1QBit

A Novel Graph-based Approach for Determining Molecular Similarity

Maritza Hernandez, Arman Zaribafiyan, Maliheh Aramon, Mohammad Naghibi

Abstract

In this paper, we tackle the problem of measuring similarity among graphs that represent real objects with noisy data. To account for noise, we relax the definition of similarity using the maximum weighted co-k-plex relaxation method, which allows dissimilarities among graphs up to a predetermined level. We then formulate the problem as a novel quadratic unconstrained binary optimization problem that can be solved by a quantum annealer. The context of our study is molecular similarity where the presence of noise might be due to regular errors in measuring molecular features. We develop a similarity measure and use it to predict the mutagenicity of a molecule. Our results indicate that the relaxed similarity measure, designed to accommodate the regular errors, yields a higher prediction accuracy than the measure that ignores the noise.

Keywords: graph similarity, molecular similarity, quadratic unconstrained binary optimization

1   Introduction

The accumulation of data in today's digital world is growing exponentially, affecting various fields, such as physics, decision sciences, astronomy, and biology. *The Economist* estimates that by the year 2020, the amount of data in the world will be seven times greater than the data that existed in 2014 [1]. The revolutionary impact of ``big data'' in biology has been tremendous, transforming it into data-rich field. The majority of the new data, including data sets of protein structures, DNA sequences, and gene expressions are complex, calling for the development of new analytical tools and algorithms to translate the data into valuable insights. The first challenge then lies in the efficient representation of biological data objects. Graphs have shown promising potential in representing data objects including DNA [2] and small molecules [3] due to their high abstractional capacity [4]. The problem of biological data analysis is then reduced to graph analysis. The goal is to search, visualize, and infer new correlations among objects to produce healthier foods, prevent diseases, and discover individually customized drugs. Therefore, meaningful metrics are required to determine similarity and dissimilarity among graphs.

Graph similarity methods are developed primarily based on the property of isomorphism in the graph theory literature [5]. A graph $G_1 = (V_1, E_1)$ is called isomorphic to a graph $G_2 = (V_2, E_2)$ if there is a bijection between the vertex sets V_1 and V_2 such that there exists a mapping between the adjacent pairs of vertices of G_1 and G_2. This is an ``edge-preserving'' bijection, consistent with the general notion of ``structure-preserving''. The main drawback of isomorphism in comparing real objects is its restrictive binary nature---it does not consider partial similarities among graphs. Two graphs are evaluated as being either exactly similar up to label permutation or not similar at all. Therefore, practical variations of isomorphism have been developed, including edit distance and maximum common subgraph (MCS) [6]. The latter has a structure-preserving feature similar to isomorphism. The MCS of two graphs G_1 and G_2 is the largest subgraph of G_1 that is isomorphic to a subgraph of G_2.

In real-world applications, particularly in biology, graphs of objects carry different types of information whose level of importance may also differ. To account for real-world properties, the MCS problem has been generalized to the labelled maximum weighted common subgraph (LMWCS) problem, where vertices and/or edges are associated with multiple labels, storing various information. For example, in a molecular similarity application, a molecular graph is built where vertices and edges usually represent atoms and atomic bonds, respectively. The vertices may have labels for atomic numbers and formal charges, and the labels of the edges may indicate the bond types. It has been shown that the MCS and LMWCS problems are equivalent to another well-known problem---the maximum independent set (MIS) of a third graph, which can be induced from the graphs being compared [7, 8]. The vertices and edges of the third graph, respectively, represent possible mappings and the conflicts between them. The goal of the MIS problem is to find the largest set of vertices such that there is no edge between all selected pairs, forming the largest conflict-free mapping.

Representing the MCS problem as the MIS problem makes it easier to further relax the definition of similarity where the data is noisy or distorted. In such contexts, the definition of similarity based on exact subgraph matching might be prohibitive, excluding useful information. For example, in three-dimensional graph representation of molecules, the position of each atom is obtained relative to the other atoms by solving an optimization problem heuristically. Therefore, the graph representation is partially accurate. A more flexible similarity measure can compensate for the effect of noise in the data, considering as similar substructures with conflicts up to a certain tolerable threshold. Even if

there is no perturbation in the data, relaxing the definition of similarity can be beneficial. For example, to compare a newly sequenced genome with a reference genome, relaxing the definition of similarity allows the regular variations that occur within populations to be accommodated. There are different relaxations of the exact subgraph matching requirement in the literature [9]. One of the relaxations is known as the maximum co-k-plex problem, where the goal is to find the largest set of vertices in the graph such that each vertex has at most $k-1$ edges connecting it to the other vertices. It is clear that the maximum co-1-plex problem is the MIS problem [8].

The majority of the similarity methods discussed above, including the MIS and maximum co-k-plex problems, are in general NP-hard, having exponentially increasing computational complexity due to the combinatorial nature of graphs [10,11]. However, since a given graph representation can be used to model a wide range of problems, including image processing and DNA sequencing, graph-based algorithms are practical. As a result, there have been efforts to find heuristics with lower complexity, develop approximation methods, restrict graph structures, or use parallel implementation strategies to solve graph similarity problems [12--14]. These methods, though designed to have polynomially bounded complexity, fail to find the optimal solutions. It is, furthermore, challenging to evaluate their performance, gaining insights on how much we lose by not using the exact algorithms in the worst case.

Recent advances in the ability to utilize quantum mechanical effects in computation have sparked interest in the research community that quantum computing could provide a speed-up in solving classical NP-hard optimization problems [15,16]. Quantum annealing is one approach to harnessing quantum effects in searching the energy landscapes of optimization problems. *D-Wave Systems* has developed a physical realization of a quantum annealer [17], which has been shown to exhibit quantum effects such as tunnelling [18] and entanglement [19]. This device finds the minimum value (i.e., ground state) for a special class of objective functions (i.e., Ising objective functions). The class of objective functions native to the hardware can be translated to quadratic functions of binary variables, and since the hardware cannot encode constraints, the device is a quadratic unconstrained binary optimization (QUBO) problem solver. As a result, there has been work in recent years to formulate various constrained optimization problems as QUBO problems [20--22].

One of the main contributions of this paper is to propose a novel QUBO problem formulation for the maximum weighted co-k-plex problem, measuring similarity among graphs. It is worth mentioning that our formulation can be solved by a quantum annealer; however, it is not our aim in this paper to study the annealer's performance. Our focus is to investigate the performance of the QUBO-based similarity measure in the context of a real molecular similarity problem. To accurately evaluate the performance of the QUBO-based measure over existing molecular similarity measures, we use an exhaustive solver to find the optimal similarity values among graphs.

The rest of the paper is organized as follows. In Section 2, we review the existing QUBO problem formulation for the MIS problem and present a novel formulation for the maximum weighted co-k-plex problem, measuring the similarity among graphs. In Section 3, we use the new model in the context of molecular similarity where the reduction of molecules to graphs and the adaptation of the novel QUBO-based formulation to molecular similarity are explained. Finally, in Section 4, we use a machine learning approach on real molecular data sets to evaluate the performance of our similarity measure in predicting mutagenicity.

2   A QUBO-based Graph Similarity Measure

The MIS problem has been formulated as a QUBO problem [23,24]; however, to our knowledge there is no QUBO problem formulation that considers the relaxation of the exact subgraph matching. In this section, we first review the existing formulation of the MIS problem. We then introduce our novel formulation for the maximum weighted co-k-plex problem and discuss its generalization to similarity problems among multiple graphs. Finally, we explain how to quantify the similarity between the labelled graphs given the solution to the QUBO problem formulation.

