

2013 Siemens Competition

Math : Science : Technology

Supplemental Form

(Please save to your hard drive, type in answers and print. Make three copies and attach one to each copy of your Research Project.)

STUDENT INFORMATION

Name of Individual or Team Leader:	Christopher Hsu
Names of Parents of Individual or Team Leader:	Qin Wang-Joy and David Joy
Name of Individual's or Team Leader's High School:	Park Tudor School
High School City and State:	Indianapolis, IN
Name of Team Member 2:	Jason Zhao
Names of Parents of Team Member 2:	Xiaowei Zhao and Jie Xu
Name of Team Member 3:	
Names of Parents of Team Member 3:	
Research Project Title:	Modeling Blood Brain Barrier Permeability Across Drug Classes Based on Fingerprirnt and Molecular Descriptor Space Similarity Searching
Project Type:	<input type="checkbox"/> Individual <input checked="" type="checkbox"/> Two-person team <input type="checkbox"/> Three-person team

PUBLICATIONS

Has this research project been published, accepted, or submitted for publication?	<input type="checkbox"/> YES - published <input type="checkbox"/> YES - accepted <input type="checkbox"/> YES - submitted <input checked="" type="checkbox"/> NO
If YES, please name the journal(s)	

ACKNOWLEDGEMENTS

You must acknowledge all individuals who assisted or advised you in any way (i.e. mentor, teacher, family members, lab assistants, other students). Please include the person's name, title, and institution and describe their participation/contribution. You are required to complete the Acknowledgements section and must include the name, title, institution, and role of any individuals who assisted with the Research Project, including family members. Receiving assistance from family members does not affect the judging in any way, but you must acknowledge their support. Failure to identify others who assisted you with any aspect of your research may result in disqualification.

Name, Title, Institution, Participation/Contribution
SAMPLE - Dr. John Smith, Professor of Biology, Sample University, Assisted with the collection of data
Mr. Jim Wikel, Head of Chemistry Dept. (retired), Eli Lilly, helped with conceptualization and genesis of project
Mr. Ryan Ritz, Teacher of Computer Science, Park Tudor School, Provided computer resources

Additionally, please identify the institutions (i.e. laboratory, university, etc.) where you performed your research, if not already acknowledged above.

Institution	City	State
Park Tudor School	Indianapolis	IN

Predicting Blood Brain Barrier Permeability across Drug Classes based on Fingerprint and Molecular Descriptor Space Similarity Searching

Abstract

The blood brain barrier (BBB) permeability of a drug plays a crucial role in drug development, and has a significant impact on the physiology of a human body. To the best of our knowledge, no studies have been reported that use the principle of molecular similarity to predict BBB permeability of drugs. Additionally, while a majority of cheminformatics similarity studies that have been conducted focus on the molecular descriptor space, our project placed an emphasis on similarity calculations based on structural fingerprints. Previous studies using molecular descriptors and structural keys have reported varying results. In the present study we supplemented molecular descriptors with molecular fingerprints which provided greater accuracy. Our objective was to create a Python program that could accurately and quickly predict BBB permeability of a molecule by comparison with other known molecules. From a dataset of 58 antihistamines and structurally related antidepressants and antianxiety drugs, our program ranked molecules by similarity score to a reference molecule. Based on this ranking, we determined the BBB permeability of the reference molecule. After testing many configurations, we found that combining the fingerprint formats, *fp2*, *fp3*, *fp4*, and *maccs* together with the molecular descriptors values of logP and molecular weight yielded the best separation between BBB permeable and impermeable drugs.

Predicting Blood Brain Barrier Permeability across Drug Classes based on Fingerprint and Molecular Descriptor Space Similarity Searching

Executive Summary

For the past decade, vast amounts of resources have been allocated to drug design, testing, and development with little return. One of the most significant considerations in drug development is interaction with the blood brain barrier (BBB), especially in regards to antihistamines, antidepressants, and anti-anxiety drugs. The blood brain barrier is a natural defense in humans that regulates what substances can enter the brain and affect the central nervous system. Designing new drugs requires accurate models of the BBB that can predict permeability among a wide variety of molecules. Antihistamines, for example are preferred to be BBB impermeable, while antidepressants and anti-anxiety drugs must cross the BBB. Therefore, there is high demand for better BBB models and prediction techniques.

Several methods and techniques researchers often use to gain greater understanding of a drug's permeability of the blood brain barrier is the application of regression analysis, machine learning methods, and various classification methods. Experimental models of the blood brain barrier are complex and multiple methods / assays may not agree.

We created a Python program that alleviates this issue through the application of molecular similarity to anticipate blood brain barrier permeability. The simplicity of our calculations allows for quick and accurate results. Our project focuses on exploring the use of this concept in a specific class of drugs whose members display both properties. In the construction of this technique, we determined what forms of comparison yield the most reliable predictions while maintaining efficiency.

Predicting Blood Brain Barrier Permeability across Drug Classes based on Fingerprint and Molecular Descriptor Space Similarity Searching

1 Introduction

In the development of antihistamine drugs, one of the most problematic aspects is the effect of the blood brain barrier, a natural defense to keep foreign substances from entering the brain⁴. For example, Benadryl, a commercial antihistamine used widely for allergy control, causes mild sedation because it crosses the blood brain barrier (BBB) and enters the brain. Newer second generation antihistamines like Zyrtec do not cross the blood brain barrier to the extent that first generation drugs like Benadryl do, and therefore, do not cause a sedative effect³. Predicting BBB permeation remains a challenge in drug design and current solutions have been less than satisfactory. Various sources have used descriptor values such as lipophilicity (log P), topological indexes, and simple atom numerations to unsubstantially anticipate BBB transport^{3, 4, 5}. Few descriptors have been shown to influence the derived relationships in a significant matter and most focus on terms describing hydrogen bonding^{4, 5}.

The term “molecular similarity” was introduced to computational chemistry in the 1990s to set the expectations that molecules with similar properties would or could be expected to elicit similar biological responses⁶. The formalization of this similarity concept in the field of cheminformatics is important to the pharmaceutical drug discovery effort in their quest for new treatments for diseases. Cheminformatics is a field of science that arose from the application of computer based informational processes in chemistry. Chemical similarity is defined using cheminformatic techniques such as distance measures in predefined molecular descriptor spaces¹¹. These distances are expressed as binary digits where 1 is defined as identity. The molecular

descriptor space is always referenced in the calculations and may be either based on 2D descriptors of a molecule or on 3D descriptors¹.

Two types of comparison for molecules are primarily used: molecular fingerprints and molecular descriptors. The most common type of molecular fingerprinting is a series of binary digits that represent the presence or absence of certain substructures in the molecule¹. Fingerprints, compared to descriptors, more accurately represent the overall structure of the molecule, while descriptors can be used to quantify bulk properties of molecules, like molecular weight, logP, and bond counts. Fingerprints and molecular descriptors can both be used to predict drug functionality, but fingerprints can bring different classes of drugs into an easily comparable frame of reference.

Our project addresses the uncertainty of molecular descriptor similarity calculations by using molecular *fingerprinting* algorithms in combination with molecular descriptor trends to identify correlations between molecular structural similarity in a simple Python program.

