Quantum Similarity: An Enhanced Framework Using AUG Weighting Technique

Mohammed Mumtaz Al-Dabbagh

Computer Engineering Department, Tishk International University, Erbil, Iraq

Corresponding author. E-mail: mohamad.aldabagh@tiu.edu.iq

Received: Dec. 07, 2023; Accepted: Feb. 23, 2024

In drug discovery, Virtual Screening (VS) encompasses the computational endeavor to unearth novel lead compounds via molecular similarity analysis. Amongst the array of techniques for ligand-based virtual screening (LBVS), similarity searching emerges as a quintessential and widely-adopted method. A prevailing assumption in many similarity search strategies posits that molecular structural features, irrespective of their biological activity, hold comparable importance. This paper delves into the AUG weighting scheme, aiming to bolster the application of quantum theory in LBVS, culminating in the formulation of a novel quantum-based similarity approach termed the QM-AUG method. Within the domain of molecular structure representation, the role of mathematical quantum space in enhancing the potency of the similarity method cannot be understated. The AUG weighting technique scrutinizes the potential consequences of adjusting weights allotted to chemical fragments, with the overarching objective of refining the quantum model's efficiency in LBVS. Methodological robustness was gauged through the recall metrics of extracted active molecules, notably within the top 1% and 5% echelons. Furthermore, comprehensive experimental evaluations using authentic datasets, specifically the MDL Drug Data Report (MDDR) and Maximum Unbiased Validation (MUV), indicate that the proposed method surpasses the performance seen with its implementation in the Bayesian Inference Network and the conventional Taninmoto coefficient.

Keywords: ligand-based; Virtual screening; Quantum-based similarity; Similarity searching method; Quantum Weighting Scheme; AUG Weighting Technique

http://dx.doi.org/10.6180/jase.202501_28(1).0018

1. Introduction

The calculation of structural similarity of molecules plays a key role in many applications in chemo-informatics such as QSAR models, docking and similarity-based virtual screening. The similarity searching is one of the tools for LBVS. The two-dimensional (2D) fingerprints are widely used by most of similarity methods to evaluate the chemical structure similarity against reference structure [1–3]. The fundamental concept underlying similarity searching is encapsulated by the similar property principle. This principle posits that molecules exhibiting structural resemblance are likely to possess analogous physiochemical and biological characteristics [1, 4]. The burgeoning significance of similarity searching applications can be primarily attributed to their pivotal role in the enhancement of lead optimization within drug discovery initiatives. In this context, the process entails the identification of the closest analogs to a primary lead compound, with the objective of discovering more efficacious compounds [5].

Over the years, several studies have introduced various similarity searching methods for chemical compounds [6–9]. The methods of similarity searching in LBVS can be used to find the similarity between large numbers of

molecules in chemical library and user-defined molecular structures using similarity coefficients such as Tanimoto [4]. The Tanimoto similarity coefficient can be considered as a benchmark similarity measure of chemical structure compounds. On the other hand, some studies proposed alternative similarity searching models for LBVS such as using Bayesian inference network [10], and using quantum concepts to develop quantum-based similarity method [11]. Other studies employed various techniques to further optimize the measure of similarity between molecules such as data fusion [12], nearest-neighbor information [13], and feature selection and weighting function [14, 15].

In the realm of Ligand-Based Virtual Screening, the efficacy of similarity searching methodologies is contingent upon three principal elements. Firstly, molecular descriptors, which encapsulate the attributes of molecules either in binary or numerical formats, are pivotal. These descriptors commonly manifest in the form of one-dimensional (1-D), two-dimensional (2-D), or three-dimensional (3D) fingerprints. Secondly, the similarity coefficient significantly influences the effectiveness of these methods. This coefficient is a numerical measure quantifying the extent of resemblance between a molecule in the database and a reference molecule [16]. Lastly, the relative significance of molecular fragments is crucial. Each chemical compound encompasses a vast array of data, inclusive of both redundant and non-essential features. The significance of these molecular fragments varies, which is typically represented through weighting techniques. In this context, a molecular fragment with a higher weight in both the database and reference structure is indicative of a greater degree of similarity compared to a fragment with a lower weight.

In the present investigation, a novel weighting technique has been implemented to enhance the efficacy of quantum-based similarity searching methodologies within Ligand-Based Virtual Screening (LBVS). The operational efficiency of the Quantum Mechanics-Augmented (QM-AUG) model is reliant on the Augmented Normalized Term Frequency (AUG) fragment-weighting function. This function is instrumental in distinguishing between various molecular fragments, focusing on their significance in ascertaining the similarity of a given molecule with others. This is achieved by attributing greater weights to certain molecular fragments during the computation of similarity. The AUG weighting function has been explored for its potential to recalibrate the weighted Hilbert space, thereby adjusting the fragment weights within molecular compounds. The QM-AUG method underwent rigorous evaluation using prominent chemo-informatics datasets, which were transformed into ECFC_4 2D-fingerprints utilizing Pipeline Pilot's software. The datasets employed for this purpose comprised the MDL Drug Data Report (MDDR) and the Maximum Unbiased Validation (MUV) datasets.

2. Related works

There are similarities between information retrieval and chemical virtual screening. In both cases, only a small percentage of the database is likely to be relevant to the user's query. Similarly, in a chemical database, only a few molecules are likely to have the same bioactivity as the reference structure. This analogy between relevance and bioactivity means that performance measures used to evaluate information retrieval systems, based on the number of relevant and non-relevant documents retrieved, can also be applied to systems for virtual screening by evaluating the number of active and inactive molecules retrieved [17, 18].