2.1 QUBO Problem Formulation of MIS

As discussed, there is a mapping between the MIS and LMWCS problems. Here, we review the QUBO problem formulation for the latter problem and then show how it can be converted into a QUBO problem for the MIS problem.

Let us consider two arbitrary graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$. To formulate the LMWCS problem as a QUBO

problem, one binary variable b_{ij}, defined below, is assigned to each possible pair $(i, j) \in V_1 \times V_2$. The weight of each pair is denoted by w_{ij}. A $0, 1$ configuration of binary variables b_{ij} depicts a common subgraph between G_1 and G_2. The vertex pairs that are assigned to 1 form the common subgraph. In other words, they are in the mapping. The goal is to find a configuration that maximizes the weight of the common subgraph, satisfying both the bijection and the user-defined requirements.

Formally, let

$$b_{ij} = \begin{cases} 1 & \text{if } (i, j) \in \text{mapping,} \\ 0 & \text{otherwise.} \end{cases}$$

The product set $V_1 \times V_2$ does not necessarily include all possible pairings between two vertex sets of the graphs---the vertex labels can exclude a specific pair from the mapping. For example, in a biology application, only the vertices with the same atomic number might be paired together.

Let us further define two sets to impose, respectively, the bijection and user-defined constraints on the mapping:

$$C = \big\{ \big((i, j), (m, n)\big) \big| \, i = m \vee j = n \big\},$$
$$C^* = \big\{ \big((i, j), (m, n)\big) \big| \, b_{ij} b_{mn} = 0 \text{ is a user-defined constraint} \big\}.$$

The QUBO problem formulation of the LMWCS problem is equivalent to the following formulation for all $a_{(i,j),(m,n)} > \min\{w_{ij}, w_{mn}\}$ [24]:

$$\max \left(\sum_{(i,j) \in V_1 \times V_2} w_{ij} b_{ij} - \sum_{\big((i,j),(m,n)\big) \in C \cup C^*} a_{(i,j),(m,n)} b_{ij} b_{mn} \right). \tag{1}$$

The first expression in the above formulation maximizes the weights of the selected pairs of vertices and the second expression penalizes the infeasible assignments due either to the bijection requirement C or the user-defined requirements C^*. An example of the latter requirements is when the user is interested in finding the maximum common clique between two graphs. The set C^* then includes $\big((i, j), (m, n)\big)$ if $(i, m) \notin E_1$ or $(j, n) \notin E_2$. It is worth mentioning that the degree of freedom represented as set C^* allows the enforcement of customized structure restrictions on the resulting MCS.

To convert Formulation (1) to an MIS formulation, we introduce a third graph called a conflict graph. Formally, let $G_c = (V_c, E_c)$ be the conflict graph of graphs G_1 and G_2, where $V_c \subset V_1 \times V_2$ and $E_c = C \cup C^*$. The vertices of the conflict graph are 2-tuples of the labelled graph vertices. All possible 2-tuples are not necessarily the vertices of the conflict graph---some pairs might be excluded by the user based on predetermined criteria. The edges of the conflict graph represent both the bijection and the user-defined requirements, defined by C and C^*, respectively. That is, there is an edge between two vertices $s = (i, j)$ and $l = (m, n)$ in the conflict graph if and only if the tuple of the corresponding pairings $\big((i, j), (m, n)\big)$ exists in $C \cup C^*$. The edges are called *conflicts* because they identify the pairs that cannot be present in the common subgraph simultaneously. Therefore, it is straightforward to see that the MIS of the conflict graph G_c corresponds to the MCS of graphs G_1 and G_2.

The QUBO problem formulation for the MIS problem of the conflict graph is

$$\max \left(\sum_{s \in V_c} w_s x_s - \sum_{(s,l) \in E_c} a_{sl} x_s x_l \right), \tag{2}$$

where x_s is a binary variable that is equal to 1 if the vertex s is included in the independent set and 0 otherwise, $w_s = w_{ij}$, and $a_{sl} > \min\{w_s, w_l\}$.

It can be demonstrated that both QUBO problem Formulations (1) and (2) are equivalent by a simple notation transformation.

2.2 QUBO Problem Formulation of Maximum Weighted Co-k-plex

A co-k-plex of a graph G is a subgraph of G in which each vertex has a degree of at most $k-1$. Therefore, the maximum weighted co-k-plex of the conflict graph G_c corresponds to the maximum weighted common subgraph of G_1 and G_2, where each possible pairing can violate at most $k-1$ constraints, either bijection or user-defined. To generalize QUBO problem Formulation (2), let us first define a star graph.

Definition 2.1. *A graph $S^k = (V, E)$ is a star graph of size k if it is a tree with $k+1$ vertices and one vertex of degree k.*

Based on the co-k-plex definition, we do not penalize the conflict edges. Each vertex in the conflict graph can have up to $k-1$ edges. Therefore, we only penalize in situations where a subset of the vertices induces a subgraph in which there is one vertex with degree greater than $k-1$. In other words, we penalize all subsets of vertices whose induced subgraph forms a star graph of size k.

Further, let us define the binary parameter $\mathcal{A}_{v_1,\ldots,v_{k+1}}$ as

$$\mathcal{A}_{v_1,\ldots,v_{k+1}} = \begin{cases} 1 & \text{if } \{v_1, \ldots, v_{k+1}\} \text{ induces } S^k, \\ 0 & \text{otherwise,} \end{cases}$$

where $v_1, \ldots, v_{k+1}$ are the vertices of the conflict graph G_c.

The QUBO problem formulation of the maximum co-k-plex is

$$\max \left(\sum_{v_i \in V_c} w_{v_i} x_{v_i} - \left(\sum_{(v_1,\ldots,v_{k+1})} a_{v_1,\ldots,v_{k+1}} \mathcal{A}_{v_1,\ldots,v_{k+1}} \prod_{i=1}^{k+1} x_{v_i} \right) \right), \tag{3}$$

where x_{v_i} is a binary variable equal to 1 if the vertex v_i is included in the MIS or 0 otherwise, and $a_{v_1,\ldots,v_{k+1}} > \min\{w_{v_1}, \ldots, w_{v_{k+1}}\}$. The k parameter is a tunable parameter that should be determined by the user.

The above formulation allows both bijection and user-defined conflicts to be present in the MIS. However, the formulation can be modified as below to exclude any bijection conflicts in the independent set of the conflict graph for $k \geq 2$:

$$\max \left(\sum_{v_i \in V_c} w_{v_i} x_{v_i} - \sum_{(v_1,v_2)\in\mathcal{C}} a_{v_1,v_2} x_{v_1} x_{v_2} - \left(\sum_{(v_1,\ldots,v_{k+1})} a_{v_1,\ldots,v_{k+1}} \mathcal{A}_{v_1,\ldots,v_{k+1}} \prod_{i=1}^{k+1} x_{v_i} \right) \right). \tag{4}$$

As mentioned in Section 2.1, the set C includes pairs of vertex sets V_1 and V_2 (alternatively, the vertices of the conflict graph) that cannot be simultaneously present in the common subgraph (alternatively, the independent set of the conflict graph). Therefore, the second expression in Formulation (4) penalizes the conflict edges due to the bijection requirement. Note that the definition of $\mathcal{A}_{v_1,\ldots,v_{k+1}}$ also slightly differs in this case: it equals 1 if vertices $v_1, \ldots, v_{k+1}$ induce a star graph of S^k whose edges are only the elements of set C^*, that is, the user-defined constraints.