2 Materials and Procedures

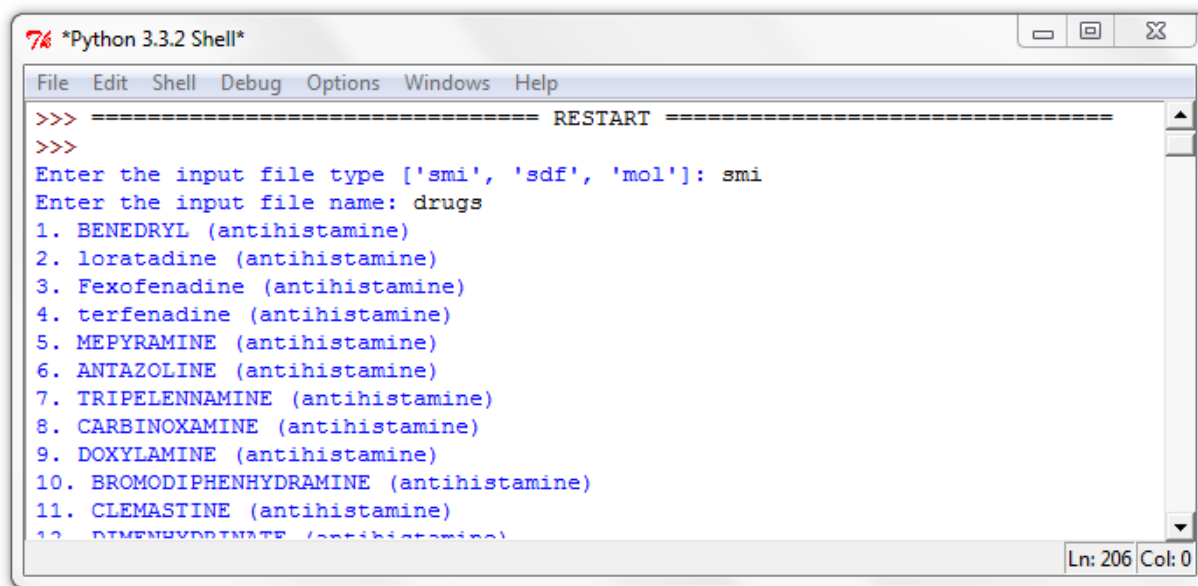
Open-Source Software

Open source cheminformatic techniques were used in the Python coding language to create a unified molecular similarity calculator. The OpenBABEL open source chemistry toolkit provides a readily available API that can be used to receive structures in a predefined file format, compute a set molecular structure fingerprint, and compare molecular fingerprints based on a calculated tanimoto coefficient. Additionally, comparison based on molecular descriptors was developed in Python to complement molecular fingerprint analysis by OpenBABEL. These three functions readily separate into system components consisting of a computer algorithm to compute molecular fingerprints, another computer algorithm to compute tanimoto coefficients

and compare the molecular fingerprints, and a reporter system to provide the results. Additionally, the scientific Python packages *Numpy*, *Scipy*, and *Scikit-Learn* as well as the open source visualization program, *Avogadro*, were utilized for mathematical computation, data structures not available in the Python library, and molecular visualization.

Procedures

In order to create a unified molecular similarity calculator, we chose to build our own console-based Python program because of the multitude of extensions available for Python. The first step was to receive the incoming structures in a predefined standard molecular format such as the SMILES or MDL SD formats. Users are prompted to enter the input file's extension to determine the file type and the input file's name at initialization (Fig. 1) and then displayed a list of all the molecules found within the specified file. The program then converts the molecular data from the input file into an internal format the OpenBABEL toolkit can use to perform calculations. Afterwards, users are prompted to select from one or any combination of the four fingerprint formats available in OpenBABEL: fp2, fp3, fp4, and maccs. Each fingerprint



```
*Python 3.3.2 Shell*
File Edit Shell Debug Options Windows Help
>>> ===== RESTART =====
>>>
Enter the input file type ['smi', 'sdf', 'mol']: smi
Enter the input file name: drugs
1. BENEDRYL (antihistamine)
2. loratadine (antihistamine)
3. Fexofenadine (antihistamine)
4. terfenadine (antihistamine)
5. MEPYRAMINE (antihistamine)
6. ANTAZOLINE (antihistamine)
7. TRIPELENNAMINE (antihistamine)
8. CARBINOXAMINE (antihistamine)
9. DOXYLAMINE (antihistamine)
10. BROMODIPHENHYDRAMINE (antihistamine)
11. CLEMASTINE (antihistamine)
12. DIMENHYDRINATE (antihistamine)
Ln: 206 Col: 0
```

Figure 1: Procedures at the initialization of our program are shown, including inputting a file type and name. Below that, the program displays all molecules (by name) found within the specified file.

contains a unique set of SMART statements. SMARTS is a line notation for specification of sub-structures in molecules. Fingerprints are calculated by examining structures in the molecule and assigning binary bits of 1 to structures that match the definition of the selected format(s). The result is a string of binary digits that can be compared with other fingerprints to create a similarity score. Since each fingerprint format defines different structures, there is variance in the similarity scores calculated from differing fingerprint formats. To increase robustness, users can opt for the program to combine formats by averaging similarity scores.

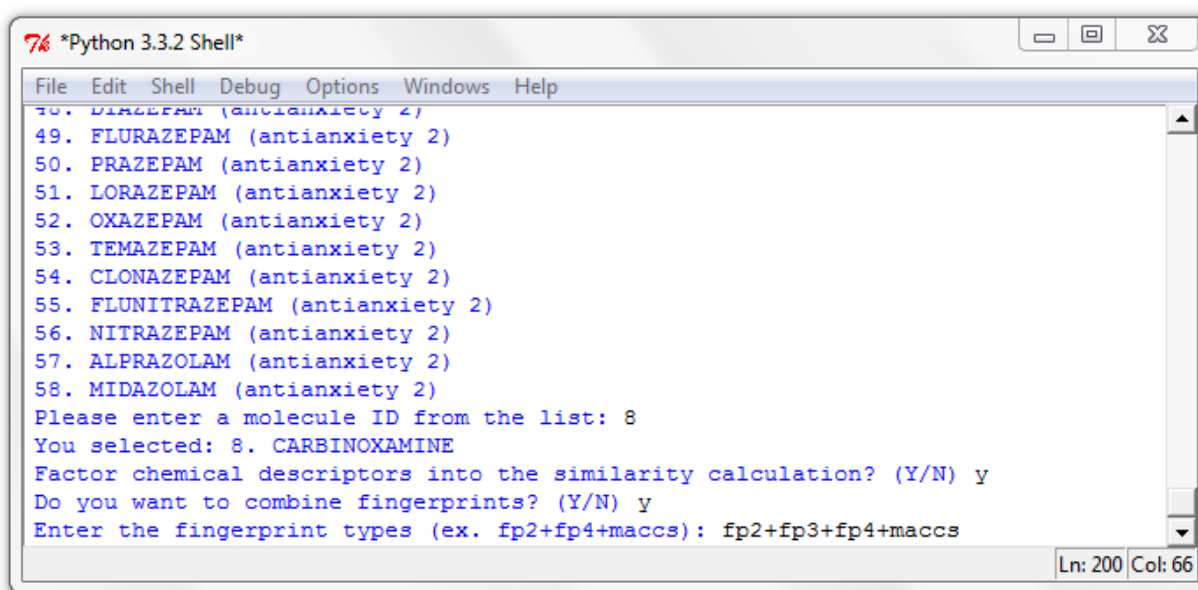


Figure 2: Users are prompted to select a molecule from the list, whether they want to include molecular descriptors in the similarity calculation, and whether they want to combine fingerprints. Below, a fingerprint combination is entered that combines the 'fp2', 'fp3', 'fp4', and 'maccs' fingerprint formats.