The field of text information retrieval employs two types of weighting schemes: *tf* (term frequency) and *idf* (inverse document frequency). In chemo-informatics, *tf* weighting assumes that molecules sharing multiple occurrences of a fragment are more similar than those sharing only a single occurrence, while *idf* weighting assumes that molecules sharing a rarely occurring fragment in the database are more similar than those sharing a commonly occurring fragment. Several studies have investigated *tf* weighting scheme, with Willett and Winterman [19] reporting that occurrence-based fingerprints were superior to incidence-based fingerprints in property prediction experiments on small QSPR and QSAR datasets. In contrast, Moock et al. [20] discovered that using *idf* weighting was successful in searching for similarities within a big chemical database. Abdo and Salim [21] integrated inverse frequency counts in their research on Bayesian inference networks for similarity search virtual screening, using them to calculate probabilities of bioactivity.

In similarity approaches used in molecular research, it is usually assumed that molecular fragments that do not have any connection to biological activity have the same level of importance as the vital ones. However, certain fragments, such as functional groups, are viewed as more significant by chemists. As a result, researchers analyze the weight of each fragment in a compound's chemical structure and assign more weight to the more important ones. Consequently, a match between two molecules based on highly weighted features contributes more to overall similarity than a match based on less significant features, according to Klinger and Austin [22] and Arif et al. [18].

Abdo and Salim [21] explored various weighting functions and introduced a fresh method for fragment weight-

ing in the Bayesian inference network used in ligand-based virtual screening. Meanwhile, Ahmed et al. [14] introduced a fragment reweighting approach that employs relevance feedback and reweighting factors to improve the retrieval recall capability of the Bayesian inference network. According to [23] selecting features can enhance the ability to detect similarities and prioritize important fragments while disregarding insignificant ones. Recently, Nasser et al. [15] employed Deep Belief Network (DBN) to reweight molecule features and improve the performance of virtual screening in identifying molecules with the greatest probabilities of activity. They investigated the importance of different molecular features and developed a method to give more weight to the important ones based on DBN, which used to calculate reconstruction feature error for all features, and then used Principal Component Analysis (PCA) to reduce the dimensions and select the features with the lowest error rates. The results showed that DBN outperformed other similarity methods in structurally heterogeneous data sets. Overall, the study aims to improve the performance of VS by selecting and giving more weight to important features.

3. Materials and methods

3.1. Standard Quantum Similarity method (SQB)

The evaluation of similarity between a chemical library and reference entities entails two fundamental components: the mathematical representation and the similarity coefficient. The mathematical representation encapsulates critical chemical information, exemplified through vectors and graphs. On the other hand, the similarity coefficients, which align with the mathematical forms, can be categorized into four types: associative, correlation, distance, and probabilistic coefficients [7, 24, 25]. In the pursuit of developing a suite of quantum-based similarity techniques within LBVS, researchers have delved into the mathematical portrayal of molecular compounds, aligning them with quantum mechanics principles. These quantum-based similarity methodologies derive their framework from the quantum probability formalism, a geometric expansion of conventional probability theory that incorporates elements such as Hilbert space, subspaces, and unit vectors. A distinctive aspect of Standard Quantum-Based (SQB) similarity methods, whether predicated on a complex or real Hilbert space, is their reliance on pure Hilbert space. The SQB approach operationalizes quantum mechanics on dual fronts: initially, by establishing a quantum framework in virtual screening through the depiction of molecular compounds and references in a complex Hilbert space, utilizing three proposed embedded techniques; subsequently, by the introduction of a novel similarity method, SQB, anchored in quantum model components [11]. In this research, the AUG weighting scheme has been adapted to synchronize with the SQB methodology. This integration leads to the formation of a weighted quantum space, which is postulated to enhance the performance of the similarity method. This aspect will be further elaborated in the subsequent section.

3.2. QM-AUG Quantum Model

The use of quantum probability was employed in the text information retrieval [26, 27]. There are several analogies between textual information retrieval and chemoinformatics [17]. The AUG (Augmented normalized term frequency) is a weighting scheme used in text information retrieval to give more weight to terms that appear frequently in a document, while discounting the effect of terms that appear frequently in all documents. The formula for AUG is given as:

$$AUG = \frac{\left(0.5 + 0.5 \times \left(\frac{tf}{\max tf}\right)\right)}{\left(1 + 0.5 \times \left(\frac{tf}{\max tf}\right)\right)} \quad (1)$$

In this context, tf denotes the raw term frequency, which refers to the frequency of occurrence of a given term within a document. Concurrently, max tf signifies the highest raw term frequency observed for any term present in the document. The constants 0.5 and 1 are employed in this framework as normalization factors, serving to standardize the values [28].

Abdo and Salim [21] leveraged the similarities between text and chemical retrieval to adapt the AUG weighting scheme for use in ligand-based virtual screening. They examined the AUG weighting function on the Bayesian inference network to enhance the retrieval effectiveness when searching chemical databases.

Hilbert space is a mathematical space used in quantum theory, which expands on the idea of Euclidean space. In order to overcome the challenges of chemical retrieval, the quantum probabilistic formalism uses a multi-dimensional representation of molecules, with probabilistic events represented as subspaces spanned by basis vectors. The components of this space can be analyzed through the geometry of chemical information space. The mathematics of Hilbert space is central to quantum theory, and the representation of molecular compounds within the space is critical for measuring the similarity of molecules. Dirac notation (bra-ket notation) is used in dealing with Hilbert space, where a vector is represented as Ket, and the transpose of Ket is represented as Bra [29]. Probability distribution over the subspace in Hilbert can be expressed using the Density operator, which is represented in its simplest form as a

one-dimensional projector. For instance, if we consider two orthonormal bases of space Ω, where the $\|v\| = \|v\| = 1$, the corresponding elements of each basis correspond to different fragments of molecules.