The objective functions of both Formulations (3) and (4) are higher-order polynomials. There are several algorithms in the literature that map higher-order polynomials to quadratic polynomials [25, and references therein].

Multiple Graph Similarity   One of the main advantages of our conflict graph-based QUBO problem formulation is that it can easily be extended to measure similarity among multiple graphs. An example of a multiple graph similarity problem in molecular biology is the comparison of an unknown molecule with several sets of molecules to find the set that is the most similar. Other examples are human face recognition, where the goal is to find an image in a database that best matches several images of the same person taken from different angles [26], and graph prototype construction in clustering and classification contexts, in order to best represent all elements of a set of labelled graphs [27].

Let us formally assume that there are n labelled graphs $G_i = (V_i, E_i)$, where $i = 1, 2, \ldots, n$, to compare. An obvious strategy is to find the maximum weighted common subgraph of graph pair G_i and G_{i+1} for $i = 1, \ldots, n-1$ and then infer the global maximum weighted common subgraph. Following this strategy, we need to build $n-1$ conflict graphs and solve their corresponding NP-hard QUBO problem formulations. However, we can generalize our graph similarity approach to find the maximum common subgraph among a set of graphs by solving only *one* QUBO problem.

The core idea of comparing the n labelled graphs is the same as before: building the conflict graph and translating it into a QUBO problem formulation. The vertex set of the conflict graph of n graphs, V_c, is a subset of the product of $V_1 \times V_2 \times \cdots \times V_n$. Therefore, each vertex of the conflict graph is an n-tuple of individual graph vertices. The tuples that do not satisfy predefined compatibility criteria are excluded. There is an edge between two vertices in the conflict graph if there is any reason imposing that the two vertices cannot be present in the similarity set at the same time. The maximum weighted independent set problem of the conflict graph can then be formulated as a QUBO problem following the same procedure as already explained.

2.3 Similarity Measure

The solution to the QUBO problem formulation presented in the previous section is a binary vector with size $|V_c|$. The non-zeros in the solution vector identify the pair of vertices present in the maximum weighted independent set of the conflict graph. To measure the similarity between graphs, there exists a large number of similarity metrics in the literature, represented mainly in terms of subset relations [3, 28]. In this paper, we use the metric

$$ \mathcal{S}(G_1, G_2) = \delta \max \left\{ \frac{|V_c^1|}{|V_1|}, \frac{|V_c^2|}{|V_2|} \right\} + (1 - \delta) \min \left\{ \frac{|V_c^1|}{|V_1|}, \frac{|V_c^2|}{|V_2|} \right\}, \quad \delta \in [0, 1], \tag{5} $$

where $|V_c^1|$ and $|V_c^2|$, respectively, denote the number of distinct vertices of G_1 and G_2 in the maximum weighted independent set of the conflict graph. This metric quantifies the contribution of each graph to the MIS. Our metric is the convex combination of two existing similarity measures---Bunk and Shearer, and Asymmetric [28]---providing the user with more flexibility (see Section 4 for more details).

In the next section, we utilize our QUBO-based graph similarity measure in the context of molecular similarity.

3   Molecular Similarity

Molecular similarity methods play an important role in many aspects of chemical and pharmaceutical research, including biological activity prediction [29], protein-ligand docking [30], and database searching [31]. Determining similarity between compounds is practical due to the *similar property principle* [32], which states that structurally similar molecules are expected to display similar properties. Therefore, various aims in biological research can be accomplished by determining structural similarity between molecules. Similarity, however, is a complicated concept to evaluate since it is not a measurable property of the molecules themselves. Indeed, it may have different interpretations depending on external criteria. This ambiguity increases the difficulty of developing an accurate computational method for determining molecular similarity.

To build a measure of molecular similarity, three basic components are required: a molecular representation that not only encodes relevant molecular features but also contains their associated weights, a method that compares molecular representations, and a function that evaluates their similarity. Many similarity measures have been proposed in the literature which can be categorized into two classes. The first class of measures uses the intuitive representation of the molecular atom-bond structure as a graph. Comparison of molecules based on graph representations is often accomplished using graph similarity techniques. As mentioned in Section 1, the computational cost of graph similarity techniques has resulted in the development of algorithms for specific graph structures. These approaches present limitations on the types of objects that can be used for comparison. However, the QUBO-based measure presented in this paper is applicable to any arbitrary graph.

The second class of measures uses a vector-based representation called fingerprint, a conventional concept in chemical informatics and related fields. Fingerprints are binary vectors that indicate the presence or absence of certain features of a molecule. Each bit can be either 1 or 0, representing whether or not the molecule contains an associated feature. Molecules modelled as fingerprints are compared using vector-based comparison measures such as Tanimoto, Cosine, Dice, and Euclidean distance [28]. Although fingerprints are easy to use and their pairwise comparisons are computationally efficient, they have certain drawbacks. For instance, fingerprints cannot be used to assess for certain whether a particular pattern is present or not in a molecular graph [28]. When a bit in the fingerprint is set to 1 for some pattern, it means that the pattern is present in the molecule only with some probability. Furthermore, fingerprint-based approaches do not usually consider underlying information about the molecular topology.

In the next section, we discuss how our graph similarity method, described in Section 2, can be applied to the molecular similarity problem.

3.1 Graph-based Molecular Similarity

To measure similarity between molecules, we discuss how to represent molecules as graphs and then build a conflict graph. Given the conflict graph, it is then straightforward to formulate the QUBO problem and quantify the similarity as described in Sections 2.2 and 2.3.

Modelling Molecules as Graphs   A direct graph representation of the structural formula of a chemical compound is often referred to as a molecular or chemical graph. In general, a molecular graph-based representation can be categorized into two classes based on the structural dimensionality of the molecule, that is, two-dimensional (2D) and three-dimensional (3D). In the former representation, each vertex is associated with one atom and there is an edge between two vertices indicating their chemical bond. The latter representation includes 3D structural information. For example, in the 3D molecular graph introduced by Raymond and Willett [3], there is an edge between each pair of vertices indicating the geometrical distance between the two corresponding atoms.

Since a graph is an abstract mathematical object, vertices and edges do not need to correlate with atoms and bonds directly. Hence, vertices could represent a group of atoms, given some meaningful criteria. Accordingly, a reduced representation of the chemical structures reduces the size of the graphs being compared and thus increases the effectiveness of the algorithms to solve the MIS (or its relaxation) problem. In this paper, we build a labelled reduced graph representation of a molecule. The steps of this reduction process are illustrated in Figure 1. The first step is to construct the atom-bond graph. We then identify ring structures of the molecules and reduce each corresponding ring's set of vertices to a single vertex. If two vertices representing rings share one or more atoms, an edge between these vertices is then added and labelled as *artificial*, emphasizing that it does not represent a natural chemical bond. Furthermore, vertices and edges of our graph representation are labelled with various information pertaining to atoms and bonds, respectively.