After displaying all the molecules in the user-selected file, our program prompts users to select a reference molecule as shown in Figure 2. This molecule will be compared to all other molecules in the set, and after the similarity calculations, molecules will be displayed in decreasing similarity from it. Our program also uses chemical descriptors to supplement fingerprint similarity score calculation. Chemical descriptors are calculated from molecular

structure data and factored into similarity score by averaging an overall score calculated from fingerprints with an overall score calculated from descriptors. Our program integrates two molecular descriptors: LogP (measure of the solubility of a molecule) and molecular weight, since these descriptors have been known to consistently predict BBB permeability well¹³. While many other descriptors exist, we did not factor them into similarity calculations; however, this is a possibility in future work. The OpenBABEL toolkit for Python, known as Pybel, was used to calculate molecular descriptor values.

At this point, initialization is complete, and the program will proceed with the calculating fingerprints, tanimoto coefficients, and descriptor values. Tanimoto calculations involve the pairwise comparison of the fingerprint of two molecules, A and B. In these calculations only the features or bits in each structure are considered. Assume N_A is the number of chemical features/bits in molecule A, N_B is the number of chemical features/bits in molecule B, and N_{AB} is the number of features/bits common to both molecule A and B, then the calculation becomes:

$$T_c = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (1)$$

These calculations are efficient and quick. A score ranging between 0 and 1.0 is produced. Identical structures will result in a score of 1.0 and it is generally accepted that scores greater than or equal to 0.85 are structurally similar molecules. In combination with fingerprint analysis, chemical descriptors such as logP and molecular weight were factored into the similarity score calculation. LogP, also known as the Partition Coefficient, measures the lipid or fat solubility of a compound and plays a key role in determining BBB diffusion; it can be calculated as follows:

$$\log P_{oct/wat} = \log \frac{[solute]_{octanol}}{[solute]_{un-ionized\ water}} \quad (2)$$

Descriptor values were factored into the similarity score by calculating the distance from the reference molecule and then normalizing the values to a scale from 0 to 1.0. Then, the

average of the values is averaged with the tanimoto coefficient produced by the fingerprint analysis to create an "averaged similarity" score. The process of calculating a similarity score can be summed up with the following equation, where n is the number of combined fingerprints, $\log P_{AB}$ is the normalized distance of the logP of A and the logP of B, and m_{AB} is the normalized distance of the molecular weights of A and B:

$$S(A, B) = \frac{\frac{1}{n} \sum_{i=1}^n T_c(A, B)_i + \frac{1}{2} (\log P_{AB} + m_{AB})}{2} \quad (3)$$

With this algorithm, an accurate similarity score can be calculated for each molecule that is compared with the reference molecule. Our program then ranks (in respect to the reference molecule) the molecules by similarity scores. Based on the permeability of the most similar drugs to the reference molecule, we can make an accurate prediction as to whether the reference molecule is BBB permeable or impermeable.

3 Results

To test our model, we used a set of well-known drugs that spanned three drug classes: antihistamines, antidepressants, and antianxiety drugs. Our dataset was divided into drugs that crossed the BBB and those that didn't. Drugs in our dataset that had their name CAPITALIZED represented drugs that were permeable to the BBB, and those whose names were lowercase were drugs that didn't cross the BBB. The set of drugs used in testing can be seen below in Table 1.

Table 1: Drugs used for Testing

#	LogP	Mol Wt.	Name	Class	SMILES
1	7.156	469.658	ebastine	1	<chem>O(CCN(C)C)C(c1ccccc1)c2ccccc2</chem>
2	6.384	471.673	terfenadine	1	<chem>O=C(OCC)N4CC/C(=C2/c1ccc(Cl)cc1CCc3cccn23)CC4</chem>
3	5.729	436.952	TIANEPTINE	2	<chem>O=C(O)C(c1ccc(cc1)C(O)CCCN2CCC(CC2)C(O)(c3ccccc3)c4ccccc4)(C)C</chem>
4	5.573	415.958	rupatadine	1	<chem>OC(c1ccccc1)(c2ccccc2)C4CCN(CCCC(O)c3ccc(cc3)C(C)(C)C)CC4</chem>
5	5.448	501.656	Fexofenadine	1	<chem>O(c1ccc(cc1)CN(c2ncccc2)CCN(C)C)C</chem>
6	5.431	390.948	MECLIZINE	1	<chem>N\1=C(\NCC/1)CN(c2ccccc2)Cc3ccccc3</chem>
7	5.362	458.57	astemizole	1	<chem>n1ccccc1N(CCN(C)C)Cc2ccccc2</chem>
8	5.042	343.89	CLEMASTINE	1	<chem>Clc1ccc(cc1)C(OCCN(C)C)c2ncccc2</chem>