The QM-AUG method employs quantum framework components to calculate molecule probability. In traditional ligand-based screening similarity searching methods, it is generally assumed that fragments of molecules that are not associated with biological activity hold the same weight as the important ones. However, chemists may consider certain chemical fragments, such as functional groups in chemical structure diagrams, as more important than others. Therefore, a weight can be assigned to each chemical fragment according to its degree of importance. Consequently, a highly weighted fragment in a match between a target structure and a database structure would contribute more to the overall similarity than a less important one. A new fragment weighting scheme was previously introduced that employed a molecular similarity searching method based on the Bayesian model [21]. In contrast, this study utilized the AUG weighting function to enhance the chemical-information space. Proper fragment weighting can significantly improve the performance of the pure space. The QM-AUG model was developed by integrating modified weighting schemes and the *pure* Hilbert space of the Standard Quantum-Based model to generate a new weighted space for the QM-AUG method. The modified Weighting Functions (WF) of molecular fragments can be given by

$$WF_{AUG}(f_i) = \left(0.5 + 0.5 \times \frac{ff_{ij}}{\max ff_j}\right) \times \frac{\log\left[\frac{m+0.5}{cf_i}\right]}{\log(m+1)} \times \frac{\min\left(ff_{ij}, ff_{ir}\right)}{\max\left(ff_{ij}, ff_{ir}\right)} \quad (2)$$

In this formulation, ff_{ij} and ff_{ir} represent the frequencies of the specific fragment within a given compound and reference structure, respectively. cf_i denotes the count of compounds that contain the containing i^{th} fragment. The $\left|c_j\right|$ signifies the size of the compound in question, measured in terms of the quantity of fragments it comprises. $\left|c_{\text{avg}}\right|$ indicates the average size, calculated based on the number of fragments, of all compounds contained within the database. Lastly, m refers to the aggregate number of compounds present in the database.

Two various types of weighting-techniques for molecular fragments have been presented in the equations above namely, local, and global weight. The local weight refers to how many times particular fragment occurs in a compound or target structure, the local fragment weight represents in the first term of the equations. In contrast, the global weight refers to how many times particular fragment occurs in the whole chemical compounds database, the global fragment weight appears in the second term of the equations. The third part introduced by our research group is integrated with local and global weight [21]. To enhance the chemical space, we proposed the addition of the weight of each fragment obtained by fragment weighting techniques to the original weights of all database compounds, while reference fragments are added based on the number of times they occur in the entire compound collection. Consequently, a new quantum space of molecular compounds is generated which plays a vital role when calculating the similarity of molecules. The ligand reweighted for collection and references compounds generated a new Hilbert space for QM-AUG method. The weighted Hilbert space created by integration of AUG weighting scheme with the pure Hilbert space for all chemical compounds in library. In contrast, the weighted Hilbert space for references compounds created by integration of the pure Hilbert space with global weighting function, which refers to how many times that molecular fragment occurs in the entire compound collection, as depicted in Fig. 1. The use of weighting function to

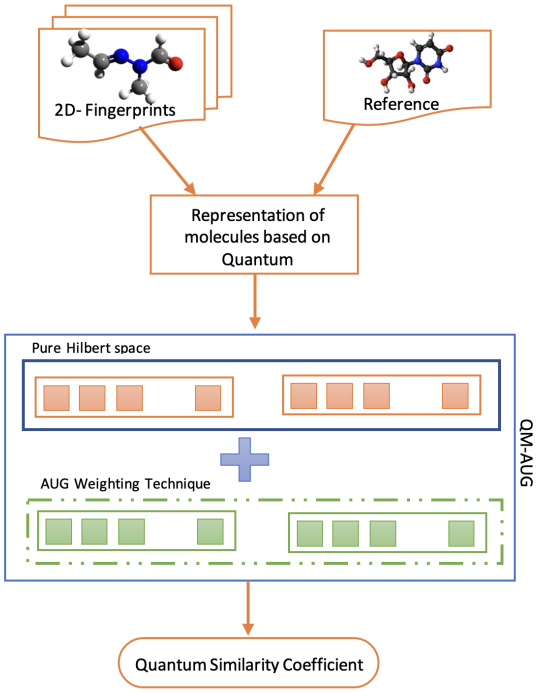

Fig. 1. Proposed QM-AUG Model

re-weight molecular fragments of compounds with quantum space led to the development of QM-AUG similarity method. The weighted spaces can be used to find the sim-

ilarity of molecules based on quantum concepts. Let us assume the probabilistic event of molecule in Hilbert space $\mathcal{H}$ is represented as a subspace S_m. A probability measure denoted as μ can be initially established for a weighted chemical-information space, which is conceptualized as a unit vector φ. This is achieved by calculating the square of the magnitude of the projection of vector φ onto a designated molecular subspace S_m. The computation of this probability is executed according to the following mathematical formulation:

$$\mu\left(S_m\right) = \mathrm{tr}(d\widehat{S}) = \|\widehat{S}\varphi\|^2 \tag{3}$$

where tr is trace operator, and $d = \varphi_i\varphi_i^{\dagger}$ is named density operator. Generally, an operator that exhibits both positive-semi-definite properties (implying $v^{\dagger}pv \geq 0$ for any vector v) and possesses a trace of one is utilized to define the probability distribution over the respective subspaces [26].