Formally, the labelled reduced chemical graph can be written as $G = (V, E, \mathcal{L}_V, \mathcal{L}_E)$, where V is the set of vertices and $E = V \times V$ is the set of edges. A set of labels, $\mathcal{L}_V : V \to \{l_1, \dots, l_8\}$, is assigned to each vertex. Each label represents a specific property of the atom or ring. The label set contains the atomic number, the number of implicit hydrogen bonds[1] associated with the atom (or ring), the formal charge, the degree of the node distinguishing covalent bond order (single, double, or triple), the set of covalent bond order between atoms in a ring, a weight denoting the number of atoms reduced to one vertex, and a 3D vector indicating the position of the atom or the geometrical centre of the ring. Likewise, a label l_e is assigned to each edge ($\mathcal{L}_E : E \to l_e$), representing whether the edge is an artificial link, or a single, double, or triple covalent bond.
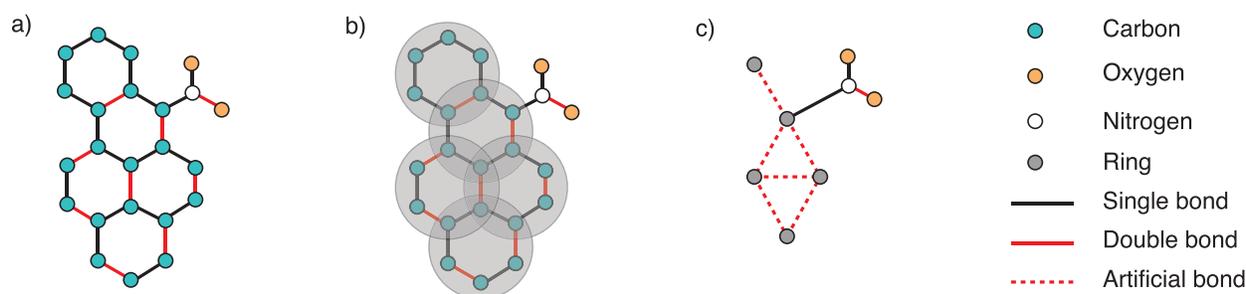

Figure 1: Steps for modelling molecules as graphs: the first step is to a) build a basic structural representation of a molecule, then b) identify ring structures, which subsequently are added as a vertex to the reduced graph as shown in c).

Building a Conflict Graph   The conflict graph $G_c = (V_c, E_c)$ of two chemical graphs $G = (V, E)$ and $G' = (V', E')$ is defined, according to Section 2.1, by its vertex set $V_c \subseteq V \times V'$ and edge set $E_c \subseteq V_c \times V_c$ as:

$$V_c = \{(v_i, v_j') \mid \mathcal{L}_V(v_i) = \mathcal{L}_V(v_j')\}, \tag{6}$$

$$E_c = \{((v_i, v_j'), (v_k, v_l')) \mid i = k \lor j = l \lor \mathcal{L}_E(v_i, v_k) \neq \mathcal{L}_E(v_j', v_l')\}, \tag{7}$$

---

[1]Explicit hydrogen atoms are added as vertices to the graph.

where $\mathcal{L}_V$ is the set of labels assigned to the vertices as specified above. The vertex set V_c includes a pair of vertices from the chemical graphs under comparison if they have matching labels. It should be noted that each of the labels can be set as a default or optional parameter according to some similarity criteria for the specific application. Here, we assume that the atomic number, the weight, and the number of hydrogen bonds associated to atoms are matching properties set by default, while the number of hydrogen bonds associated to rings (RH), the bond order within a ring (RB), the formal charge (FC), and the degree of the nodes (DN) are optional parameters. The first two conditions associated with the creation of edges in (7) imply that a conflict edge between two vertices (v_i, v_j') and (v_k, v_l') is added if one of the vertices from the chemical graphs has been matched twice, that is, if the bijection requirement is violated. The third condition in (7) indicates that a conflict edge is added if there is a mismatch between the edge labels of the two chemical graphs. The set of conflict edges can be extended to include 3D molecular information. Since molecular compounds are inherently 3D, it is expected that the incorporation of 3D information would be preferable over 2D. To include 3D information, we consider the geometrical distance between a pair of atoms. That is, an edge between two vertices in the conflict graph (e.g., between (v_i, v_j') and (v_k, v_l')) is added if the distance between the vertices v_i and v_k of the graph G is not comparable to the distance between vertices v_j' and v_l' of the graph G'. To determine if two distances are comparable, we introduce a user-defined threshold distance d_t. Thus, a new condition can be added to (7),

$$E_c = \left\{ ((v_i, v_j'), (v_k, v_l')) \mid |d(v_i, v_j) - d(v_j', v_l')| > d_t \right\}, \tag{8}$$

where $d(v_i, v_j)$ is the geometrical distance between the vertices v_i and v_k. The distance $d(v_j', v_l')$ is defined similarly. When $d_t = 0$, a restrictive 3D representation is modelled, penalizing in situations where the distances between atoms in corresponding graphs are not equal. The conflict graph in this case has the highest density. Its density reduces as we allow for the modelling of increasingly relaxed 3D representations by increasing d_t. We eventually reach a model that represents the 2D structure when d_t is greater than any of the distances between atoms. After the conflict graph is built, a QUBO problem formulation can be obtained as explained in Sections 2.2 and 2.3.

In the following section, we present a case study applying our molecular similarity method to predict the mutagenicity of molecules.

4   Case Study: Prediction of Mutagenicity

In this section, we use a supervised machine learning approach to build a novel model that predicts the mutagenicity of chemical molecules. The predictive model is based on the proposed QUBO-based molecular similarity method described in Section 3. We then present some results, validating the performance of the predictive model using two real data sets and compare it with the conventional fingerprint-based approach.

4.1 Predictive Model

In the pharmaceutical industry, the liability of drugs needs to be assessed before commercialization. To that end, researchers attempt to estimate a number of desired properties associated with absorption, distribution, metabolism, excretion, and toxicity for collections of drugs by using their molecular structures. One of the properties related to toxicology is mutagenicity. Detecting mutagenicity is a crucial step in drug discovery because mutagenic chemicals may potentially be carcinogenic. That is, they may damage DNA, resulting in mutations that can cause cancer. Using preclinical experimental tests to predict the mutagenicity of a molecule is significantly time-consuming and expensive. Therefore, developing robust *in silico* predictive techniques is of considerable importance in decreasing the time between early development and launch phases in the life cycle of a drug.

Our predictive model applies the κ-nearest neighbours (κ-NN) statistical method as the classifier to assign a ``mutagen'' or ``non-mutagen'' label to a molecule. Although there are various classifiers in the machine learning literature, we use the κ-NN classifier because of its simplicity and its well-known high performance in the context of computational biology [33, 34]. The κ-NN classifier measures the pairwise distance between an unknown molecule i and a data set of molecules with known mutagenicity labels. It then selects the κ closest molecules to molecule i from the data set and counts the contributions of mutagenic and non-mutagenic neighbours. Finally, the label with the highest contribution is assigned to molecule i. The pairwise distance between molecule i and molecule j in the data set, denoted by $\mathcal{D}_{ij}$, is equal to $1 - \mathcal{S}(G_i, G_j)$, where G_i and G_j are the corresponding graphs of molecules i and j, respectively, and $\mathcal{S}(G_i, G_j)$ is the similarity measure between the two graphs, quantified using Equation (5). The contribution of neighbour j is measured

by its weight being equal to $1/\mathcal{D}_{ij}$. Although various weight functions have been proposed for the κ-NN classifier, our early experimentation showed that using the inverse of the distance as a weight function results in better performance.

