*Health Biotechnology and Biopharma (2023), 7(3): 62-76*

*Original Research Article*

# Introducing an innovative algorithm based on the abilities of the human perception system towards profiling similarity search of drugs and drug-like compounds such as fluoxetine

Setareh Ghavami [1,2], Ghazaleh Ghavami [1], Hassan Toozandehjani [2], Soroush Sardari [1,*]

[1]Drug Design and Bioinformatics Unit, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran; [2]Department of Psychology, School of Humanities, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran

***Corresponding author:** Soroush Sardari; Drug Design and Bioinformatics Unit, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran. E- mail: sardari@pasteur.ac.ir*

**DOI: 10.22034/HBB.2023.24**

**Received:** February 28, 2023; **Accepted:** July 15, 2023

## ABSTRACT

The current research had been finalized based on the employing interdisciplinary aspects in addition to utilizing skills of cheminformatics in addition to psychosomatic facts towards introducing the innovative Translator Machine System (TMS-CH) for converting small molecule structures to melodic composition and image color combination on the subject of completing inventive pattern and process for chemical structure similarity search. The percentage of true answers for the entire tests among 200 participants confirmed the ability of the proposed TMS-CH system for evaluating the matching level between tested structures. The perception of chemical structures via visual and auditory representations by bio-mimicking human systems during the current investigation could be an efficient strategy to develop findings related to ligand-based similarity search methods.

***Keywords*:** Virtual screening, similarity search, pattern recognition, drug-like compounds

## INTRODUCTION

Cheminformatics is the *in silico* technique that is widely utilized in pharmaceutical companies in the process of drug discovery in addition to numerous applications in chembio-computational research [1]. During the current investigation, the cheminformatics-based knowledge was combined with perception concepts related

to psychology science to introduce an innovative cheminformatics-based similarity search algorithm in the frame of intra science concept.

Virtual screening as one of the applicable aspects in the field of cheminformatics is all the time more utilized as a cost-effective approach to high-throughput screening that utilizes numerous computational techniques to highlight a certain population from the pool of small molecule structures and filter large chemical datasets to certify those structures have the prior probabilities regarding bioactivity and drug-like properties [2].

An example of ligand-based approaches as one of the major kinds of the virtual screening process is the similarity search methods. One of the first concepts underlying similarity-based virtual screening is introduced by Johnson and Maggiora as similar property theory states [3]. A basic virtual screening approach computes the symmetry between the identified template and every structure in a chemical library ranks the database compounds in diminishing instruction of the similar properties and then performs actual filtering on just the high-ranked database structures [4].

Similarity Search issue is one of the major issues in software engineering. It has various applications in content, video and picture recovery, pattern recognition, bioinformatics, web seek, fingerprint databases, and numerous others. In this issue, an example is given and the calculation looks through the database, or the Web, to restore all or most, contingent upon whether the hunt is correct or rough, of the information protests that are "close" to that example as indicated by some semantics of closeness [5]. In another variety of this issue, the calculation might be solicited to recover a predefined number from the nearest questions that example [5]. A gullible answer for this issue is to contrast the example and each question in the database and restore every one of the items that are like that example. This similitude is delineated utilizing an essential idea which is the comparability measure or its more grounded shape; the separation metric [5]. With the expansion of more affordable stockpiling units with an expanding putting away limit, the measure of information in current databases is large to the point that productivity and viability have turned into the principal issues in assessing the execution of various similarity search algorithms [5].

During current research to achieve the possibility for overcome pointed challenges regarding similarity search via biochemical databases, a novel similarity search algorithm was introduced for the process of recognizing patterns of chemical compound building blocks utilizing bio-mimicking visual and auditory capabilities of humans for recognition of input patterns towards investigating further novel advances in the field of chemical similarity search methodology based on the 2D structure of chemical compounds as a Simplified Molecular-Input Line-Entry System (SMILES).

