} \to \mathcal{F}$ that takes a molecule from the set $\mathcal{M}$ and produces its corresponding fingerprint in the set $\mathcal{F}$. На мой взгляд написано слишком кратко про фингерпринты. Кажется, если человек не в теме, то он может не понять, что это за сопоставление молекулам бинарных строк, зачем оно нужно и почему оно эффективно реализует этап filter

To make the substructure search process more efficient, we propose organizing these fingerprints in a binary search tree, denoted as $\mathbb{T}$. The tree is binary and complete, having a specific depth d.

In this tree, the root, left, and right subtrees of a node $\mathtt{v}$ are represented as $\mathbb{T}.\mathtt{root}$, $\mathtt{v}.\mathtt{left}$, and $\mathtt{v}.\mathtt{right}$, respectively. Each node also has a set of all leaves in its subtree, denoted as $\mathtt{v}.\mathtt{leaves}$. Each leaf ℓ in the tree $\mathbb{T}$ holds a set $\ell.\mathtt{set}$ of fingerprints. A unique concept to our approach is the centroid, $\mathtt{v}.\mathtt{centroid}$, recorded at each node v. The centroid is defined as a fingerprint F for which $F[i] = 1$ if and only if there exists another fingerprint F' in the subtree of $\mathtt{v}$ such that $F'[i] = 1$. This is represented as

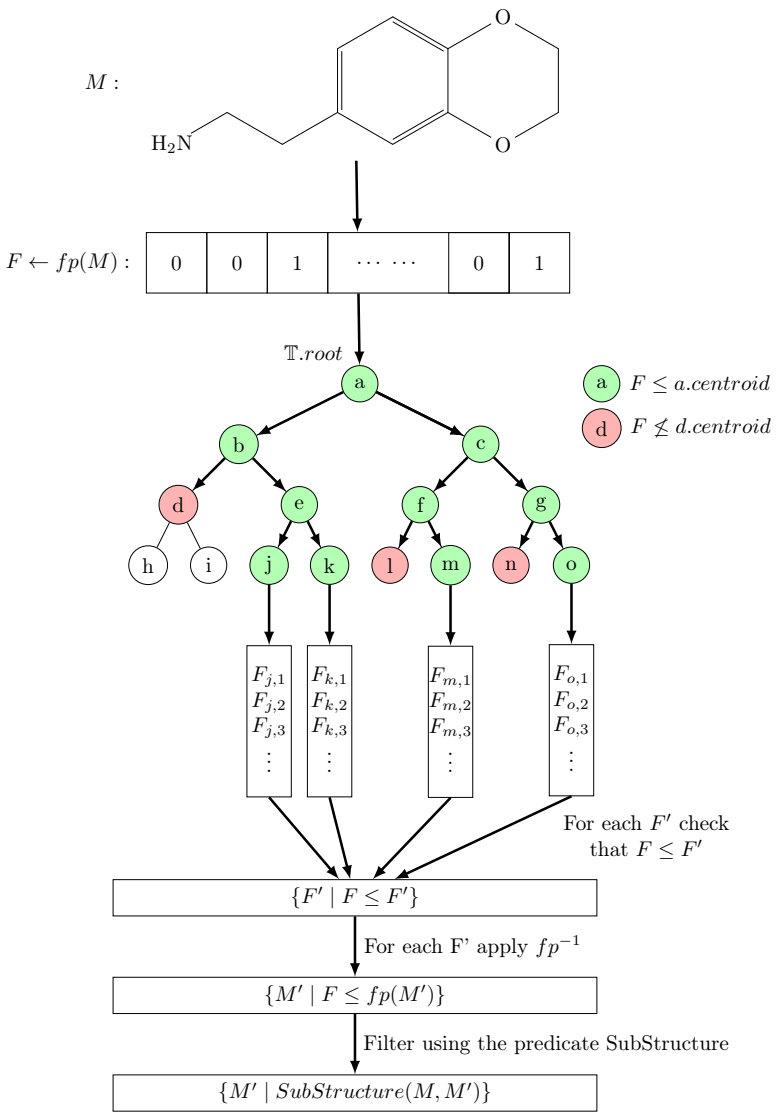
$$\mathtt{v}.\mathtt{centroid} = \bigvee_{\ell \in \mathtt{v}.\mathtt{leaves}} \bigvee_{F \in \ell.\mathtt{set}} F.$$

This concept of the centroid is inspired by BallTree literature.