To measure the similarity between two molecules, we need to assign values to the optional parameters, including the threshold distance between atoms (d_t), the number of hydrogen bonds associated to rings (RH), the bond order within a ring (RB), the formal charge (FC), and the degree of the nodes (DN). While the first parameter takes any positive real number, the last four are assigned a value of 1 if they are taken into account in constructing the conflict graph, or 0 otherwise. Furthermore, the co-k-plex relaxation parameter (i.e., $k = 1$ or $k > 1$) should be determined. Different combinations of parameter values may result in different similarity measures, changing the performance of the predictive model as a consequence. For each combination, we use a k-fold cross-validation technique to find the predictive model's performance. The combination yielding the best performance determines the best set of parameter values.

The cross-validation is performed on a data set provided by Xu et al. [35] that contains 7617 molecules, 4252 of them mutagenic and 3365 non-mutagenic. The data set is partitioned into five folds, where, initially, the first fold is reserved as the validation set and the other folds are used as the training set. To assign a ``mutagen'' or ``non-mutagen'' label to each molecule in the validation set, we use the 3-NN classifier. The performance of the classifier is then evaluated using four performance metrics, defined below. This procedure is repeated five times, each fold acting once as a validation set. Finally, the performance of the predictive model is equal to the average performance of the classifier over five folds. The number of folds and neighbours are chosen as five and three, respectively, because our early experiments showed that different values of folds and neighbours do not have a significant effect on the performance of the classifier.

The four performance metrics used are sensitivity, specificity, accuracy, and precision. They are based on counting four numbers in the validation set: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Sensitivity (true positive rate) is equal to $TP/(TP + FN)$, the conditional probability that a molecule is labelled ``mutagen'' given it truly is mutagenic. Specificity (true negative rate) equals $TN/(TN + FP)$, the conditional probability that a molecule is labelled as ``non-mutagen'' given it truly is non-mutagenic. Accuracy is equal to $(TP + TN)/(TP + FP + TN + FN)$, the probability of predicting a molecule's label correctly. Precision is the likelihood of labelling a truly mutagenic molecule ``mutagen'', and is equal to $TP/(TP + FP)$. To compare predictive models, we first consider the accuracy metric. However, due to the accuracy paradox, a high accuracy value does not guarantee that a classifier is good. A good classifier not only has high accuracy, but also reasonable performance in terms of the other three metrics. Furthermore, since sensitivity is negatively correlated with specificity and precision, a good classifier is the one with balanced performance in terms of all metrics.

Before discussing the results, we provide an example to show how the similarity measure between two molecules changes as the parameter values change.

Illustrative Example   We measure the similarity between two molecules, ``1,8-Dinitrobenzo[e]pyrene'' (Mol 0) and ``6-Nitro-7,8,9,10-tetrahydrobenzo[pqr]tetraphene'' (Mol 1), for different parameter values. The molecules are represented in Figure 2.
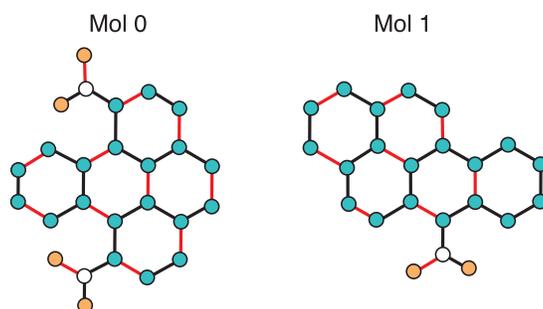

Figure 2: The molecular representations of molecules 0 and 1.

Table 1 shows the conflict graphs and the similarity measures of Mol 0 and Mol 1 for different parameter values when $\delta = 0.5$. The black, green, and blue edges are due to bijection, distance, and edge label conflicts, respectively.

As expected, the conflict graphs and similarity measures change for different values of optional parameters, distance
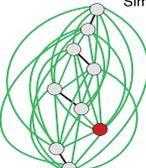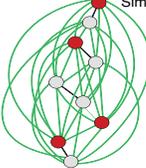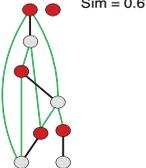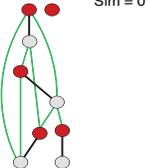
Optional Parameters	$d_t = 0$		$d_t = 1.5$	
	$k = 1$	$k = 4$	$k = 1$	$k = 4$
FC = 1, DN = 1	Sim = 0.24	Sim = 0.57	Sim = 0.61	Sim = 0.61
RH = 1, FC = 1	Sim = 0.24	Sim = 0.57	Sim = 0.61	Sim = 0.66
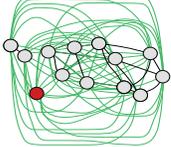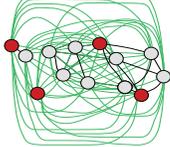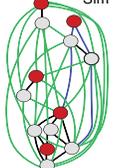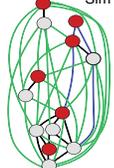
Table 1: The conflict graphs and similarity measures of Mol 0 and Mol 1 for different parameter values when $\delta = 0.5$.

thresholds, and co-k-plex relaxations. More specifically, for given optional parameters, the conflict graphs are denser for lower distances and the similarity values increase or remain the same as the similarity criteria become more relaxed by increasing either d_t or k.

4.2 Results and Discussion

In this section, we use the training data set of 7617 molecules to find the best set of parameters, then validate the predictive model performance on two test sets, and finally compare the predictive model with the fingerprint-based approach. To find the best set of parameters, we investigate how different combinations of the optional molecular labels RH, RB, FC, and DN change the performance metrics. We then report on the effect of different distance threshold and co-k-plex relaxation values, d_t and k, respectively.

Molecular Feature Selection   As discussed earlier, there are four optional molecular features---RH, RB, FC, and DN---that can be considered in 16 different layouts to build the conflict graph. Each layout setting results in a different conflict graph size (see Table 1). In this paper, we use an exhaustive solver to solve the QUBO problem formulations of conflict graphs whose number of vertices is less than or equal to twenty. The molecules that have a larger conflict graph are excluded from the cross-validation. Therefore, the number of molecule pairs used in each replication of the cross-validation differs for each of the different layouts. To make a fair comparison of the effect of different molecular features, we have excluded the pairs whose similarity cannot be quantified in at least one of the layout settings from the training sets of all layouts. In other words, we have considered the same molecule pairs in all layouts, which we refer to as ``reduced pairs set''.

Figure 3 illustrates the mean accuracy of our 3-NN classifier with $k = 1$ and $d_t = 1.5$, along with error bars, for different layouts, considering the total number of pairs and reduced number of pairs solved. As shown, Layout 0, where none of the molecular features are taken into account, has the lowest number of pairs solved (the two accuracy plots are the same). Since excluding the molecular features increases the number of matching vertices in the conflict graph, it is more likely that there are more molecule pairs whose conflict graphs have more than twenty vertices in Layout 0 and their QUBO problem formulations cannot be solved. Figure 3 also shows that for the rest of the layouts, the accuracy is greater when the total number of pairs solved is considered.