## MATERIALS AND METHODS

### *Translational technique for altering the chemical structure of music and image*

Based on previously reported algorithms in addition to the concept of characters associated with SMILES, we cluster and encode each SMILES character with the aim of presenting an innovative similarity search algorithm for the advancement of chemical compound pattern recognition. Then, in addition to the classical note parameters (Do, Re, Mi, Fa, Sol, La, Si), in addition to two aspects such as the peak pattern, I presented a decisive melody theme for the SMILES character. In terms of sharp concepts, SMILES characters, in addition to the image color composition of chemical structures, had individual colors as points to express a specific melody composition.

### *Algorithm coding trend for presenting the translator machine system*

Using Microsoft Visual Studio 2013-12.021005.1 within the framework of the C# programming language, an innovative translation engine system called Translational Modern System-Chemicals (TMS-CH) was used to encode compound structures into melodic sounds in addition to color schemes.

Initially, two separate TMS-CHs were developed to convert protein structures into melodic compositions and pictorial color combinations, termed Translational Modern System Chemicals Music (TMS-CH-M) and Translational Modern System-Chemicals Image (TMS-CH-I). Additionally, the TMS-CH-M/TMS-CH-I Executive Edition has the ability to generate music and image formats as system output for ease of use with point systems.

### *Assortment and preparation of pattern sample and test chemical compounds*

In the beginning, to investigate a novel potential approach toward drug-based therapy of depression by introducing new active chemical molecules with higher activity and fewer side effects, fluoxetine was chosen as the pattern sample. Following that, the SMILES sequence of fluoxetine was explored based on a similarity structure search program via PubChem databases provided by the NCBI PubChem Structure Search homepage. As the outcome, two chemical structures with 70 % and 5 % similarity with the pattern sample were presented. In conclusion, the SMILES string of pattern sample in addition to both resulted in similar (similarity index 70 % and 5 %) structures as test sets were introduced to TMS-CH-M/TMS-CH-I for generating music and image formats for the entire pattern sample and test compounds.

***Measurement of the visual and auditory perceptions to use as a bio-mimic pattern towards constructing the innovative similarity search algorithm***

During current research, an innovative pattern recognition method has been introduced for compound similarity search based on the concept of biomimicry insight by modeling the human perceptual system in finding similarities. To achieve this goal, two human perceptual abilities as visual and auditory were assessed to identify similarities between samples and test sets in the form of images and music recognizable by the human perceptual system.

Initially, a statistical society with 200 participants, well-distributed by age, sex, education, and marital status, was nominated as the empirical population. Later, auditory and visual assessment methods were planned.

Regarding utilization of the human visual perception system, in addition to items in the test set that included the sample as particularly similar structures (100 % similarity index), a combination of fluoxetine-generated images was selected as a control group (sample Introduce compounds with 77 % and 5 % similarity to the pattern into the population evaluated in a uniform standard situation. Relevant tests were recorded by recipient subjects on a unique form with a specific code for each participant (codes were recognized for each participant on the registration form, consent form, and sensory assessment form).

Methods cited also include: It was supplemented for auditory cognitive assessment by presenting the generated melodic construct (less than 30 s) as a sharp condition for visual cognitive assessment.

Finally, the results of the evaluation test for both the perceptual rating based on one-way Anova and the test with GraphPad Prism 5.0.

***Statistical analysis***

The presented data were examined by GraphPad Prism 5.0 with a 95 % confidence interval and Microsoft Office Excel. Statistical comparisons were carried out by ANOVA and post hoc Tukey's test. The differences were considered significant when $p<0.05$.