9	4.826	382.883	loratadine	1	<chem>O(CCN(C)C)C(c1ncccc1)(c2cccc2)C</chem>
10	4.636	287.398	CYPROHEPTADINE	1	<chem>BrC1ccc(cc1)C(OCCN(C)C)c2cccc2</chem>
11	4.593	314.852	CLOMIPRAMINE	2	<chem>Clc1ccc(cc1)[C@](OCC[C@@H]2N(C)CCC2)(c3cccc3)C</chem>
12	4.552	298.446	ALIMEMAZINE	1	<chem>O(CCN(C)C)C(c1cccc1)c2cccc2</chem>
13	4.304	284.419	PROMETHAZINE	1	<chem>n1cccc1C(c2cccc2)CCN(C)C</chem>
14	4.252	284.439	IPRINDOLE	2	<chem>Clc1ccc(cc1)C(c2ncccc2)CCN(C)C</chem>
15	4.217	263.377	NORTRIPTYLINE	2	<chem>Clc1ccc(cc1)[C@@H](c2ncccc2)CCN(C)C</chem>
16	4.186	294.434	TRIMIPRAMINE	2	<chem>BrC1ccc(cc1)[C@@H](c2ncccc2)CCN(C)C</chem>
17	4.169	277.403	AMITRIPTYLINE	2	<chem>n3c(\C(=C\CN1CCCC1)c2ccc(cc2)C)cccc3</chem>
18	4.147	292.418	DIMETINDENE	1	<chem>n1cccc1C(C=3c2cccc2CC=3CCN(C)C)C</chem>
19	4.117	334.251	BROM...DRAMINE	1	<chem>c1c(cccc1)C(c2cccc2)N3CCN(CC3)C</chem>
20	4.108	388.888	bepotastine	1	<chem>Clc1ccc(cc1)C(c2cccc2)N3CCN(CC3)C</chem>
21	3.989	266.381	DESIPRAMINE	2	<chem>Clc1ccc(cc1)C(c2cccc2)N3CCN(CC3)CCOCCO</chem>
22	3.962	279.376	DOXEPIN	2	<chem>Clc1ccc(cc1)C(c2cccc2)N3CCN(CC3)Cc4cccc(c4)C</chem>
23	3.953	348.438	acrivastine	1	<chem>S2c1cccc1N(c3c2cccc3)CC(N(C)C)C</chem>
24	3.952	309.425	ketotifen	1	<chem>S2c1cccc1N(c3c2cccc3)CC(C)CN(C)C</chem>
25	3.94	280.407	IMIPRAMINE	2	<chem>c43\C(=C1/CCN(C)CC1)c2cccc2\C=C/c3cccc4</chem>
26	3.928	319.239	DEXB...NIRAMINE	1	<chem>n4c3\C(=C1/CCN(C)CC1)c2cccc2CCc3ccc4</chem>
27	3.855	278.391	TRIPOLIDINE	1	<chem>O=C3c1sccc1C(\c2c(cccc2)C3)=C4/CCN(C)CC4</chem>
28	3.819	274.788	DEXC...NIRAMINE	1	<chem>Fc1ccc(cc1)Cn2c5cccc5nc2NC4CCN(CCC3ccc(OC)cc3)CC4</chem>
29	3.819	274.788	CHLORPHENAMINE	1	<chem>Clc1ccc(cc1)C(c2cccc2)N3CCN(CC3)CCOCC(=O)O</chem>
30	3.76	325.767	MIDAZOLAM	3	<chem>Clc1cc5c(cc1)\C(=C3/CCN(Cc2cncc(c2)C)CC3)c4ncccc4CC5</chem>
31	3.646	290.402	AZATADINE	1	<chem>O=C5/C=C\N=C(\N(C)C4CCN(c1nc3cccc3n1Cc2ccc(F)cc2)CC4)N5</chem>
32	3.553	300.826	CHLORCYCLIZINE	1	<chem>O=C(O)\C=C\c3nc(\C(=C\CN1CCCC1)c2ccc(cc2)C)ccc3</chem>
33	3.506	387.878	FLURAZEPAM	3	<chem>O=C(c1ccc(cc1)C(C)(C)C)CCCN4CCC(OC(c2cccc2)c3cccc3)CC4</chem>
34	3.477	432.493	mizolastine	1	<chem>Clc1ccc(cc1)[C@H](OC2CCN(CCCC(=O)O)CC2)c3ncccc3</chem>
35	3.435	324.804	PRAZEPAM	3	<chem>OC(c1cccc1)(c2cccc2)C4C3CCN(CC3)C4</chem>
36	3.403	290.788	CARBINOXAMINE	1	<chem>c12c(C[C@@H]3c4c2cccc4CCN3C)ccc(O)c1O</chem>
37	3.354	255.355	DIMENHYDRINATE	1	<chem>C1(\c2c(CCc3c1cccc3)cccc2)=C/CCNC</chem>
38	3.354	255.355	BENEDRYL	1	<chem>c3cc2c(/C(c1c(cccc1)CC2)=C\CCN(C)C)cc3</chem>
39	3.202	293.403	quifenadine	1	<chem>N1(c2cc(Cl)ccc2CCc2c1cccc2)CCCN(C)C</chem>
40	3.165	240.343	PHENIRAMINE	1	<chem>c1cc3c(cc1)CCc2c(cccc2)N3CCCN(C)C</chem>
41	3.135	315.711	CLONAZEPAM	3	<chem>O=C2c1c(cccc1)N(c3c(N2CCN(C)C)cccc3)C</chem>
42	3.024	388.888	zyrtec_citirizine	1	<chem>O3c1cccc1C(c2c(cccc2)C3)=CCCN(C)C</chem>
43	3.016	308.765	ALPRAZOLAM	3	<chem>c1cc3c(cc1)CCc2c(cccc2)N3CCCN(C)C</chem>
44	2.932	374.904	HYDROXYZINE	1	<chem>c13c(n(c2cccc12)CCCN(C)C)CCCCC3</chem>
45	2.923	270.369	DOXYLAMINE	1	<chem>C1(c2c(S(=O)(=O)N(c3c1cccc3)C)cc(Cl)cc2)NCCCCCCC(O)=O</chem>
46	2.899	266.381	CYCLIZINE	1	<chem>c1cc3c(cc1)CCc2c(cccc2)N3CC(C)CN(C)C</chem>
47	2.788	267.322	AMOMORPHINE	2	<chem>Clc1ccc2\N=C(/C/[N+])(/[O-])=C(\c2c1)c3cccc3)NC</chem>
48	2.736	299.755	CHLO...EPOXIDE	3	<chem>c12c(N(C(=O)CN=C1c1cccc1)C)ccc(c2)Cl</chem>
49	2.675	321.158	LORAZEPAM	3	<chem>Fc3cccc3C/2=N/CC(=O)N(c1c\2cc(Cl)cc1)CCN(CC)CC</chem>
50	2.658	285.384	MEPYRAMINE	1	<chem>Clc4cc\1c(N(C(=O)C/N=C/1c2cccc2)CC3CC3)cc4</chem>
51	2.654	284.74	DIAZEPAM	3	<chem>C1(=NC(C(=O)Nc2c1cc(cc2)Cl)O)c1c(Cl)cccc1</chem>
52	2.65	255.358	TRIPLENNAMINE	1	<chem>c1ccc(cc1)C2=NC(C(=O)Nc3c2cc(cc3)Cl)O</chem>
53	2.617	295.379	DIBENZEPIN	2	<chem>CN1c2ccc(cc2C(=NC(C1=O)O)c3cccc3)Cl</chem>

54	2.572	313.283	FLUNITRAZEPAM	3	<chem>C=1(c2c(NC(CN1)=O)ccc([N+](=O)[O-])c2)c1c(Cl)cccc1</chem>
55	2.481	281.266	NITRAZEPAM	3	<chem>[O-][N+](=O)c3cc\1c(N(C(=O)C/N=C/1c2ccccc2F)C)cc3</chem>
56	2.459	265.353	ANTAZOLINE	1	<chem>[O-][N+](=O)c3cc\1c(NC(=O)C/N=C/1c2ccccc2)cc3</chem>
57	2.022	286.713	OXAZEPAM	3	<chem>n12c3c(C(c4ccccc4)=NCc1nnc2C)cc(cc3)Cl</chem>
58	1.973	300.74	TEMAZEPAM	3	<chem>n12c3c(C(c4c(F)cccc4)=NCc1cnc2C)cc(cc3)Cl</chem>

Table 1: Displayed is the list of drugs used for the testing and optimization of our program along with the accompanying SMILES string, logP value, and molecular weight. Drugs are tagged with one of three classes: antihistamine (1), antidepressant (2), and antianxiety (3). The full names of drug #s 19, 26, 28, and 48 are BROMODIPHENHYDRAMINE, DEXBROMPHENIRAMINE, DEXCHLORPHENIRAMINE, and CHLORDIAZEPOXIDE respectively.

Data

To analyze all similarity scores, we generated multiple NxN matrices – where N is the number of molecules – that display the averaged similarity score between each molecule. Comparison tests were run using each of the four fingerprints as well as with all possible combinations of those fingerprints. Additionally, matrices with molecular descriptors and without molecular descriptors were generated and analyzed. The tanimoto score from fingerprints combined with molecular descriptor values creates an “averaged score”. Values in the matrices were also color-coded (0.0 = red, 1.0 = green) with a gradient generating algorithm to aid in visualization. The matrix can be seen in Figure 4 in the *Illustrations* section, although the matrix was so large, we can only displaying a smaller sampling of it. Trends and patterns are clearly visible in the matrix making it a valuable component in visual analysis.

Qualitative Analysis

In our qualitative analysis, we looked at separation between BBB permeable and BBB impermeable drugs. Below (Table 2) are results from our similarity list program that has a user select a reference molecule. In this iteration, we chose CARBINOXAMINE as the reference molecule, which is displayed at the top of the list at number 0.