We can further observe that when reduced pairs (blue plots) are considered, the highest accuracy achieved across all layouts is between $[0.71, 0.74]$, with Layout 0 having the highest accuracy. When the total number of pairs solved (red plots) are considered, the highest accuracy for all layouts slightly increases. The highest accuracy obtained across all layouts in this case is between $[0.72, 0.74]$, where Layout 7 has the highest accuracy. It is worth mentioning that Layouts 0, 1, 3, 8, 12, and 13 also have a mean accuracy of 0.74. However, we report Layout 7 as being the best since it has a slightly higher accuracy---it is 0.744 while the others are between 0.735 and 0.740. Given that the standard error of the accuracy is almost 1% and the difference between the highest accuracy in Layouts 0 and 7 is 0.004, it is reasonable to consider both Layouts 0 (RH = 0, RB = 0, FC = 0, DN = 0) and 7 (RH = 1, RB = 0, FC = 0, DN = 1) as the best layouts for investigating the impact of the other parameters.
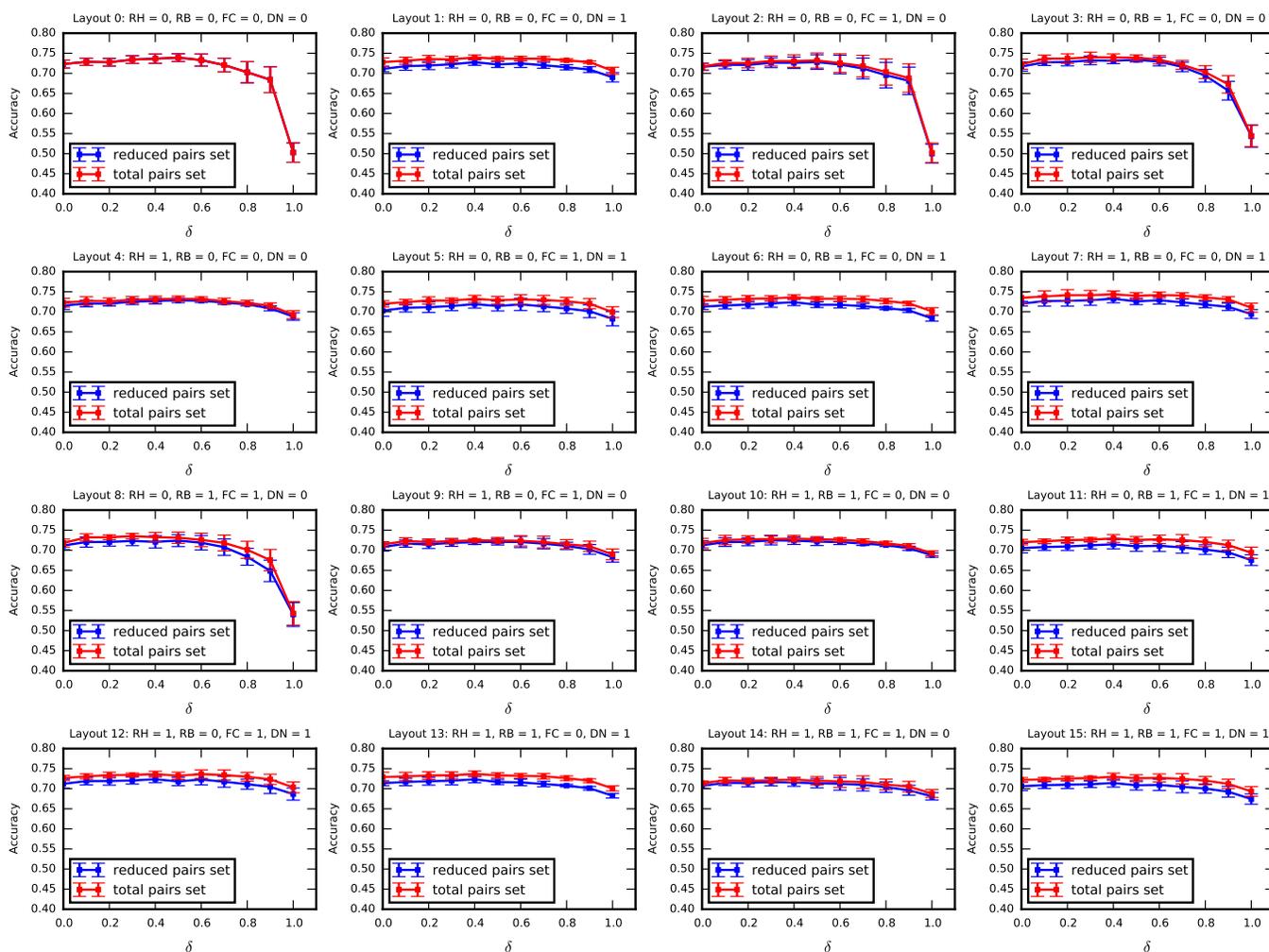
Figure 3: Accuracy of 16 different layouts of molecular features with $d_t = 1.5$ and $k = 1$ in two cases: total number of pairs (total pairs set) and reduced number of pairs (reduced pairs set) solved.

Figure 3 also illustrates that the accuracy of all layouts, either with a reduced pairs set or total pairs set, is robust with respect to $\delta \in [0.3, 0.5]$. Recall that different values of δ in Equation (5) yield different similarity measures, affecting the performance of the classifier as a consequence. The robustness of the accuracy across a range of δ values affords us the advantage of being able to choose δ such that the classifier has reasonable performance with respect to the other three metrics---precision, sensitivity, and specificity. Table 2 shows the four metrics for Layouts 0 and 7, where $\delta \in [0.3, 0.5]$. As expected, specificity and precision are negatively correlated with sensitivity. We could therefore choose δ to balance for the metric that is most relevant to the context of our classifier. In our predictive model of mutagenicity, it is more desirable to have a high sensitivity (i.e., most of the mutagenic molecules are identified) at a potential loss of precision and specificity where some of the non-mutagenic molecules are retrieved as mutagenic. Given Table 2, setting $\delta = 0.4$ seems reasonable in the context of our case study.

δ	Layout 0				Layout 7			
	Accuracy	Precision	Sensitivity	Specificity	Accuracy	Precision	Sensitivity	Specificity
0.3	0.73	0.76	0.78	0.68	0.74	0.76	0.80	0.68
0.4	0.74	0.76	0.77	0.70	0.74	0.76	0.80	0.68
0.5	0.74	0.77	0.76	0.72	0.74	0.76	0.79	0.69

Table 2: Accuracy, precision, sensitivity, and specificity for Layouts 0 and 7 and different δ values.

Distance Threshold Parameter   As mentioned in Section 3.1, the distance threshold parameter d_t allows some information regarding the 3D structure of the molecules to be incorporated into the conflict graph. The number of edges in the conflict graph of two molecules decreases as d_t increases. Therefore, the size of the maximum weighted independent set of the conflict graph might increase, capturing more similarities between molecules. However, it is clear that there is an upper bound on d_t where the conflict graph does not change. Therefore, the similarity measures and consequently the performance of the classifier remain constant. Figure 4 illustrates the four performance metrics of the classifier for Layouts 0 and 7 and six different values of d_t (0, 0.5, 1, 1.5, 5, and 10) when $k = 1$, $\delta = 0.4$. We see that the classifier performs poorly at $d_t = 0$, while its performance substantially improves for $d_t = 0.5$ and then remains almost the same for higher values of d_t.