## RESULTS

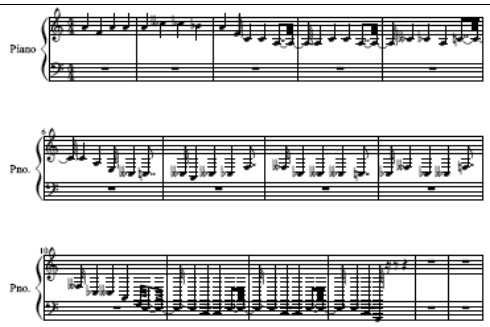
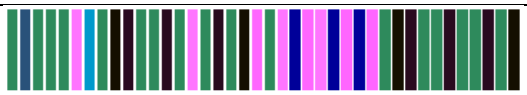
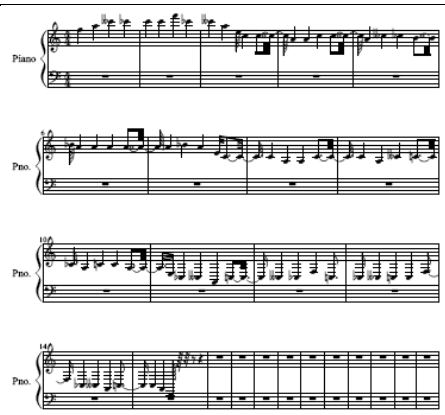
As the cited, in the current investigation, based each character is related to the SMILES string for chemical compounds (Table 1), the inventive pattern for clustering and coding of SMILES characters was examined to build the unique translator machine system named as TMS-CH for translating the chemical compound structure to melodic composition in addition to image color combination based on the pointed new translational technique (Table 2).

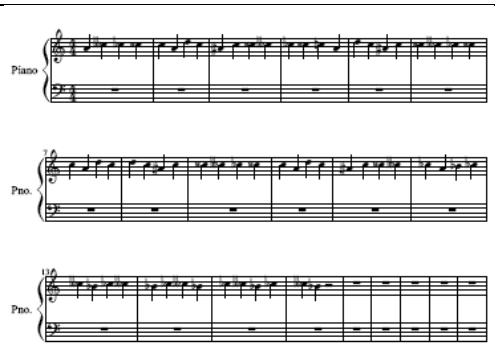
**Table 1**. Resulted code from translating process of SMILES into music notes and color combinations in the frame of the novel presented system

| Parameters of SMILES | Code for color combination* | Code for music note |
|---|---|---|
| = | #290A1F | CC |
| - | #3D0F2E | C |
| # | #14050F | CCC |
| ( | #FF75FF | C ♭ ♭ C ♭ |
| ) | #FF66FF | C ♭ C ♭ ♭ |
| [ | #A319FF | C*C |
| ] | #AD33FF | C C* |
| + | #660000 | D* |
| _ | #F0B2B2 | D |
| / | #660066 | C# |
| @ | #6600FF | DC |

| | | |
|---|---|---|
| . | #993399 | . |
| > | #D14719 | D♭ ♭D♭ |
| < | #D65C33 | D♭D♭ ♭ |
| * | #FFFFFF | 7A |
| : | #CC00CC | CC#C |
| \| | #9900FF | D# |
| 0 | #000000 | E* |
| 1 | #140F00 | 1E |
| 2 | #291F00 | 2E |
| 3 | #3D2E00 | 3E |
| 4 | #523D00 | 4E |
| 5 | #664C00 | 5E |
| 6 | #7A5C00 | 6E |
| 7 | #8F6B00 | 7E |
| 8 | #A37A00 | 8E |
| 9 | #B88A00 | 9E |
| C | #2E8A5C | A |
| c | #2E8A5C | A |
| P | #66FF33 | A* |
| N | #29527A | F |
| n | #29527A | F |
| S | #00FF99 | B♭ ♭ |
| O | #0099CC | B |
| o | #0099CC | B |
| Cl | #669900 | B* |
| H | #996633 | F* |
| F | #000099 | F♭ ♭ |