Our search process is designed to locate all fingerprints F' in the set $\mathcal{F}$ where F is a submask of F'. This search relies on the relation $F_1 \leq F_2$ for fingerprints F_1, F_2 that holds true if and only if for every $i \in 1, 2, \ldots, \mathsf{fl}$, $F_1[i] \leq F_2[i]$.

The search starts from the root and recursively descends into both subtrees. Note that here we can improve the performance by parallelizing this step to explore both subtrees simultaneously. We stop the recursive descent if we reach a node $\mathtt{v}$ where $F \not\leq \mathtt{v}.\mathtt{centroid}$. Conversely, if we reach a leaf ℓ and $F \leq \ell.\mathtt{centroid}$, we add to $\mathcal{F}_M$ the set $\{F' \in \ell.\mathtt{set} \mid \mathsf{fp}(M) \leq F'\}$.

Following the generation of $\mathcal{F}_M$, the next phase involves examining each M' in $\bigcup_{F' \in \mathcal{F}_M} \mathsf{fp}^{-1}(F')$. The objective is to determine whether each M' is a substructure of M. This determination is made by employing external algorithms to verify the predicate $\mathtt{SubStructure}(M', M)$, which is true if and only if M' is a substructure of M.


M :
O
H_2N
O
$F \leftarrow fp(M)$:
0
0
1
⋯ ⋯
0
1
$\mathbb{T}.root$
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
a $F \leq a.centroid$
d $F \not\leq d.centroid$
$F_{j,1}$ $F_{j,2}$ $F_{j,3}$ ⋮
$F_{k,1}$ $F_{k,2}$ $F_{k,3}$ ⋮
$F_{m,1}$ $F_{m,2}$ $F_{m,3}$ ⋮
$F_{o,1}$ $F_{o,2}$ $F_{o,3}$ ⋮
For each F' check that $F \leq F'$
$\{F' \mid F \leq F'\}$
For each F' apply fp^{-1}
$\{M' \mid F \leq fp(M')\}$
Filter using the predicate SubStructure
$\{M' \mid SubStructure(M, M')\}$

Further details on the BallTree and the utilization of the tree in the substructure search process will be provided in the subsequent sections.

The pseudocode for the fingerprint search function in the tree is described in Algorithm 1. The pseudocode for the function that searches for superstructures of a given molecule is described in Algorithm 2.

Algorithm 1 Searching for all matching fingerprints in a subtree

Require: $\mathtt{v}$ is a tree vertex, F is a fingerprint
Ensure: $\{F' \in \bigcup_{\ell \in \mathtt{v.leaves}} \ell.\mathtt{set} \mid F \leq F'\}$

procedure FindInSubtree($\mathtt{v}, F$)
  if $F \not\leq \mathtt{v.centroid}$ **then**
    return $\varnothing$
  else if $\mathtt{v}$ is leaf **then**
    return $\{F' \in \mathtt{v.set} \mid F \leq F'\}$
  else
    $\mathtt{left} \leftarrow$ FindInSubtree($\mathtt{v.left}, F$)
    $\mathtt{right} \leftarrow$ FindInSubtree($\mathtt{v.right}, F$)
    return Concatenate($\mathtt{left}, \mathtt{right}$)
  end if
end procedure

Algorithm 2 Searching for all superstructures of a given molecule

Require: M is a molecule
Ensure: $\{M' \in \mathcal{M} \mid \mathtt{SubStructure}(M, M')\}$

procedure FindMetaStructures(M)
  $F \leftarrow \mathsf{fp}(M)$
  $F_M \leftarrow$ FindInSubtree($\mathbb{T}.\mathtt{root}, F$)
  return $\{M' \in \bigcup_{F' \in F_M} \mathsf{fp}^{-1}(F') \mid \text{SubStructure}(M, M')\}$
end procedure

2.2 Building the tree

To start, let's create a trivial tree with a single node, denoted as $\mathbb{T}.\mathtt{root}$. Assign $\mathbb{T}.\mathtt{root}.\mathtt{set} = \mathcal{F}$. Next, we will inductively split the leaves of the tree into two parts, thereby adding new nodes to the tree.