3.3. Experimental Datasets

The screening experiments in this study were conducted using two distinct databases: the *MDL Drug Data Report (MDDR)* [30] and Maximum Unbiased Validation (MUV) [31]. Each molecule within these databases was transformed into Pipeline Pilot ECFC_4 (extended connectivity fingerprints, subsequently folded to a dimensionality of 1024 bits) [32]. The MDDR dataset comprises a total of 102,516 molecules, segmented into three distinct subsets: MDDR-DS1, MDDR-DS2, and MDDR-DS3. The MDDR-DS1 subset encompasses 11 activity classes, characterized by a mix of structural homogeneity and heterogeneity among their active compounds. In contrast, the MDDR-DS2 subset is composed of 10 homogeneous activity classes, whereas MDDR-DS3 contains 10 heterogeneous activity classes. Comprehensive details pertaining to these subsets are presented in Tables 1 to 3. Each row in these tables delineates an activity class, enumerating the count of molecules associated with it, along with the class's diversity. This diversity metric was calculated as the mean pairwise Tanimoto similarity, derived from all molecule pairs within a class, utilizing the ECFC_4 fingerprints.

The MUV dataset is comprised of 17 distinct activity classes, each encompassing up to 30 active compounds and 15,000 inactive ones. An analysis of the diversity within this dataset reveals a predominance of highly diverse or more heterogeneous activity classes. Detailed information regarding the MUV dataset is delineated in Table 4. For each of the datasets - MDDR-DS1, MDDR-DS2, MDDR-DS3, and MUV - the screening experiments were executed using ten reference structures. These structures were randomly selected from each activity class. Subsequently, these structures were unified and employed in the assessment of three similarity methods: Tanimoto coefficient (TAN), Binary Augmented (BIN-AUG), and the proposed Quantum Mechanics-Augmented (QM-AUG) method. The effectiveness of these methods was gauged based on the recall results of active compounds retrieved. These results were averaged across each dataset for all active molecules, particularly focusing on the top 1% and 5% of the ranked list generated from the similarity searches.

4. Results and discussion

The results for the searches of MDDR-DS1, MDDR-DS2, MDDR-DS3, and MUV are shown in Tables 5 to 8 respectively, the left-hand part of each table reported the results of top 1%, while the right-hand part contains the results of top 5%. The results of QM-AUG, SQB TAN, and BIN-AUG are presented in these tables for both cut-offs. The AUG weighting scheme was applied in this study to improve the SQB method, which was previously introduced based on the pure Hilbert space [11] and was also applied to the Bayesian Inference Network model by Abdo and Salim [21]. In addition, the results were compared with the benchmark of the standard Tanimoto coefficient (TAN). Each row in each table corresponds to one activity class, with the best recall value highlighted in each row. The best recall values and the means for all activity classes from MDDR-DS1, MDDR-DS2, MDDR-DS3, and MUV are presented in Figs. 2 to 5, using cutoffs of both 1% and 5%. The highest values are highlighted in a distinct color.

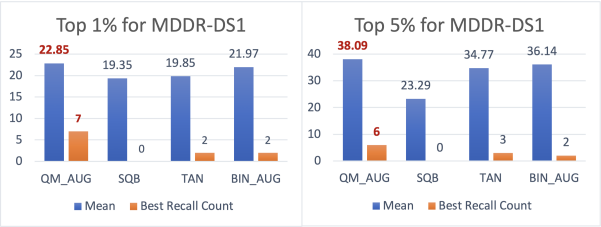

Fig. 2. Comparative Analysis of Mean Recall and Best Recall Instances for Top 1% and 5% for MDDR-DS1 Dataset

The recall values for MDDR-DS1, documented in Table 5 for cutoffs of 1% and 5%, reveal a notable superiority of the QM_AUG approach over the benchmark TAN method and other existing methodologies. When comparing different strategies, the AUG weighting scheme, integrated with a quantum mechanics (QM) model, consistently outperforms both the standard quantum-based model (SQB) and the AUG combined with the Bayesian Inference Network model (BIN_AUG) in terms of retrieval recall results.

Table 1. MDDR-DS1 structure activity classes

Activity Class	Activity Index	Active molecules	Pairwise similarity
Renin inhibitors	31420	1130	0.290
HIV protease inhibitors	71523	750	0.198
Thrombin inhibitors	37110	803	0.180
Angiotensin II AT1 antagonists	31432	943	0.229
Substance P antagonists	42731	1246	0.149
Substance P antagonists	06233	752	0.140
5HT reuptake inhibitors	06245	359	0.122
D2 antagonists	07701	395	0.138
5HT1A agonists	06235	827	0.133
Protein kinase C inhibitors	78374	453	0.120
Cyclooxygenase inhibitors	78331	636	0.108

Table 2. MDDR-DS2 structure activity classes

Activity Class	Activity Index	Active molecules	Pairwise similarity
Adenosine (A1) agonists	07707	207	0.229
Adenosine (A2) agonists	07708	156	0.305
Renin inhibitors 1	31420	1300	0.290
CCK agonists	42710	111	0.361
Monocyclic lactams	64100	1346	0.336
Cephalosporins	64200	113	0.322
Carbacephems	64220	1051	0.269
Carbapenems	64500	126	0.260
Tribactams	64350	388	0.305
Vitamin D analogous	75755	455	0.386

Table 3. MDDR-DS3 structure activity classes

Activity Class	Activity Index	Active molecules	Pairwise similarity
Muscarinic (M1) agonists	09249	900	0.111
NMDA receptor antagonists	12455	1400	0.098
Nitric oxide synthase inhibitors	12464	505	0.102
Dopamine -hydroxylase inhibitors	31281	106	0.125
Aldose reductase inhibitors	43210	957	0.119
Reverse transcriptase inhibitors	71522	700	0.103
Aromatase inhibitors	75721	636	0.110
Cyclooxygenase inhibitors	78331	636	0.108
Phospholipase A2 inhibitors	78348	617	0.123
Lipoxygenase inhibitors	78351	2111	0.113

Fig. 2 delineates a methodological comparison, highlighting both mean and peak retrieval values for each activity class within MDDR-DS1 across both cutoff thresholds. Analyzing the top 1% cutoff, QM_AUG emerges as the leading approach, delivering the most optimal retrieval recall results in 7 out of 11 activity classes, whereas both TAN and BIN_AUG achieve this distinction in just 4 classes, split equally between the two methods. Moving to the 5% threshold, QM_AUG dominates in 6 classes, with TAN excelling in 3, and BIN_AUG in 2. When considering average values, QM_AUG stands out, registering scores of 22.85 and 38.09 for the 1% and 5% cutoffs, respectively.