**Table 2**. Translational process of template/test chemical compounds to music note and color combination

| | |
|---|---|
| **SMILES steering for template** | CNCCC(OC1=CC=C(C=C1)C(F)(F)F)C1=CC=CC=C1 |
| Music note for pattern sample |  |
| Color combination for pattern sample |  |
| SMILES steering for test chemical compound with 70 % similarity to pattern sample | NC(=N)C1=CC=C(OCCCCCOC2=CC=C(C=C2)C(N)=N)C=C1 |
| Music note for test chemical compound with 70 % similarity to pattern sample |  |
| Color combination for test chemical compound with 70 % similarity to pattern sample |  |
| SMILES steering for test chemical compound with 5 % similarity to pattern sample | C([C@H]([C@H]([C@@H]([C@H](CO)O)O)O)O)O |

| | |
|---|---|
| Music note for test chemical compound with 5 % similarity to pattern sample |  |
| Color combination for test chemical compound with 5 % similarity to pattern sample |  |

Following that, the translated format of fluoxetine as the pattern sample as well as two chemical structures with 77 % and 5 % similarity with the pattern sample to melodic composition and image color combination were utilized to measure the visual and auditory perceptions in the definite statistical population. These patterned assessments and subsequent related profitable findings (Figures 1-4) were employed as an establishment plan for bio-mimic designing and introducing the innovative pattern for chemical compound similarity search.

## DISCUSSION

The fundamental phase during the early stages of the drug discovery procedure is the identification of possible active compounds with desirable biological activity in addition to drug-like parameters confirmed by certain evaluation tests [6]. The pointed phase traditionally has been complicated by given in vitro assays in the frame of High-Throughput Screening (HTS) operations [6]. This statement has led to a wide range of restrictions such as the limited size of compound libraries that have been evaluated by *in vitro* assays. This kind of limitation was in contradiction with the increasing capabilities of the synthetic chemistry and screening technologies in addition to creating and spreading numerous new diseases and illnesses followed by occurring critical need to discover and propose new drugs parallel to optimizing existing medicines [6].

In the early 1990s to overcome the mentioned challenges in addition to reducing the time and costs needed for the lead identification procedure, scientists introduced and employed a virtual screening approach as an *in silico* comparable of HTS. Computational screening-based virtual screening methods can be utilized to eradicate undesired compounds from the available pools of structures with huge amounts and decline the size of the structure population for bioassays [6-7]. On the other hand, virtual screening approaches can play the key role as a selective computational robot to assist in collecting the procedure of potential new bioactive structures among great virtual databases. As the basic definitions, Walters *et al*. described the virtual screening as "automatically evaluating very large libraries of compounds using computer programs" [8].

Commonly, the approach of virtual screening is classified into two concepts including receptor-based and ligand-based methods. On the subject of current research focus, ligand-based techniques are employed to achieve the physicochemical-related information subliminally accessible in identified ligands by this means of relying on a set of reference structures with recognized bioactivity, molecular descriptors in addition to an appropriate quantity of structure similarity [6]. These aspects led to becoming virtual screening as an essential part of the drug discovery/design procedure in the latest decades with the brilliant capacity of designing and optimizing targeted combinatorial databases in addition to enriching libraries of existing structures that yield a unique pharmacological profile [8].

As the pointed out, the realization of ligand-based virtual-screening computations is affected extremely by the quality-based ability of the employed algorithm for recognizing patterns of chemical compounds and relevant approach to identify diverse/similar structures among the pools of data [7].