More formally, for each leaf node ℓ of the tree, we will divide $\ell.\mathtt{set}$ using a specific function called SplitFingerprints: $\mathcal{F}_l, \mathcal{F}_r \leftarrow \text{SplitFingerprints}(\ell.\mathtt{set})$ ($\mathcal{F}_l \sqcup \mathcal{F}_r = \ell.\mathtt{set}$). Next, we will recursively build trees for $\ell.\mathtt{left}, \ell.\mathtt{right}$ using the sets $\mathcal{F}_l, \mathcal{F}_r$.

We will continue splitting the leaves in this manner until $\mathbb{T}$ becomes a full binary tree with depth d. The pseudocode for the algorithm described above can be found in 3.

Algorithm 3 Building the tree

Require: $\mathcal{F}$ is the set of all fingerprints, d is the depth of the tree
Ensure: $\mathbb{T}$ is the BallTree for the superstructure fingerprint search

procedure BUILDTREE($\mathcal{F}, d$)
  $\mathtt{v} \leftarrow$ new node
  if $d = 1$ **then**
    $\mathtt{v.set} \leftarrow \mathcal{F}$
    $\mathtt{v.centroid} \leftarrow \bigvee_{F \in \mathcal{F}} F$
    return v
  else
    $\mathcal{F}_l, \mathcal{F}_r \leftarrow$ SPLITFINGERPRINTS($\mathcal{F}$)
    $\mathtt{v.left} \leftarrow$ BUILDTREE($\mathcal{F}_l, d-1$)
    $\mathtt{v.right} \leftarrow$ BUILDTREE($\mathcal{F}_r, d-1$)
    $\mathtt{v.centroid} \leftarrow \mathtt{v.left.centroid} \vee \mathtt{v.right.centroid}$
    return v
  end if
end procedure

Algorithm 4 Algorithm for splitting fingerprints in parts during tree construction

Require: set $\mathcal{F}$ of fingerprints to be split
Ensure: the split $\mathcal{F}_l, \mathcal{F}_r$ of the set $\mathcal{F}$

procedure SPLITFINGERPRINTS($\mathcal{F}$)
  $j \leftarrow \arg\min_i \{||\mathcal{F}| - 2k| \mid k = \#\{F \in \mathcal{F} \mid F_i = 1\}\}$
  $\mathcal{F}_l \leftarrow \{F \in \mathcal{F} \mid F[j] = 0\}$
  $\mathcal{F}_r \leftarrow \{F \in \mathcal{F} \mid F[j] = 1\}$
  if $|\mathcal{F}_l| > \lfloor \frac{n}{2} \rfloor$ **then**
    $\mathcal{F}_r \leftarrow \mathcal{F}_r \cup$ TAKELASTELEMENTS($\mathcal{F}_l, |\mathcal{F}_l| - \lfloor \frac{n}{2} \rfloor$)
    $\mathcal{F}_l \leftarrow$ DROPLASTELEMENTS($\mathcal{F}_l, |\mathcal{F}_l| - \lfloor \frac{n}{2} \rfloor$)
  else if $|\mathcal{F}_r| > \lceil \frac{n}{2} \rceil$ **then**
    $\mathcal{F}_l \leftarrow \mathcal{F}_l \cup$ TAKELASTELEMENTS($\mathcal{F}_r, |\mathcal{F}_r| - \lceil \frac{n}{2} \rceil$)
    $\mathcal{F}_r \leftarrow$ DROPLASTELEMENTS($\mathcal{F}_r, |\mathcal{F}_r| - \lceil \frac{n}{2} \rceil$)
  end if
  return $\mathcal{F}_l,\ \mathcal{F}_r$
end procedure

We want to perform the splits in such a way that, on average, the search often prunes branches during the traversal. That is, the **if** statement in line 2 of algorithm 1 should be executed frequently. Let's discuss the function SplitFingerprints in more detail.

Initially, one might consider selecting a specific bit j and assigning all fingerprints F such that $F[j] = 0$ to the left subtree, and those with $F[j] = 1$ to the right subtree. In this case, when searching for superstructures of the fingerprint F', if $F'[j] = 1$, the entire left subtree would be cropped. However, in practice, this approach leads to significant differences between the left and right parts after a few splits, making it difficult to create a deep and balanced tree. Unfortunately, a shallow or unbalanced tree does not offer substantial improvements over a full search, as it barely eliminates any search branches.