In Tables 6 and 7, recall values are recorded for homogenous and heterogeneous datasets, respectively, according to specified cutoffs. Upon examining the homogenous MDDR-DS2 dataset, the introduced QM_AUG method manifests a clear advantage over both the TAN and SQB methods. Specifically, QM_AUG achieves a mean recall value of 81.03 at the 1% cutoff and 93.94 at the 5% threshold. In contrast, the SQB method, which exhibits performance metrics closely aligned with the TAN method, records 62.41 for the top 1% and 76.53 for the top 5%. Furthermore, as depicted in Fig. 3, QM_AUG leads in a majority of the activity classes. While the performance of QM_AUG closely

Table 4. MUV structure activity classes

Activity Class	Activity Index	Pairwise similarity
S1P1 rec. (agonists)	466	0.117
PKA (inhibitors)	548	0.128
SF1 (inhibitors)	600	0.123
Rho-Kinase2 (inhibitors)	644	0.122
HIV RT-RNase (inhibitors)	652	0.099
Eph rec. A4 (inhibitors)	689	0.113
SF1 (agonists)	692	0.114
HSP 90 (inhibitors) 30	712	0.106
ER-a-Coact. Bind. (inhibitors)	713	0.113
ER-b-Coact. Bind. (inhibitors)	733	0.114
ER-a-Coact. Bind. (potentiators)	737	0.129
FAK (inhibitors)	810	0.107
Cathepsin G (inhibitors)	832	0.151
FXIa (inhibitors)	846	0.161
FXIIa (inhibitors)	852	0.150
D1 rec. (allosteric modulators)	858	0.111
M1 rec. (allosteric inhibitors)	859	0.126

Table 5. Outcomes of retrieval for Top 1% and 5% within the MDDR-DS1 dataset

Activity Index	QM-AUG	SQB	TAN	BIN-AUG	QM-AUG	SQB	TAN	BIN-AUG
	1%				5%			
31420	76.17	70.03	69.69	74.82	87.44	75.97	83.49	87.48
71523	28.52	25.58	25.94	26.8	55.38	29.76	48.92	51.44
37110	22.98	9	9.63	24.03	44.75	22.41	21.01	49.53
31432	38.74	37.34	35.82	39.63	76.07	38.6	74.29	76.05
42731	22.15	17.34	17.77	20.68	31.14	22.9	29.68	25.55
06233	13.65	10.75	13.87	12.57	23.81	14.53	27.68	20.36
06245	6.82	6.03	6.51	6.15	16.09	6.84	16.54	12.74
07701	11.75	8.25	8.63	11.45	28.93	11.55	24.09	26.35
06235	11.07	9.14	9.71	9.48	24.09	11.66	20.06	21.44
78374	13.65	13.65	13.69	10.86	20.64	15.46	20.51	17.7
78331	5.91	5.78	7.17	5.23	10.74	6.54	16.2	8.93

Table 6. Outcomes of retrieval for Top 1% and 5% within the MDDR-DS2 dataset

Activity Index	QM-AUG	SQB	TAN	BIN-AUG	QM-AUG	SQB	TAN	BIN-AUG
	1%				5%			
07707	73.91	58.5	61.84	72.67	74.56	70.39	70.39	74.81
07708	97.55	55.61	47.03	98.13	100	64.97	56.58	99.61
31420	81.28	62.22	65.1	77.58	94.98	87.04	88.19	94.63
42710	74.64	83	81.27	73.64	95.18	89.18	88.09	92.45
64100	88.57	80.73	80.31	90.2	99.19	94.59	93.75	98.25
64200	68.48	53.13	53.84	68.48	99.02	81.34	77.68	97.29
64220	65.39	34.61	38.64	66.94	91.49	48.11	52.19	91.29
64500	81.12	29.04	30.56	82.32	95.12	47.68	44.8	96.16
64350	81.63	81.86	80.18	80.57	91.65	87.96	91.71	91.87
75755	97.73	85.4	87.56	97.67	98.26	94.07	94.82	98.24

parallels that of BIN_AUG, the latter method stands out in five activity classes, achieving a mean recall of 80.82 at the 1% cutoff. However, for the 5% threshold, BIN_AUG emerges as the frontrunner in only three classes.

In contrast, analyzing the heterogeneous MDDR_DS3 dataset, the proposed QM_AUG method marginally surpasses the TAN in mean recall values at the 1% cutoff. However, for the 5% threshold, the TAN exhibits a slight

Table 7. Outcomes of retrieval for Top 1% and 5% within the MDDR-DS3 dataset

Activity Index	QM-AUG	SQB	TAN	BIN-AUG	QM-AUG	SQB	TAN	BIN-AUG
	1%				5%			
09249	11.69	9.92	12.12	8.81	20.14	21.4	24.17	15.44
12455	6.75	5.12	6.57	6.53	9.51	8.1	10.29	9.56
12464	7.28	5.56	8.17	5.67	18.69	10.56	15.22	15.48
31281	15.71	10.29	16.95	15.9	24.1	15.14	29.62	24.19
43210	6.58	5.31	6.27	5.97	13.5	14.47	16.07	12.06
71522	6.67	3.03	3.75	6.18	14.21	9.2	12.37	11.97
75721	19.83	15.24	17.32	19.17	28.17	22.27	25.21	28.88
78331	5.34	5.48	6.31	5.06	10.38	12.03	15.01	9.89
78348	9.37	9.67	10.15	7.01	22.55	22.72	24.67	16.46
78351	12.32	10.03	9.84	12.05	12.91	11.95	11.71	12.85