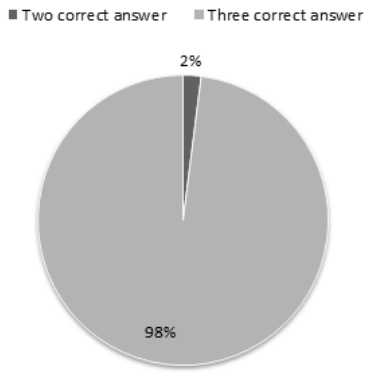



**Figure 1**. Percentage of correct answers for image color combinations of fluoxetine.

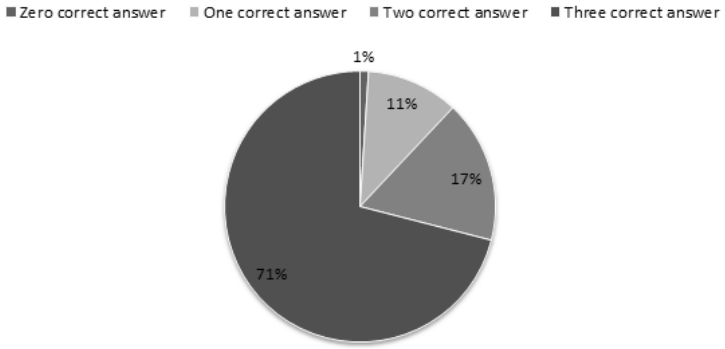



**Figure 2**. Percentage of correct answers for the melodic composition of fluoxetine.

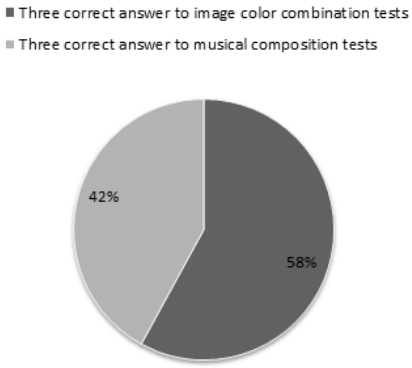



**Figure 3**. Compare the percentage of the correct answers to all tests for fluoxetine

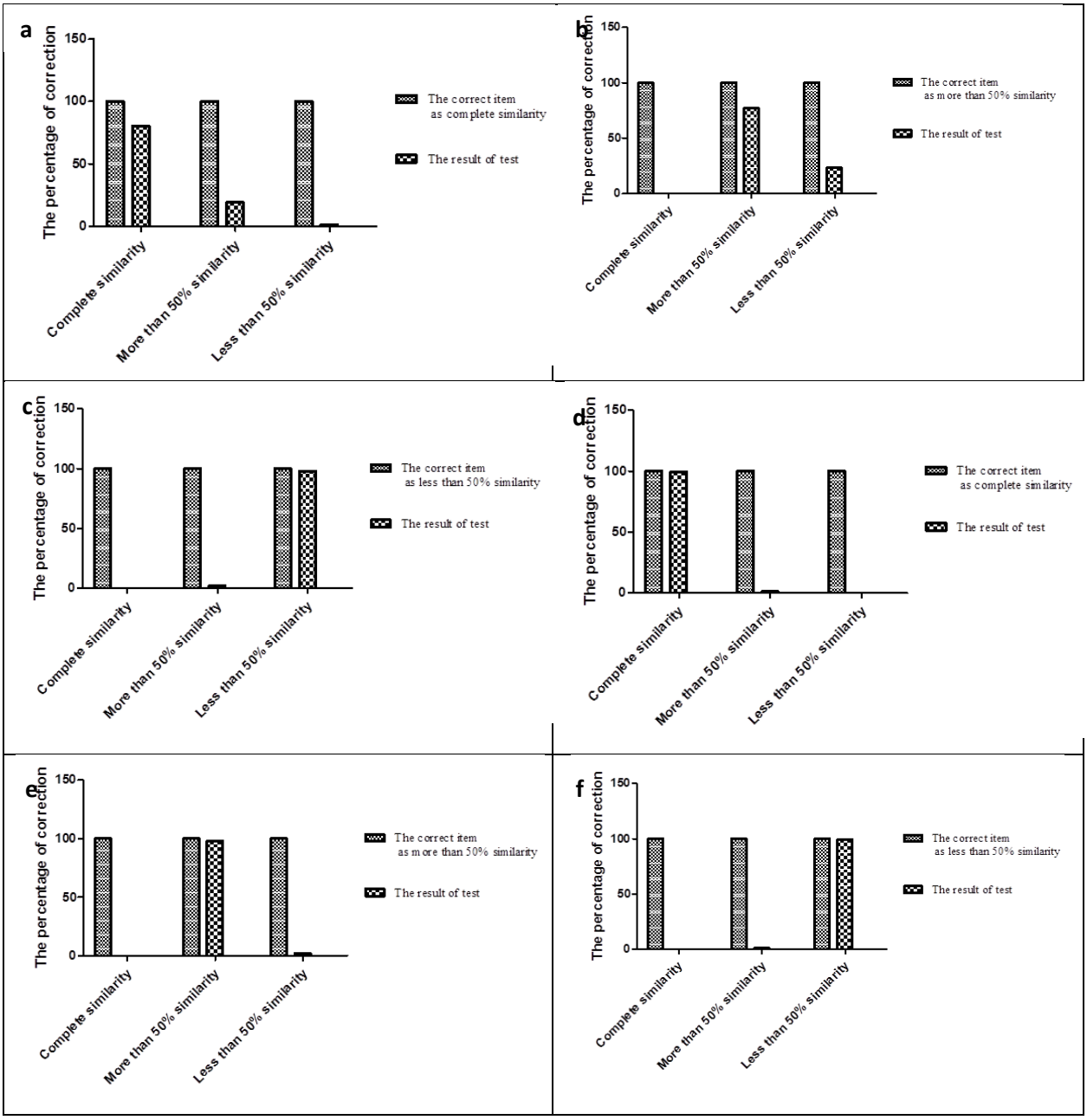



**Figure 4**. The correlation percentage of the tests for the melodic composition of fluoxetine (**a**) The correct answer is complete similarity, (**b**) Correct answer is more than 50 % similarity, (**c**) Correct answer is less than 50 % similarity, the correlation percentage of the tests for color arrangements of fluoxetine (**d**) The correct answer is complete similarity, (**e**) Correct answer is more than 50 % similarity, (**f**) Correct answer is less than 50 % similarity.

This fact has confirmed the need to develop an innovative similarity search algorithm to navigate high-dimensional chemical spaces, and train similarity calculations on certain compound classes, in addition, to detecting overlapped similarity relationships [3].

Supreme similarity approaches depend on the employing of chemical descriptors and reference spaces including string-based similarity searching, and simplified 2 Dimensional (2D) molecular graph symbols for similarity searching. Similarity searching utilizing 2D fingerprints such as SMILES fingerprint during virtual screening process towards early stages of lead-discovery programs [3].

Originally, 2D fingerprints were advanced for similarity searching utilizing single and multiple template molecules that with using multiple templates, all templates are identified actives. Nonetheless, even molecules found to be most similar to a single reference compound in an initial similarity search can be included, irrespective of activity [9]. Ongoing examinations to build unique fingerprint search presentations employing numerous template structures have much focused on methodologies to either scale or normal fingerprints and on the assessment of elective scoring plans, specifically, closest neighbor strategies and data combination [10].