Table 2: Similarity List Result

#	Score	Name	Class	#	Score	Name	Class
0	1.0	CARBINOXAMINE	1	29	0.635	zyrtec_citirizine	1
1	0.885	DOXYLAMINE	1	30	0.632	PRAZEPAM	3
2	0.811	DIMENHYDRINATE	1	31	0.631	MIDAZOLAM	3
3	0.811	BENEDRYL	1	32	0.627	PROMETHAZINE	1
4	0.801	CHLORPHENAMINE	1	33	0.626	DIAZEPAM	3
5	0.794	DEXCHLORPHENIRAMINE	1	34	0.616	CYPROHEPTADINE	1
6	0.774	CHLORCYCLIZINE	1	35	0.613	AMOMORPHINE	2
7	0.752	BROMODIPHENHYDRAMINE	1	36	0.613	ketotifen	1
8	0.739	PHENIRAMINE	1	37	0.603	ALIMEMAZINE	1
9	0.736	DEXBROMPHENIRAMINE	1	38	0.602	acrivastine	1
10	0.725	AZATADINE	1	39	0.601	CHLORDIAZEPOXIDE	3
11	0.721	TRIPOLIDINE	1	40	0.595	NORTRIPTYLINE	2
12	0.717	MEPYRAMINE	1	41	0.594	CLONAZEPAM	3
13	0.699	CYCLIZINE	1	42	0.591	FLURAZEPAM	3
14	0.698	DIMETINDENE	1	43	0.590	ANTAZOLINE	1
15	0.688	IPRINDOLE	2	44	0.577	loratadine	1
16	0.686	DOXEPIN	2	45	0.576	LORAZEPAM	3
17	0.679	quifenadine	1	46	0.568	TEMAZEPAM	3
18	0.679	TRIPLENNAMINE	1	47	0.565	OXAZEPAM	3
19	0.677	IMIPRAMINE	2	48	0.553	NITRAZEPAM	3
20	0.668	CLEMASTINE	1	49	0.542	FLUNITRAZEPAM	3
21	0.666	HYDROXYZINE	1	50	0.526	MECLIZINE	1
22	0.655	TRIMIPRAMINE	2	51	0.495	mizolastine	1
23	0.655	DIBENZEPIN	2	52	0.465	rupatadine	1
24	0.654	ALPRAZOLAM	3	53	0.399	astemizole	1
25	0.654	bepotastine	1	54	0.327	TIANEPTINE	2
26	0.644	CLOMIPRAMINE	2	55	0.287	ebastine	1
27	0.642	AMITRIPTYLINE	2	56	0.287	Fexofenadine	1
28	0.635	DESIPRAMINE	2	57	0.281	terfenadine	1

Table 2: The results of our similarity list program with the reference molecule as CARBINOXAMINE and the fingerprint formats as fp2, fp3, fp4, and maccs. Molecular descriptor values are also factored in to the score. Class 1 is antihistamine, class 2 is antidepressant, and class 3 is antianxiety.

Looking through the table, one can see that BBB permeable drugs are grouped together well. Drug classes also form smaller clusters within the larger groups. Non-BBB permeable drugs, however, do not perform as well. There is clear clustering at the end of the list, but there are also many BBB impermeable drugs scattered throughout the permeable drugs. These scattered BBB impermeable drugs represent a sect of molecules that have unusual shapes and

characteristics, therefore making it hard to identify similarities and differences. The similarity lists were also used for optimization. The BBB permeable compounds may be clustering due to the selectivity of the BBB for specific molecular properties reflected in the fingerprints and property descriptors. The scatter of the BBB impermeable compounds may reflect the multitude of ways compounds are excluded from brain penetration.

Using Avogadro to visualize these molecules, we further tested the accuracy of our similarity score algorithm. Below (Figure 5) are six images of molecules from the comparison labeled A-F. Three (A-C) were categorized as similar by our program, while three others (D-F) were categorized as dissimilar. Plugging in data from the similarity list to the molecular visualization application, Avogadro, we can see that the most similar and dissimilar molecules to the reference, CARBINOXAMINE (A).

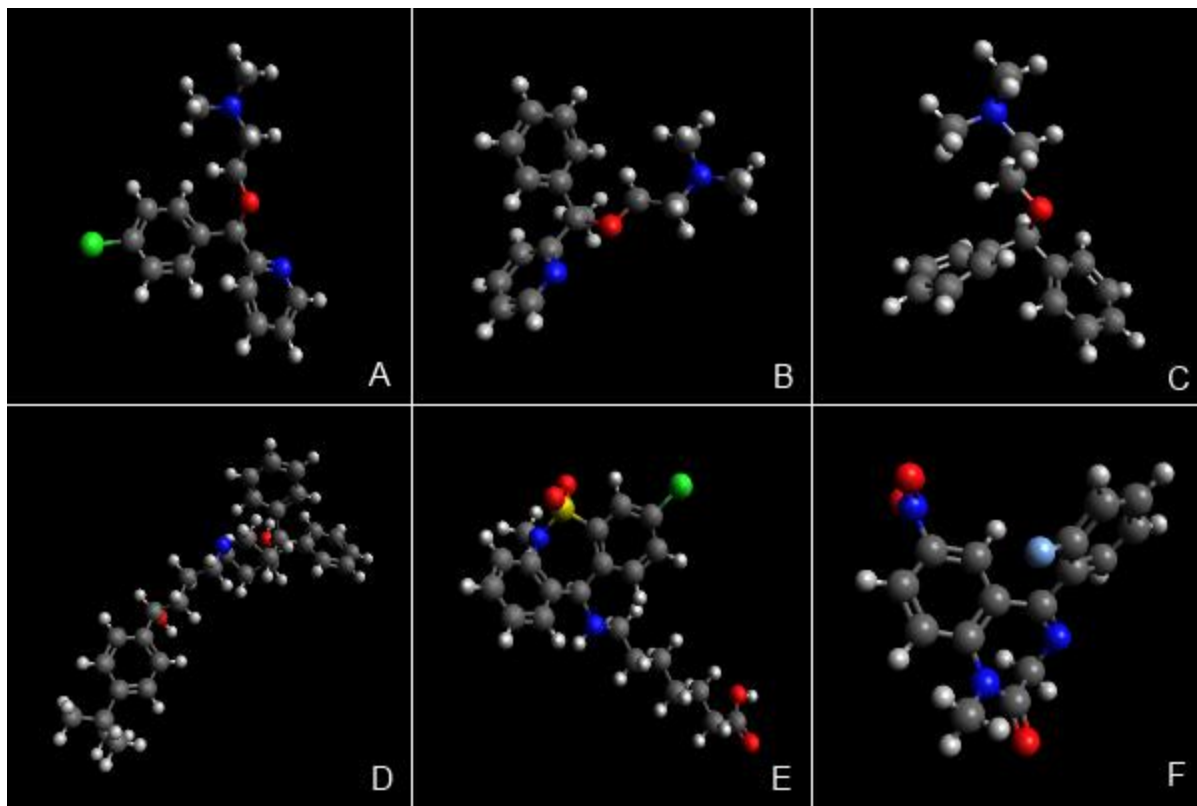


Figure 5: The molecules shown in this figure from top-left to bottom-right are CARBINOXAMINE (A), DOXYLAMINE (B), DIMENHYDRINATE (C), terfenadine (D), TIANEPTINE (E), and FLUNITRAZEPAM (F). Dark gray is carbon, light gray is hydrogen, red is oxygen, blue is nitrogen, yellow is sulfur, teal is noble gasses, and green is fluorine/chlorine. From this diagram, one can see the similarity among A, B, and C (classified as similar) and the dissimilarity between the top row and the bottom row (classified as dissimilar).