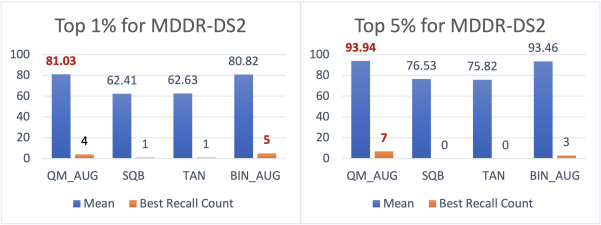

Fig. 3. Comparative Analysis of Mean Recall and Best Recall Instances for Top 1% and 5% for MDDR-DS2 Dataset

performance edge over QM_AUG, registering a mean of 18.4. In terms of both mean and the tally of peak recall values at the 5% cutoff, as illustrated in Fig. 4, QM_AUG clearly outshines both the SQB and BIN_AUG methods. Additionally, it is noteworthy that both the QM_AUG and TAN methods attain top results in five distinct activity classes at the 1% cutoff.

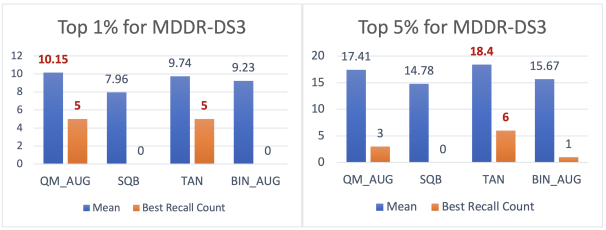

Fig. 4. Comparative Analysis of Mean Recall and Best Recall Instances for Top 1% and % for MDDR-DS3 Dataset

In the examination of the MUV dataset, as delineated in Table 8, the outcomes for the proposed methodology at both 1% and 5% were showcased. The results revealed a marginal distinction across the methods, indicating that no method demonstrated unequivocal dominance over the others. Specifically, at the 1% threshold, the introduced technique displayed marginally enhanced performance, identifying eight activity classes out of 17, with an average of 4.5. This average was on par with the TAN method and bore proximity to BIN_AUG. It's noteworthy that the QM_AUG results surpassed those of the standard QM (SQB) model. Conversely, when assessed at the 5% threshold, the SQB model emerged as superior, even though the mean results of all models closely aligned. The employment of the AUG weighting approach on both the QM and BIN models yielded optimal results in four distinct classes for each. In terms of activity class performance, SQB and TAN were predominant in 8 and 6 classes, respectively, as shown in Fig. 5.

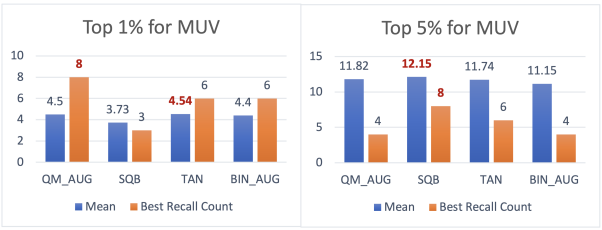

Fig. 5. Comparative Analysis of Mean Recall and Best Recall Instances for Top 1% and 5% for MUV Dataset

The principal aim of the advocated method is to ascertain the efficacy of the quantum-based model in leveraging the AUG weighting function, subsequently determining retrieval efficiency. A visual assessment of recall metrics serves as a tool for contrasting the proficiency of assorted search algorithms. Nonetheless, the Kendall W test of concordance offers a more numerical metric for analysis [33]. This specific test ascertains the consistency of judgments across a set of evaluators when ranking various entities. In this context, the activity classes function as the evaluators, while the recall rates of the diverse weighting functions act as the entities being ranked. The test yields key values:

Table 8. Outcomes of retrieval for Top 1% and 5% within the MUV dataset

Activity Index	QM-AUG	SQB	TAN	BIN-AUG	QM-AUG	SQB	TAN	BIN-AUG
	1%				5%			
466	1.38	2.41	3.1	1.38	7.24	6.9	5.86	6.55
548	9.31	7.59	8.62	8.62	23.79	21.03	22.76	20.69
600	5.17	2.41	3.79	5.17	15.17	10.34	11.38	15.17
644	7.24	7.24	7.59	6.9	17.24	17.24	17.59	14.83
652	3.45	2.07	2.76	3.45	8.97	8.62	7.93	9.31
689	4.14	2.07	3.79	3.79	9.31	8.28	9.66	9.66
692	1.03	0.69	0.69	1.38	4.48	6.21	4.83	3.79
712	3.45	4.14	4.14	2.76	11.03	16.9	10.34	10.34
713	2.76	2.41	3.1	2.76	6.21	7.24	7.24	6.55
733	3.79	1.38	3.45	3.45	7.59	8.97	8.97	8.97
737	1.72	1.38	2.41	1.72	8.62	12.41	8.28	8.62
810	2.41	2.41	2.07	2.07	10	10.34	6.9	8.97
832	6.55	4.48	6.55	6.21	14.48	11.38	13.1	13.45
846	11.38	8.97	9.66	12.07	24.48	23.45	28.62	24.14
852	8.97	8.62	12.41	8.62	19.31	18.62	21.38	16.55
858	1.72	3.1	1.72	1.72	4.14	7.93	5.86	4.83
859	2.07	2.07	1.38	2.76	8.97	10.69	8.97	7.24

the Kendall coefficient (W), Chi-Square (X2), and the corresponding significance level (P). The P-value reflects the likelihood of the coefficient's value arising fortuitously. If this value is deemed significant (using thresholds of 0.01 or 0.05) it permits the comprehensive ranking of the evaluated entities. The outcomes of the Kendall test for datasets MDDR(DS1-DS3) and MUV are delineated in Table 9, incorporating both the 1% and 5% cutoffs. While Fig. 5 depicted the performance-based Ranking of Methods Insights from the Kendall W Test.