Intuitively, we would expect the accuracy for a given layout, δ, and k to be a concave and monotonic function of d_t. The accuracy increases as d_t increases, but it is bounded above since the similarity values between molecules would eventually be invariant at higher d_t values.
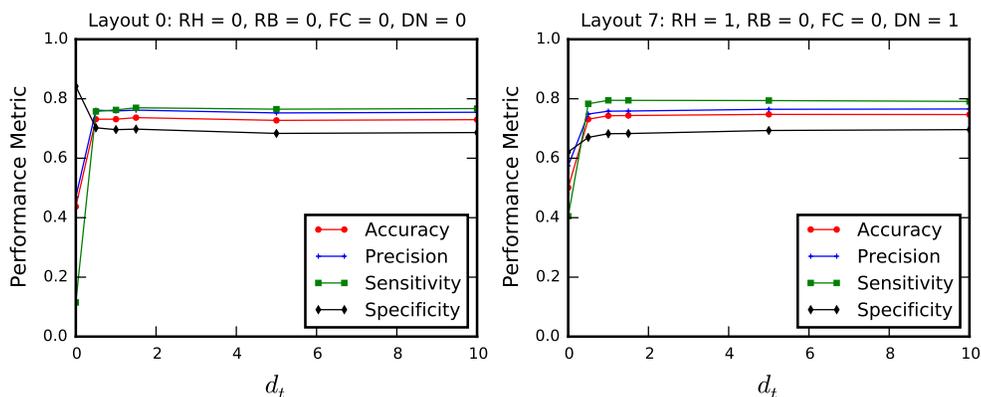

Figure 4: Accuracy, precision, sensitivity, and specificity for Layouts 0 and 7 and different d_t values when $k = 1$ and $\delta = 0.4$.

Co-k-plex Relaxation   At higher values of k, we allow more conflicts to be present in the independent set of the conflict graph. As a result, the similarity measure between two molecules increases. Table 3 shows the accuracy of Layouts 0 and 7 for different k and d_t values, where $\delta = 0.4$. The highest accuracy value is 0.75 for Layout 7, $k = 3$, and $d_t = 1.5$.

d_t	Layout 0					Layout 7				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0	0.44	0.48	0.72	0.65	0.65	0.50	0.61	0.67	0.69	0.71
0.5	0.73	0.72	0.71	0.70	0.71	0.73	0.74	0.74	0.74	0.74
1	0.73	0.73	0.72	0.71	0.71	0.74	0.74	0.74	0.74	0.74
1.5	0.74	0.73	0.71	0.71	0.70	0.74	0.74	**0.75**	0.74	0.74

Table 3: Accuracy of Layouts 0 and 7 for different k and d_t values, where $\delta = 0.4$.

Table 3 also shows that when $d_t = 0$, the accuracy substantially improves at higher values of k for both Layouts 0 and 7. It changes from a useless (i.e., randomly assigning mutagenicity labels to molecules) to a useful classifier. For distance $d_t = 0$, the accuracy in Layout 7 has a strictly increasing trend, whereas the accuracy in Layout 0 reaches its highest value at $k = 3$ and then decreases. We can further observe that for $d_t > 0$, the difference in accuracy diminishes as k increases. This observation can be justified since we know that the number of conflict edges decreases for higher values of d_t and that higher values of k are more beneficial when the conflict graph is denser. To conclude, we expect that the accuracy would be a non-monotonic function of k for a given d_t. That is, there would be eventually a $k_{\max}$ such that the performance of the classifier for all $k > k_{\max}$ is lower than its performance at $k_{\max}$, ultimately reaching a constant limit.

External Validation and Benchmarking   Our investigation indicates that the set of parameter values including Layout 7, $d_t = 1.5$, $k = 3$, and $\delta = 0.4$ obtains the highest accuracy of 0.75 ± 0.01. We use these parameter values to validate our QUBO-based maximum weighted co-k-plex relaxation model on two external test sets. The first test set is a balanced

Test Set	Performance Metric	Method		
		QUBO-based Maximum Weighted Co-3-plex	QUBO-based Maximum Weighted Co-1-plex	MACCS Fingerprint
Set 1	Accuracy	**0.81**	0.80	0.76
	Precision	0.79	0.77	0.69
	Sensitivity	0.84	0.85	0.95
	Specificity	0.78	0.75	0.58
Set 2	Accuracy	**0.80**	0.79	0.77
	Precision	0.78	0.77	0.70
	Sensitivity	0.83	0.85	0.95
	Specificity	0.77	0.74	0.60

Table 4: Accuracy, precision, sensitivity, and specificity of the 3-NN classifier, where the QUBO-based maximum weighted co-3-plex, the QUBO-based maximum weighted co-1-plex, and the MACCS fingerprint-based approaches are used.

set provided by Xu et al. [35], containing 234 molecules with the same number of mutagens and non-mutagens. The second test set is built using the sets given by Hansen et al. [36] and Xue et al. [35]. Hansen et al. constructed a data set comprising 6512 molecules, which we compare with our training set of 7617 molecules, excluding the molecules that are common to both. Combining the 106 molecules that are not in our training set with 731 molecules of another set given by Xue et al. forms a new set of molecules. The second test set contains 300 molecules randomly chosen from the new set where half are mutagenic and half are non-mutagenic.

The performance of our similarity model on both test sets is compared with a fingerprint-based approach. We use the MACCS fingerprint, a vector of 166 bits representing molecular features. Table 4 shows the four performance metrics of the 3-NN classifier on both test sets, where the QUBO-based maximum weighted co-3-plex, the QUBO-based maximum weighted co-1-plex, and the fingerprint-based approaches are used. In Table 4, set 1 and set 2 refer to the first test set and the second test set, respectively. The fingerprint-based approach measures the similarity between two molecules as ``similarity $= 1 -$ distance''. We use the Euclidean distance between two molecular vectors. For a fair comparison, we exclude the molecule pairs whose similarity values are not found in Layout 7 from the training set of the fingerprint-based method. As shown, both QUBO-based maximum weighted co-k-plex methods have better performance than the fingerprint-based method in three metrics---accuracy, precision, and specificity---in both test sets. Although the fingerprint-based method has a higher sensitivity in both sets, its low specificity value implies that the majority of non-mutagenic molecules are blindly labelled as mutagens, increasing the sensitivity. It is worth mentioning that our graph-based maximum weighted co-k-plex relaxation similarity measures not only yield a higher classification quality, but also provide complete information on the common molecular substructures while the fingerprint-based measure reports only one value as the similarity measure without providing any insights.

Table 4 also shows the superiority of the maximum weighted co-3-plex relaxation method over the maximum weighted co-1-plex relaxation. The higher accuracy of the maximum weighted co-3-plex relaxation method provides evidence that accounting for the noisy data results in a more accurate prediction of mutagenicity.

5 Conclusions

In this paper, we address the problem of measuring similarity among graphs with noisy data whose graph representations are partially accurate in the context of predicting molecular mutagenicity. To account for data distortion, we develop a novel QUBO problem formulation for the maximum weighted co-k-plex problem to measure the similarity of graphs, relaxing the requirement that the subgraph matching be exact. The relaxation allows for the existence of dissimilarities among graphs up to a predefined level. The QUBO-based similarity measure can also be extended to the problem of measuring similarity among multiple graphs.