From basic visual analysis, it is clear in the figure above that DOXYLAMINE (B) and DIMENHYDRINATE (C), the two most similar drugs from the similarity calculations, are indeed very similar to the reference molecule. All three have oxygen atoms in their center, and cyclic carbon-nitrogen/carbon rings. Molecule D, E, and F are clearly different; molecule D is elongated which would hinder its ability to be BBB permeable, while molecules E and F have different atoms like sulfur and noble gasses which are chemical differences that likely wouldn't affect the molecules ability to pass through the BBB. The visual analysis further affirms the accuracy of the results from the similarity list.

Quantitative Analysis

Multidimensional Scaling (MDS) is a way to visualize the similarity of individual classes of a dataset. This technique takes in an NxN distances matrix and creates a map where objects are placed in a manner that represents the distances in the matrix. The results of the MDS analysis can be seen in Figure 6 in the *Illustrations*. Each molecule is represented as a point and similarity is reflected by how close points are to each other. Clustering of similar molecules can be clearly seen, and wholly dissimilar molecules maintain a distance from all other points.

4 Discussion

Variability in Descriptors

Cheminformatics literature describes a plethora of molecular descriptors that have been proposed and used over the past 20 years. These molecular descriptors can be classified according to the type of representation used by the fingerprinting algorithm. Chemical structures are typically represented in one or more format consisting of a one-dimensional (1D) format, a two-dimensional (2D) format, or a three-dimensional (3D) format. A one-dimensional format (i.e. SMILES format) is a very concise format and molecular descriptors are quickly

computed. These molecular descriptors typically relate bulk properties like molecular weight, molar refractivity, logP (logarithm of the octanol/water partition coefficient), etc. A two-dimensional (2D) format (i.e. MDL SDF file), is significantly larger in size. Two-dimensional based descriptors describe properties such as connectivity indices, path counts, bond counts, ring count, etc. A three-dimensional (3D) format is very similar to the 2D format and is generated by a specific program starting from either the 1D or 2D molecular representation. Three-dimensional (3D) descriptors depend on conformations of molecules and include descriptor values for solvent accessible surface areas, principal moment of inertia, Van der Waals volume, etc. Three-dimensional (3D) descriptors are more compute expensive and have the added complexity of multiple 3D conformations of a structure. Thus, to achieve the fastest similarity calculations without compromising effectiveness, we used one-dimensional formats.

Molecular descriptors can also be classified according to their environment into the five following classes: (i) constitutional, (ii) topological, (iii) geometrical, (iv) electronic, and (v) quantum-chemical. Constitutional descriptors are fragment additive and reflect properties similar to the 1D derived properties (ex. molecular weight). The topological descriptors are derived from either 1D or 2D structure representations using the mathematical graph theory applied to the scheme of atoms connections of the structure. Typical examples of topological descriptors are Kier-Hall Indices, Balaban index, Burden numbers, B-Cut values, etc. Both constitutional and topological descriptors are facile to calculate. The geometrical, electronic, and quantum-chemical descriptors are more computationally intensive to compute and describe the molecule in a higher level of theory. For our algorithm, we used molecular weight (constitutional) and logP (electronic) descriptors.

As with the wide selection of molecular fingerprints and descriptors, Cheminformatics literature describes the use of a plethora of similarity measures. Investigators have examined which type of similarity measures perform best in a given situation, and results are variable. Currently, the widely accepted method for measure similarity is the Tanimoto coefficient (T_c), also known as the Jaccard coefficient. Another method for measuring similarity is the Tversky Index, which is an asymmetrical similarity measure that is a generalization of the Tanimoto Coefficient. We did not use the Tversky index when performing any similarity calculations, but the measurement could be used in future similarity research.

Distinguishing Factors of Structural Keys and Molecular Fingerprints

Our research distinguishes itself in its application of molecular fingerprints in similarity calculations for the prediction of BBB penetration. Before the advent of fingerprints, *structural keys* were the first type of screen employed for high-speed screening of chemical databases. Structural keys are represented as *boolean arrays*. As the name implies, a *structural key* is a bitmap in which each bit represents the presence (TRUE) or absence (FALSE) of a specific structural feature (pattern).

To make a structural key, one decides which structural features are important, assigns a bit of the bitmap to each, then generates a bitmap for each molecule in the database. Constructing structural keys is often a time-consuming process. A substructure search for each pattern represented in the bitmap and for each molecule in the database must be performed.

The list of patterns that one might use is long. Some examples are:

- The presence/absence of each element, or if an element is common (ex. nitrogen) (3), several bits might represent "at least 1 N", "at least 2 N", and so on.
- Unusual or important electronic configurations (ex. sp^3 carbon, triple-bonded nitrogen.)

- Rings and ring systems, such as cyclohexane, pyridine, or naphthalene.

Structural keys, however, suffer from a lack of generality. Patterns included in the key have a negative effect on the search speed across a database: An effective choice will screen out virtually all structures that aren't of interest, greatly increasing search speed, whereas a poor choice will cause many "false hits," which slows searching to a crawl. The choice of patterns also depends on the nature of the queries to be made.

Fingerprints resolve this lack of generality by avoiding the idea of pre-defined patterns. A fingerprint is a boolean array, but unlike a structural key there is no assigned meaning to each bit. Unlike a structural key with its pre-defined patterns, fingerprint patterns are generated from the molecule itself. Fingerprinting examines the molecule and generates the following:

- a pattern for each atom
- a pattern representing each atom and its nearest neighbors (plus the bonds that join them)
- a pattern representing each group of atoms / bonds connected by paths up to 7 bonds long

For practical purposes, the number of patterns one might encounter is infinite, but the number produced for any molecule can be easily handled by a computer. In spite of the difference between the meaning of a fingerprint's bits and a structural key's bits, both share an important feature. Similarly to structural keys, we can use simple boolean operations on fingerprints to screen molecules, making a fingerprint comparison an extremely fast screen for substructure searching. Thus, fingerprints can be used for extremely accurate, yet efficient calculations for similarity that can immediately benefit drug research.

There are four widely accepted classifications of fingerprint identification: *fp2*, *fp3*, *fp4*, and *maccs*. *fp2* fingerprints are path-based fingerprints which index small molecule fragments, whereas *fp3*, *fp4*, and *maccs* fingerprints are identified using SMARTS queries stored in native

files that are often determined by drug development companies. These various SMARTS queries reference patterns in molecule branching, bonding, and ligands.

Rather than focusing on single descriptor classes or even individual fingerprint analysis, we questioned whether a comprehensive method of comparison would yield results that were consistent with past trials and, perhaps, accurate to an even greater extent. We used our similarity matrix to visualize and understand the effectiveness of comparison among individual fingerprint classes.

Optimization and Testing

Since similarity models vary in accuracy across different drug classes, our program was optimized for the three relatively similar drug classes: antihistamines, antidepressants, and antianxiety. While our program can extrapolate and perform similarity calculations on any class of drugs, accuracy cannot be guaranteed unless testing and optimization is performed. Antihistamine drugs were chosen as our initial experimental drug class, for the BBB barrier permeability attribute either applies or does not apply to a given antihistamine molecule. This duality in antihistamine drugs serves well for similarity comparisons that aim to predict drug activity. All antidepressant and antianxiety drug tests were used as test cases for our program's algorithm, for all drugs in the two drug classes were previously known to be BBB permeable.

5 Conclusions

Our project has addressed a number of weaknesses in molecular similarity for a few popular market drug classes, but it has also highlighted several opportunities for further understanding of molecular similarity in fields outside of pharmaceutical development.