In the evaluation of MDDR and MUV datasets utilizing the Kendall W concordance test, the AUG_QM method's ranking emerges prominently, notably surpassing its predecessors SQB, TAN, and BIN_AUG at the 1% cutoff. Upon reviewing the datasets, it's observed that, for the top 1% threshold, the significance test (P) values consistently fell below 0.05. This underscores the significance of the proposed method across all scenarios with a 1% cutoff. Consequently, a holistic assessment of the methodologies denotes the supremacy of the AUG_QM approach over earlier studies, notably the benchmark TAN. A comprehensive ranking reveals the preeminence of the proposed method across DS1, DS2, DS3, and MUV datasets, as visualized in Fig. 6. Contrarily, data assessed at the 5% cutoff exhibits more variance. In this bracket, AUG-QM stands out in MDDR-DS1 and MDDR-DS2, with P values below 0.05. Yet, TAN and SQB respectively lead in MDDR-DS3 and MUV, yielding P values of 0.118 and 0.512, both surpassing the 0.05 threshold. Notably, AUG-QM outperforms BIN-AUG in these instances. In summation, of the eight scenarios evaluated, AUG-QM clinched the top spot in six, settling for second in the remaining two.

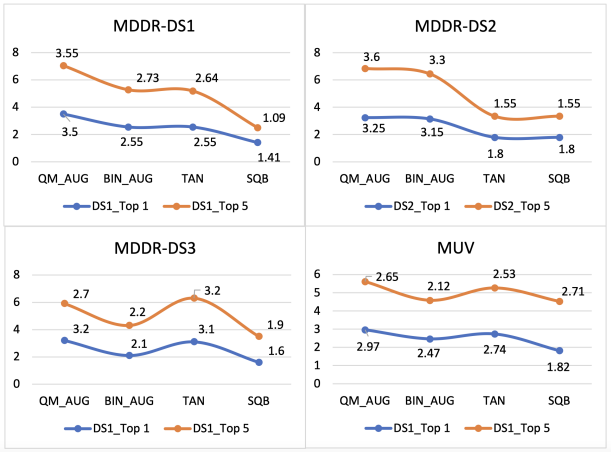

Fig. 6. Evaluating and Ranking Methodological Performance through Kendall's W Analysis

5. Conclusion

Various similarity search methodologies often presuppose that molecular structural attributes, irrespective of their relevance to biological activity, possess equivalent significance to the pivotal features. The relative prominence among disparate fragments of a chemical compound can be demarcated through a strategic weighting framework. This study delves into the ramifications of recalibrating weights assigned to chemical fragments, aiming to optimize the performance of the quantum model in ligand-based virtual screening. Our research introduces a novel

Table 9. Raking of similarity methods QM-AUG,SQB, TAN, and BIN-AUG using Kendall W Test results for MDDR (DS1-DS3) and MUV at the top 1% and 5%

Dataset	Cutoffs	W	X^2	P	Rank Methods
MDDR-DS1	1%	0.44	14.61	0.002	QM_AUG > BIN_AUG > TAN > SQB
	5%	0.63	20.78	0.000117	QM_AUG > BIN_AUG > TAN > SQB
MDDR-DS2	1%	0.39	11.9	0.008	QM_AUG > BIN_AUG > SQB > TAN
	5%	0.73	22.15	0.000061	QM_AUG > BIN_AUG > SQB > TAN
MDDR-DS3	1%	0.36	10.92	0.012	QM_AUG > TAN > BIN_AUG > SQB
	5%	0.19	5.88	0.118	TAN > QM_AUG > BIN_AUG > SQB
MUV	1%	0.16	8.5	0.037	QM_AUG > TAN > BIN_AUG > SQB
	5%	0.045	2.3	0.512	SQB > QM_AUG > TAN > BIN_AUG

quantum-centric similarity methodology, denominated as the QM-AUG method, that recalibrates molecular fragments leveraging the AUG weighting mechanism. This integration with the quantum similarity model has culminated in a newly weighted Hilbert molecular space, designed to amplify the efficacy of retrieval outcomes inherent to the conventional quantum similarity SQB approach. Empirical evidence underscores that the AUG weighting methodology markedly elevates retrieval effectiveness, outperforming its application in the Bayesian Inference Network and the standard Taninmoto coefficient. The findings spotlight QM-AUG's superior performance in MDDR-DS1 and MDDR-DS2 across both threshold limits. For the more diverse MDDR-DS3 dataset, QM-AUG outpaced its counterparts at the 1% cutoff, albeit marginally trailing behind TAN at the 5% threshold. Conversely, MUV dataset results bore alignment with extant literature, mirroring TAN benchmark across both evaluated thresholds.

References

[1] P. Willett, (2009) *"Similarity methods in chemoinformatics"* **Annual Review of Information Science and Technology *43***(1): 1–117. DOI: 10.1002/aris.2009.1440430108.

[2] R. P. Sheridan, (2007) *"Chemical similarity searches: when is complexity justified?"* **Expert opinion on drug discovery 2**(4): 423–430.

[3] L. Y. E. Ekaney, D. B. Eni, and F. Ntie-Kang, (2021) *"Chemical similarity methods for analyzing secondary metabolite structures"* **Physical Sciences Reviews *6***: 247–264. DOI: 10.1515/psr-2018-0129.