Therefore, we suggest the following method: we will still select the bit as mentioned above, but we will divide the fingerprints in a way that ensures the sizes of the resulting partitions match. For instance, if the optimal division of n fingerprints yields parts with sizes $n_0, n_1 (n_0 < n_1 \ \wedge \ n_0 + n_1 = n)$, then all values with zero will be assigned to the left partition, while the values with one will be distributed to achieve final left and right partition sizes of $\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil$ respectively. If $n_0 > n_1$, we will proceed symmetrically. The algorithm for the SplitFingerprints function can be found in the pseudocode 4.

3 Benchmarks

In this study, we have carried out comprehensive benchmarking to assess the performance of our algorithm, which is an extension of the Bingo fingerprinting system, in comparison with established indices, namely Bingo [Pavlov et al., 2010] and Sachem/Lucy [Kratochvíl et al., 2018]. В Sachem статье есть ещё сравнения Sachem/OrChem, Sachem/eCDK, RDKit, pgchem. Можно их тоже перенести в бенчмарки чтобы база сравнений была шире. Our benchmarking process was performed under the following conditions:

- OS: Ubuntu 22.04
- Processor: Intel Xeon E5-2686 v4 (Broadwell)
- Clock speed: 2.7 GHz
- RAM: 120 GB

The query dataset used for benchmarking was retrieved from `https://hg.sr.ht/~dalke/sqc/browse?rev=tip`, which contains 3488 relevant queries for substructure search. Ten queries were excluded due to various issues, resulting in a final set of 3478 compounds. Кажется, плохо так сильно обобщать проблемы. Мы выкинули те молекулы, на которых наш алгоритм работает заведомо плохо. Поэтому выкинув эти 10 молекул мы улучшили свои результаты. С другой стороны 10 штук почти не влияют на статистику, поэтому может и нет смысла описывать подробности.

We believe that this comparison is appropriate because the Sachem search in the referenced study was conducted in a similar manner. However, there are some notable differences between our testing conditions and those of the referenced studies. Our testing was performed on a processor with a higher clock speed (2.7 GHz versus 2.6 GHz), the database used in our tests was larger (113M compared to 94M), and our search was conducted for the first 10,000 results. Ещё одно отличие: Sachem выкинули 159 запросов из 3488.

Despite these differences, the benchmark results provide a meaningful comparison of the relative efficiencies of the three systems. For a single-threaded in-memory execution Надо либо добавить в сравнительную таблицу версию, работающую на жёстком диске, либо убрать приписку "in-memory", потому что без версии во внешней памяти она не имеет смысла. Однако, из-за возникших технических трудностей, на данный момент у меня нет качественно проделанных замеров версии с жёстким диском , our algorithm demonstrates competitive performance, and it also shows the potential for parallelization, exhibiting substantial improvements when executed on 16 threads in memory.

The table below summarizes these benchmark timings, providing a clear comparison between our algorithm, Bingo, and Sachem/Lucy. Bingo и Sachem тестировались тоже с искусственным ограничением на один поток. О чём сделал приписку в таблице

%	Our Algorithm, single-threaded (s) in-memory	Our Algorithm, 16-threaded (s) in-memory	Bingo, single-threaded (s)	Sachem/Lucy, single-threaded (s)
50%	2.17443	0.337058	1	-
60%	3.83995	0.525275	-	-
70%	4.87392	0.677609	-	-
80%	6.71327	0.895	10	-
85%	-	-	-	1
90%	12.9814	1.65519	-	-
95%	30.2751	3.75599	100	10
98%	-	-	-	10
100%	-	-	-	<80

This thorough analysis offers valuable insights into the performance and potential scalability of our algorithm, especially when it comes to parallel computing. мне не кажется, что способность к распараллеливанию является нашим преимуществом, так как другие подходы тоже умеют работать параллельно

4 Further Development

Fingerprints currently form the basis of our algorithm, but they do have certain limitations which don't make them the ideal fit for our tree-based approach.