We have found that primarily using molecular fingerprints with supporting molecular descriptor values yields more accurate and consistent results than descriptor values alone when

comparing antihistamine, antianxiety, and antidepressant drug classes. The use of molecular descriptors like LogP and molecular weight in similarity calculations has shown to produce reliable results and we have expanded upon that with the addition of molecular fingerprint algorithms. Our program has added another layer to cheminformatic similarity calculations, increasing robustness and accuracy. Using this more robust and accurate approach, we saw that predicting BBB permeability increased proportionally to the accuracy of the similarity calculations. We propose the range of drug classes that the principle of similarity applies to can be expanded, possibly into fields such as agrochemicals. In future projects, we would like to explore the advantages that fingerprint similarity calculations could bring to chemical development outside of pharmaceuticals.

6 Illustrations

	BENEDRYL (antihistamine)	loratadine (antihistamine)	Fexofenadine (antihistamine)	terfenadine (antihistamine)	MEPYRAMINE (antihistamine)	ANTAZOLINE (antihistamine)	TRIPLENNAMINE (antihistamine)	CARBINOXAMINE (antihistamine)	DOXYLAMINE (antihistamine)	BROMODIPHENHYDRAMINE (antihistamine)	CLEMASTINE (antihistamine)	DIMENHYDRINATE (antihistamine)	PHENIRAMINE (antihistamine)
BENEDRYL (antihistamine)	1.0	0.518	0.353	0.364	0.669	0.62	0.692	0.82	0.829	0.834	0.659	1.0	0.731
loratadine (antihistamine)	0.518	1.0	0.513	0.506	0.494	0.43	0.457	0.624	0.564	0.633	0.716	0.518	0.506
Fexofenadine (antihistamine)	0.353	0.513	1.0	0.844	0.312	0.292	0.291	0.372	0.347	0.453	0.548	0.353	0.344
terfenadine (antihistamine)	0.364	0.506	0.844	1.0	0.317	0.278	0.298	0.377	0.359	0.461	0.56	0.364	0.36
MEPYRAMINE (antihistamine)	0.669	0.494	0.312	0.317	1.0	0.73	0.88	0.732	0.753	0.605	0.541	0.669	0.699
ANTAZOLINE (antihistamine)	0.62	0.43	0.292	0.278	0.73	1.0	0.797	0.613	0.658	0.52	0.476	0.62	0.665
TRIPLENNAMINE (antihistamine)	0.692	0.457	0.291	0.298	0.88	0.797	1.0	0.701	0.746	0.567	0.504	0.692	0.785
CARBINOXAMINE (antihistamine)	0.82	0.624	0.372	0.377	0.732	0.613	0.701	1.0	0.898	0.775	0.71	0.82	0.755
DOXYLAMINE (antihistamine)	0.829	0.564	0.347	0.359	0.753	0.658	0.746	0.898	1.0	0.723	0.653	0.829	0.78
BROMODIPHENHYDRAMINE (antihistamine)	0.834	0.633	0.453	0.461	0.605	0.52	0.567	0.775	0.723	1.0	0.766	0.834	0.603
CLEMASTINE (antihistamine)	0.659	0.716	0.548	0.56	0.541	0.476	0.504	0.71	0.653	0.766	1.0	0.659	0.552
DIMENHYDRINATE (antihistamine)	1.0	0.518	0.353	0.364	0.669	0.62	0.692	0.82	0.829	0.834	0.659	1.0	0.731
PHENIRAMINE (antihistamine)	0.731	0.506	0.344	0.36	0.699	0.665	0.785	0.755	0.78	0.603	0.552	0.731	1.0
CHLORPHENAMINE (antihistamine)	0.697	0.593	0.399	0.412	0.685	0.64	0.729	0.812	0.753	0.668	0.646	0.697	0.894
DEXCHLORPHENIRAMINE (antihistamine)	0.693	0.59	0.397	0.409	0.679	0.637	0.723	0.805	0.748	0.664	0.655	0.693	0.884
DENBROMPHENIRAMINE (antihistamine)	0.646	0.618	0.443	0.455	0.654	0.591	0.677	0.753	0.7	0.744	0.672	0.646	0.838
TRIPOLIDINE (antihistamine)	0.666	0.637	0.402	0.413	0.663	0.649	0.695	0.733	0.73	0.638	0.627	0.666	0.781
DIMETINDENE (antihistamine)	0.632	0.654	0.422	0.436	0.651	0.596	0.664	0.712	0.692	0.655	0.637	0.632	0.759
CYCLIZINE (antihistamine)	0.753	0.486	0.347	0.363	0.733	0.745	0.8	0.714	0.76	0.639	0.576	0.753	0.794
CHLORCYCLIZINE (antihistamine)	0.709	0.572	0.402	0.415	0.688	0.666	0.713	0.779	0.689	0.696	0.673	0.709	0.734
HYDROXYZINE (antihistamine)	0.636	0.614	0.463	0.474	0.606	0.563	0.574	0.694	0.657	0.669	0.646	0.636	0.554
MECLIZINE (antihistamine)	0.516	0.688	0.591	0.606	0.499	0.491	0.52	0.587	0.522	0.624	0.731	0.516	0.541
PROMETHAZINE (antihistamine)	0.615	0.548	0.385	0.391	0.649	0.625	0.659	0.645	0.634	0.623	0.599	0.615	0.641
ALIMEMAZINE (antihistamine)	0.582	0.572	0.417	0.425	0.609	0.587	0.615	0.626	0.604	0.618	0.621	0.582	0.617
CYPROHEPTADINE (antihistamine)	0.615	0.645	0.434	0.442	0.586	0.566	0.599	0.639	0.604	0.63	0.648	0.615	0.657

Figure 4: Displayed is a sampling of the NxN matrix generated by our program. This view provides raw data for a given comparison; all N molecules in a file are compared to every other molecule in the file. This table was also used in the MDS algorithm by converting it to a distance matrix. This matrix can also be used to spot patterns and trends, for example, we can clearly see from this view that Fexofenadine and terfenadine are quite dissimilar from every other molecule, but similar to each other.