The QUBO-based similarity measure is then used to quantify the similarity of molecules. We discuss efficient approaches to the abstraction of molecules as labelled graphs and build their corresponding conflict graphs, incorporating various molecular features, including 3D structural information. To assess the performance of the developed measure, two real-molecule data sets are used to predict mutagenicity, exploiting a machine learning

approach. Our results demonstrate that relaxing the definition of similarity by setting $k = 3$ yields the highest prediction accuracy. Investigation of the results further indicates that the accuracy of mutagenicity prediction is poor when the strict 3D representation is modelled. There is also a maximum value for the level of relaxation in the maximum weighted co-k-plex formulation where the prediction accuracy is the highest. Finally, comparison of the QUBO-based similarity measure with the existing fingerprint-based measure in the literature shows the superiority of our measure.

Although the developed similarity measure can be solved by a quantum annealer, we use an exhaustive solver to find the optimal QUBO-based similarity measures. We will pursue the study of the potential speed-up of the quantum annealing solver on the maximum weighted co-k-plex problem in future work.

Acknowledgements

References

[1] M. Wall. Big Data: Are you ready for blast-off? *BBC News*, March 2014.

[2] B. Kehr, K. Trappe, M. Holtgrewe, and K. Reinert. Genome alignment with graph data structures: A comparison. *BMC Bioinformatics*, 15:1--20, 2014.

[3] J. W. Raymond and P. Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16:521--533, 2002.

[4] M. A. Eshera and K-S Fu. An image understanding system using attributed symbolic representation and inexact graph-matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:604--618, 1986.

[5] L. A. Zager and G. C. Verghese. Graph similarity scoring and matching. *Applied Mathematics Letters*, 21(1):86 -- 94, 2008.

[6] H. Bunke. Graph matching: Theoretical foundations, algorithms, and applications. In *Proceedings of Vision Interface, Montreal*, pages 82--88, 2000.

[7] F. N. Abu-Khzam, N. F. Samatova, M. A. Rizk, and M. A. Langston. The maximum common subgraph problem: Faster solutions via vertex cover. In *IEEE/ACS International Conference on Computer Systems and Applications*, pages 367--373, 2007.

[8] B. Balasundaram and F. Mahdavi Pajouh. Graph theoretic clique relaxations and applications. In Panos M. Pardalos, Ding-Zhu Du, and Ronald L. Graham, editors, *Handbook of Combinatorial Optimization*, pages 1559--1598. Springer New York, 2013.

[9] B. Balasundaram, S. Butenko, and I. V. Hicks. Clique relaxations in social network analysis: The maximum k-plex problem. *Operations Research*, 59(1):133--142, 2011.

[10] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, New York, 1979.

[11] R. G. Downey and M. R. Fellows. Fixed-parameter tractability and completeness II: On completeness for $w[1]$. *Theoretical Computer Science*, pages 109--131, 1995.

[12] Y-K Wang, K-C Fan, and J-T Horng. Genetic-based search for error-correcting graph isomorphism. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 27:588--597, 1997.

[13] Arvind Gupta and Naomi Nishimura. The complexity of subgraph isomorphism for classes of partial k-trees. *Theoretical Computer Science*, 164(1):287--298, 1996.

[14] G. Kolliasa, M. Satheb, O. Schenkc, and A. Gramaa. Fast parallel algorithms for graph similarity and matching. *Journal of Parallel and Distributed Computing*, 74:2400--2410, 2014.

[15] Giuseppe E Santoro and Erio Tosatti. Optimization using quantum mechanics: Quantum annealing through adiabatic evolution. *Journal of Physics A: Mathematical and General*, 39(36):R393, 2006.

[16] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, Joshua Lapan, Andrew Lundgren, and Daniel Preda. A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem. *Science*, 292(5516):472--475, 2001.

[17] MW Johnson, MHS Amin, S Gildert, T Lanting, F Hamze, N Dickson, R Harris, AJ Berkley, J Johansson, P Bunyk, et al. Quantum annealing with manufactured spins. *Nature*, 473(7346):194--198, 2011.

[18] Sergio Boixo, Vadim N Smelyanskiy, Alireza Shabani, Sergei V Isakov, Mark Dykman, Vasil S Denchev, Mohammad Amin, Anatoly Smirnov, Masoud Mohseni, and Hartmut Neven. Computational role of multiqubit tunneling in a quantum annealer. 2015. arXiv:1502.05754.

[19] T. Lanting, A. J. Przybysz, A. Yu. Smirnov, F. M. Spedalieri, M. H. Amin, A. J. Berkley, R. Harris, F. Altomare, S. Boixo, P. Bunyk, N. Dickson, C. Enderud, J. P. Hilton, E. Hoskinson, M. W. Johnson, E. Ladizinsky, N. Ladizinsky, R. Neufeld, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, S. Uchaikin, A. B. Wilson, and G. Rose. Entanglement in a quantum annealing processor. *Phys. Rev. X*, 4:021041, May 2014.

[20] E. G. Rieffel, D. Venturelli, B. O'Gorman, M. B. Do, E. M. Prystay, and V. N. Smelyanskiy. A case study in programming a quantum annealer for hard operational planning problems. *Quantum Information Processing*, 14:1--36, 2015.

[21] D. Venturelli, D. J. J. Marchand, and G. Rojo. Quantum annealing implementation of job-shop scheduling. 2015. arXiv:1506.08479.

[22] G. Rosenberg, P. Haghnegahdar, P. Goddard, P. Carr, K. Wu, and M. L. Prado. Solving the optimal trading trajectory problem using a quantum annealer. 2015. arXiv:1508.06182.

[23] Endre Boros and Peter L Hammer. Pseudo-boolean optimization. *Discrete applied mathematics*, 123:155--225, 2002.

[24] C. Vicky. Minor-embedding in adiabatic quantum computation: I. The parameter setting problem. *Quantum Information Processing*, 7(5):193--209, 2008.

[25] Endre Boros and Aritanan Gruber. On quadratization of pseudo-boolean functions. In *ISAIM*, 2012.

[26] E. Bengoetxea. *Inexact Graph Matching Using Estimation of Distribution Algorithms*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2002.

[27] Albert S. Ribalta. *Multiple Graph Matching and Applications*. PhD thesis, Universitat Rovira I Virgili, 2012.

[28] D. Baum. *A Point-Based Algorithm for Multiple 3D Surface Alignment of Drug-Sized Molecules*. PhD thesis, Free University of Berlin, 2007.

[29] H. Van De Waterbeemd and E. Gifford. Admet in silico modelling: Towards prediction paradise? *Nature Review Drug Discovery*, 2:192--204, 2003.

[30] F. S. Kuhl, G. M. Crippen, and D. K. Friesen. A combinatorial algorithm for calculating ligand binding. *Journal of Computational Chemistry*, 5:24--34, 1984.

[31] P. Willett. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *Journal of Computational Biology*, 6:447--457, 1999.

[32] Mark A. Johnson and Gerald M. Maggiora. *Concepts and applications of molecular similarity*. Wiley, New York, 1990.

[33] D. K. Slonim. From patterns to pathways: Gene expression data analysis comes of age. *Nature Genetics*, 32:502--508, 2002.

[34] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, and L. M. Sturla. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436--442, 2002.

[35] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee, and Y. Tang. In silico prediction of chemical Ames mutagenicity. *Journal of Chemical Information and Modeling*, 52(11):2840--2847, 2012.

[36] Katja Hansen, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. Benchmark data set for in silico prediction of Ames mutagenicity. *Journal of Chemical Information and Modeling*, 49(9):2077--2081, 2009.