There is the theory that instead of computer-based capabilities for similarity search approaches, human mind properties may be practical in assembling algorithms to introduce a superior algorithm for recognizing chemical patterns toward a successful similarity search process in addition to intermittently altering the ending findings to imitate patterns with epitomized capability [11].

On the subject of pointed theory, it would be possible that introduce the extra generation of similarity search methods more perfect based on the integration of different concepts of sciences such as the current investigation deliberated based on the bio-mimicking approach in addition to employing SMILES strings in the frame of 2D fingerprinting of chemical compounds towards introducing innovative recognition methodology for similarity search of chemical compound patterns.

The credible benefits of employing the auditory presentation of scientific evidence as a source to progress analytical approaches have been accurately established since 1994 [12-13]. The related investigators confirmed the complexity in addition to the proficiency of the human

auditory system to take aspects of the sounds based on using numerous complex layers of empathetic modules [12]. It would be notable because of selection analytical auditory approaches over visual-based analytical methods by scientists, noting related studying has been finalized for translating DNA and protein sequences to visual color-based combinations despite much-performed investigations for converting DNA and protein structures to music/sounds compositions [12-13].

On the other hand, there was no scientific evidence in previous research regarding converting/translating chemical structures into music/sound and visual color-based formats as well as presenting pattern recognition algorithms for chemical structure similarity search based on human auditory and visual perception systems.