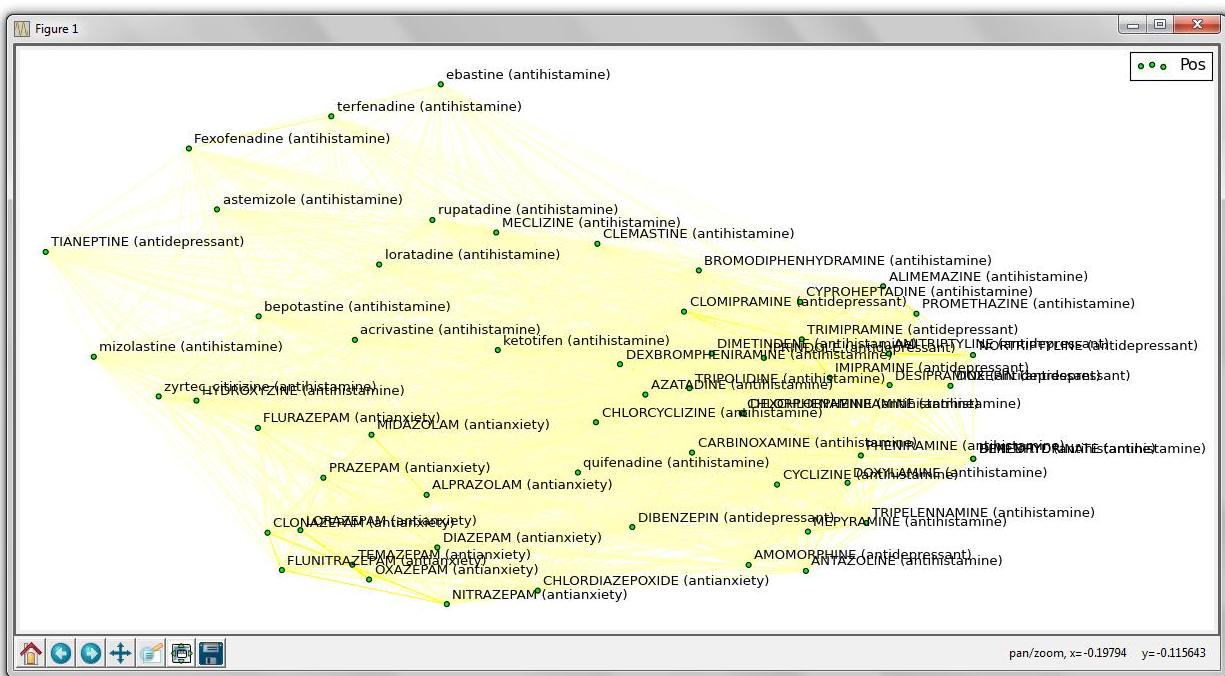


Figure 6: This distance map that the MDS algorithm generated represents the similarity between molecules. Clearly, similar drugs are clustering into semi-tight collections, while dissimilar drugs are separated from clusters. This map is interactive, so crowded areas can be zoomed in as demonstrated in Figure 7.

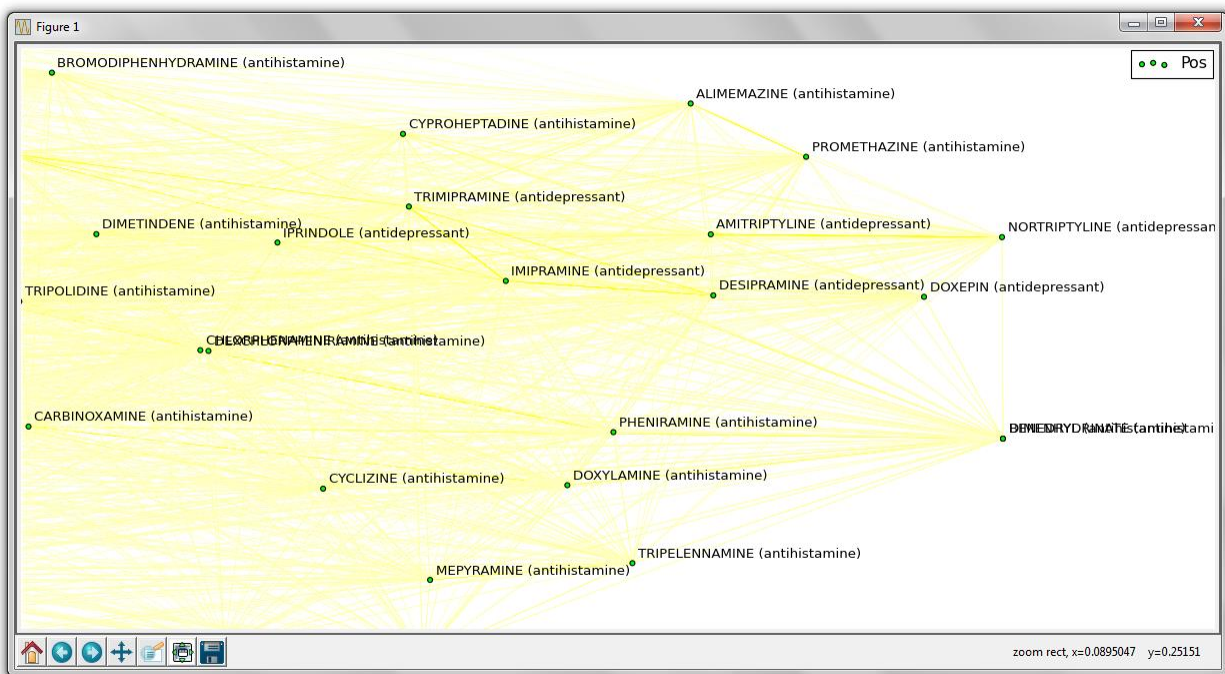


Figure 7: A zoomed portion of MDS generated similarity map of Figure 6. From this view, we can clearly see the similarity of CHLORPHENIRAMINE and DEXCHLORPHENIRAMINE based on their very close distance. Zooming in even further might be required to read the point labels.

References

- [1] Ulf Norinder, Markus Haeberlein, *Computational approaches to the prediction of the blood–brain distribution*, *Advanced Drug Delivery Reviews* 54 (2002) 291–313
- [2] Patrizia Crivori, Gabriele Cruciani, Pierre-Alain Carrupt, Bernard Testa, *Predicting Blood-Brain Barrier Permeation from Three-Dimensional Molecular Structure*, *Journal of Medicinal Chemistry* 2000, 43, 2204-2216
- [3] Eric Deconinck, Menghui H. Zhang, Danny Coomans, and Yvan Vander Heyden, *Classification Tree Models for the Prediction of Blood-Brain Barrier Passage of Drugs*, *Journal of Chemical Information and Modeling*, 2006, 46, 1410-1419
- [4] M. Ecemis, James H. Wikel, Christopher Bingham, and Eric Bonabeau, *A Drug Candidate Design Environment Using Evolutionary Computation*, *Evolutionary Computation*, IEEE Transactions on , vol.12, no.5, pp.591,603, Oct. 2008
- [5] Stephen A. Hitchcock, Lewis D. Pennington, *Structure-Brain Exposure Relationships*, *Journal of Medicinal Chemistry* 2006, 49:26
- [6] A. M. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, New York: John Wiley & Sons 1990 [ISBN0-471-62175-7](#).
- [7] Gerhard F. Ecker, Christian R. Noe, *In Silico Prediction Models for Blood-Brain Barrier Permeation*, *Current Medicinal Chemistry*, 2004, 11, 1617-1628
- [8] Gré'gori Gerebtzoff, Anna Seelig, *In Silico Prediction of Blood-Brain Barrier Permeation Using the Calculated Molecular Cross-Sectional Area as Main Parameter*, *Journal of Chemical Information and Modeling* 2006, 46, 2638-2650
- [9] Travis T. Wager, Ramalakshmi Y. Chandrasekaran, Xinjun Hou, Matthew D. Troutman, Patrick R. Verhoest, Anabella Villalobos, and Yvonne Will, *Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes*, *ACS ChemNeurosci* 2010 1(6): 420-434
- [10] Hongmao Sun, *A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption*, *Journal of Chemical Information and Modeling* 2004, 44, 748-757
- [11] A. M. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, New York: John Wiley & Sons 1990 [ISBN0-471-62175-7](#).
- [12] Hassan Pajouhesh, George R. Lenz, *Medicinal Chemical Properties of Successful Central Nervous System Drugs*, *The Journal of the American Society for Experimental NeuroTherapeutics* 2005 October; 2(4): 541–553.

[13] Noel M O'Boyle; Michael Banck; Craig A James; Chris Morley; Tim Vandermeersch; Geoffrey R Hutchison, *Open BABEL: An open chemical toolbox*, Journal of Cheminformatics **2011**, 3:33

[14] Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek and Geoffrey R Hutchison, *Avogadro: An advanced semantic chemical editor, visualization, and analysis platform*, Journal of Cheminformatics **2012**, 4:17.