Firstly, the fingerprint's condensed nature is aimed to ensure efficient computation, which often leads to grouping together several characteristics. For

instance, a single attribute in a fingerprint often encapsulates multiple individual elements because these isolated items, while lacking substantial filtering power across the entire dataset, might be relevant for specific subsets. However, the fingerprint structure doesn't account for such instances. Contrarily, our approach could accommodate more complex functions, even if they operate slower than traditional filtering methods—for example, using a fingerprint variant that doesn't amalgamate different elements.

Secondly, fingerprints are designed to provide a universal filter across the entire dataset. This results in a significantly pared-down set of attributes applicable to the entire database. For example, Bingo utilizes 2584 attributes, which intuitively seems insufficient to capture all peculiarities of a 113M-sized molecule dataset. Even a substantially enlarged fingerprint variant wouldn't be able to cover all exceptional cases. In contrast, our approach, by dealing with subsets, can extract a unique characteristic for a tree node relevant to the set in the given subtree, thus allowing for a much more effective coverage of the existing data nuances.

As a result, a potential enhancement of our algorithm might involve the use of a specific attribute in each tree node. Depending on its presence or absence, the search continues in both subtrees or only in the right subtree. This attribute would be chosen in advance to approximately bisect the set in the subtree. A leaf would contain several characteristics which would be examined when filtering elements from the leaf.

Employing the method described above, we could potentially improve the false positive rate, as the selected attributes would be relevant to the examined subsets. Moreover, these attributes could be utilized during verification, possibly resulting in substantial improvements in the verification stage speed, thanks to the relevance of these attributes to the molecule subsets.

5 Conclusion

The current version of our approach can serve as an extension to a fingerprint, enhancing the filtering speed by avoiding exhaustive enumeration. Moreover, the tree's ability to cluster molecules enables a more detailed examination of cluster-specific attributes, an aspect that existing algorithms struggle with, as they aim to find optimal ways to generalize across the entire dataset. Therefore, our approach could potentially be used in the future to improve both the false-positive rate and the verification speed.

References

[Barnard, 1993] Barnard, J. M. (1993). Substructure searching methods: Old and new. *Journal of Chemical Information and Computer Sciences*, 33(4):532–538.

[Bonchi et al., 2011] Bonchi, F., Perego, R., Silvestri, F., Vahabi, H., and Venturini, R. (2011). Exemplar queries: Give me an example of what you need. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2097–2100. ACM.

[Clarkson, 1994] Clarkson, K. L. (1994). Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, 2(2006):15–59.

[Cordella et al., 2004] Cordella, L. P., Foggia, P., Sansone, C., and Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1367–1372.

[Klein et al., 2014] Klein, K., Werth, M., and Kriege, N. M. (2014). Efficient subgraph mining in cheminformatics and elsewhere. *Informatik Spektrum*, 37(1):9–16.

[Kratochvíl et al., 2018] Kratochvíl, M., Vondrášek, J., and Galgonek, J. (2018). Sachem: a chemical cartridge for high-performance substructure search. *Journal of Cheminformatics*, 10(1):27.

[Kuramochi and Karypis, 2001] Kuramochi, M. and Karypis, G. (2001). Frequent subgraph discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 313–320. IEEE.

[Ledley et al., 1964] Ledley, R. S., Lusted, L. B., and Schulman, J. D. (1964). Computer-based medical decision making: The use of computers for the identification of chemical substructures. *Science*, 146(3647):1043–1045.

[Nijssen and Kok, 2004] Nijssen, S. and Kok, J. N. (2004). A quickstart in frequent structure mining can make a difference. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 647–652. ACM.

[Omohundro, 1989] Omohundro, S. M. (1989). Five balltree construction algorithms. Technical report, International Computer Science Institute.

[Pavlov et al., 2010] Pavlov, D., Rybalkin, M., and Karulin, B. (2010). Bingo from scitouch llc: chemistry cartridge for oracle database. *Journal of Cheminformatics*, 2(1):F1.

[Shasha et al., 2002] Shasha, D., Wang, J. T., and Giugno, R. (2002). Algorithmics and applications of tree and graph searching. In *Proceedings of the Twenty-First ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 39–52.

[Ullmann, 1976] Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)*, 23(1):31–42.

[Weininger, 1988] Weininger, D. (1988). Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

[Yan and Han, 2002] Yan, X. and Han, J. (2002). gspan: Graph-based substructure pattern mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724. IEEE.