Current investigation was performed in parallel with the study finalized by the same research group regarding introducing a protein translation method for converting amino acid sequences into music and color image based on human perception patterns [14].

A immense percentage of true answers to the entirety tests for evaluating the human visual and auditory capabilities to recognize patterns among 200 participants in the statistical community regarding recognition of the number of matching levels between the generated image combination of fluoxetine as the control set (pattern sample) in addition to the items of the test set counting the pattern sample as the exact 100 % similar structure (99 % true answer for image combination and 80 % for music composition) and the chemical structures with 77 % (98 % true answer for image combination and 77 % for music composition) in addition to 5 % similarity (99 % true answer for image combination and 98 % for music composition) with pattern sample, confirmed the presented translational method for similarity search in the current investigation.

Furthermore, the resulted evidence during the current search showed that the percentage of answer correction (42 % correct answer) for identification of melodic composition based on recognition capacities of the auditory system for input patterns was lower than the percentage of answer correction (58 % correct answer) in related to image color combinations regarding visual system abilities. This information may demonstrate the necessity for performing further investigation to present a novel operative algorithm for recognizing patterns of small-molecule

similarity search process regarding the bio-mimicking human visual system.

## CONCLUSION

The essential part of the early stages of the drug discovery process is the representation of probable active structures with appropriate biological activity in addition to drug-like factors confirmed by certain evaluation assays.

For the first time, the patterned strategy in the current investigation based on the concept of intra scientific studying and utilizing both abilities of cheminformatics and psychological aspects led to present the inimitable translator machine system named as TMS-CH for translating SMILES strings of chemical structures to melodic composition in addition to image color combination towards creating innovative pattern and system for small molecule similarity search.

Indeed, during the current investigation, the perception of chemical structures utilizing visual and auditory illustrations by a human was predicted to progress research related to the chemical structure similarity search methods toward profitable drug design process.

Therefore, the same further investigations can be practical and important as innovative developments to improve the ligand-based virtual screening field.

## REFERENCES

[1]. Friedman JH, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Statist*, 2000, 28: 337–407.

[2]. Varnek A, Baskin I. Chemoinformatics as a theoretical chemistry discipline. *Mol Inform*, 2011, 30(1): 20–32.

[3]. Willett P, Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today*, 2006, 11(23-24): 1046-53.

[4]. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today*, 2007, 12(5-6): 225-33.

[5]. Kearnes S, Riley P, McCloskey K, Berndl M, Pande V. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des*, 2016, 30: 595–608.

[6]. Schomburg K, Ehrlich HC, Stierand K, Rarey M. From structure diagrams to visual chemical patterns. *J Chem Inf Model*, 2010, 27; 50(9): 1529-35.

[7]. Hähnke V, Hofmann B, Grgat T, Proschak E, Steinhilber D, Schneider G. PhAST: pharmacophore alignment search tool. *J Comput Chem*, 2009, 15; 30(5): 761-71.

[8]. Klebe G, Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today*, 2006, 11(13-14): 580-94.

[9]. Muegge I, Oloff S. Advances in virtual screening. *Drug Discov Today*, 2006, 3(4): 405-11.

[10]. Xue L, Godden JW, Stahura FL, Bajorath J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J Chem Inf Comput Sci*, 2003, 43(4): 1218-25.

[11]. A, Floersheim P, Acklin P, Jacoby E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci*, 2003, 43(2): 391-405.

[12]. Garcia-Ruiz MA, Guiterrez-Pulido JR. An overview of auditory display to assist comprehension of molecular information. *Interact Comput*, 2006, 18: 853–68.

[13]. Shi XJ, Cai YY, Chan CW. Electronic music for biomolecules using short music phrases. *Leonardo*, 2007, 40 (2): 137–41.

[14]. Ghavami S, Toozandehjani H, Ghavami G, Sardari S. Innovative protein translation into music and color image applicable for assessing protein alignment based on bio-mimicking human perception system. *Int J Biol Macromol*, 2018, 119: 896-901.