[4] M. A. Johnson and G. M. Maggiora, (1990) *"Concepts and applications of molecular similarity"* **John Wiley Sons: New York,NY,USA**:

[5] S.-Q. Yang, Q. Ye, J.-J. Ding, M.-Z. Yin, A.-P. Lu, X. Chen, T.-J. Hou, and D.-S. Cao, (2021) *"Current advances in ligand-based target prediction"* **Wiley Interdisciplinary Reviews: Computational Molecular Science *11***(3): e1504.

[6] A. Bender, H. Y. Mussa, R. C. Glen, and S. Reiling, (2004) *"Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier"* **Journal of Chemical Information and Computer Sciences *44***(1): DOI: 10.1021/ci034207y.

[7] A. Maldonado, J. P. Doucet, M. Petitjean, and B.-T. Fan, (2006) *"Molecular similarity and diversity in chemoinformatics: From theory to applications"* **Molecular Diversity *10***(1): 39–79. DOI: 10.1007/s11030-006-8697-1.

[8] A. Abdo and M. Pupin, (2021) *"LINGO-DL: a text-based approach for molecular similarity searching"* **Journal of Computer-Aided Molecular Design *35***(5): DOI: 10.1007/s10822-021-00383-9.

[9] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, (2014) *"Molecular similarity in medicinal chemistry"* **J Med Chem *57***: DOI: 10.1021/jm401411z.

[10] A. Abdo, B. Chen, C. Mueller, N. Salim, and P. Willett, (2010) *"Ligand-Based Virtual Screening Using Bayesian Networks"* **Journal of Chemical Information and Modeling *50***(6): 1012–1020. DOI: 10.1021/ci100090p.

[11] M. M. Al-Dabbagh, N. Salim, M. Himmat, A. Ahmed, and F. Saeed, (2015) *"A quantum-based similarity method in virtual screening"* **Molecules *20***(10): 18107–18127. DOI: 10.3390/molecules201018107.

[12] P. Willett, (2006) *"Enhancing the Effectiveness of Ligand-Based Virtual Screening Using Data Fusion"* **QSAR Combinatorial Science *25***(12): 1143–1152. DOI: 10.1002/qsar.200610084.

[13] J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, (2005) *"Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information"* **Journal of medicinal chemistry *48***(22): 7049–7054.

[14] A. Ahmed, A. Abdo, and N. Salim, (2012) *"Ligand-based Virtual screening using Bayesian inference network and reweighted fragments"* **The Scientific World Journal**:

[15] M. Nasser, N. Salim, H. Hamza, F. Saeed, and I. Rabiu, (2020) *"Features Reweighting and Selection in ligand-based Virtual Screening for Molecular Similarity Searching Based on Deep Belief Networks"* **Advances in Data Science and Adaptive Analysis *12***(03n04): 2050009–2050009. DOI: 10.1142/s2424922x20500096.

[16] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett, (2012) *"Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets"* **Journal of Chemical Information and Modeling *52***(11): 2884–2901. DOI: 10.1021/ci300261r.

[17] P. Willett, (2000) *"Textual and chemical information processing: different domains but similar algorithms"* **Information Research *5***(2):

[18] S. M. Arif, J. D. Holliday, and P. Willett. "The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening". In: *Advances in Mathematical Chemistry and Applications: Revised Edition*. ***1***. Elsevier Inc., 2015, 92–112. DOI: 10.1016/B978-1-68108-198-4.50005-9.

[19] P. Willett and V. Winterman, (1986) *"A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity Measures of Inter-Molecular Structural Similarity"* **Quantitative Structure-Activity Relationships *5***(1): 18–25.

[20] T. E. Moock, D. L. Grier, W. D. Hounshell, G. Grethe, K. Cronin, J. G. Nourse, and J. Theodosiou, (1988) *"Similarity searching in the organic reaction domain"* **Tetrahedron Computer Methodology *1***(2): 117–128.

[21] A. Abdo and N. Salim, (2010) *"New fragment weighting scheme for the bayesian inference network in ligand-based virtual screening"* **Journal of chemical information and modeling *51***(1): 25–32.

[22] S. Klinger and J. Austin. "Weighted superstructures for chemical similarity searching". In: *Proceedings of the 9th Joint Conference on Information Sciences*. 2006.

[23] M. Vogt, A. M. Wassermann, and J. Bajorath, (2010) *"Application of information—Theoretic concepts in chemoinformatics"* **Information *1***(2): 60–73.

[24] J. D. Holliday, C. Hu, and P. Willett, (2002) *"Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings"* **Combinatorial chemistry high throughput screening *5***(2): 155–166.

[25] G. Maggiora and V. Shanmugasundaram. "Molecular Similarity Measures". In: *Chemoinformatics*. Ed. by J. Bajorath. ***275***. Methods in Molecular Biology™. Humana Press, 2004. Chap. 1, 1–50. DOI: 10.1385/1-59259-802-1:001.

[26] C. J. v. Rijsbergen. *The Geometry of Information Retrieval*. UK: Cambridge University Press, 2004.

[27] M. Melucci and K. van Rijsbergen. "Quantum mechanics and information retrieval". In: *Advanced topics in information retrieval*. Germany: Springer Berlin Heidelberg, 2011, 125–155.

[28] G. Salton and C. Buckley, (1988) *"Term-weighting approaches in automatic text retrieval"* **Information processing management *24***(5): 513–523.

[29] P. A. M. Dirac. *The principles of quantum mechanics*. Oxford university press, 1981.

[30] *MDL Drug Data Report (MDDR)*. Web Page.

[31] S. G. Rohrer and K. Baumann, (2009) *"Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data"* **Journal of chemical information and modeling *49***(2): 169–184.

[32] *Pipeline Pilot Software : SciTegic Accelrys Inc*. Computer Program.

[33] P. Legendre, (2005) *"Species associations: the Kendall coefficient of concordance revisited"* **Journal of agricultural, biological, and environmental statistics *10***(2): 226–245.