Spring 2020

Utilizing Virtual High-Throughput Screens to Identify Promising Inhibitors of Complement Factor C1s

Lyndsey Schmucker
lns41@zips.uakron.edu

Utilizing Virtual High-Throughput Screens to Identify Promising Inhibitors of Complement Factor C1s

Honors Course: 4200 497-001

Recipient: The University of Akron Williams Honors College

Authors: Lyndsey Schmucker

A version of this work was previously published with authors:
Jon J. Chen & Dr. Donald P. Visco Jr.

5/2/20

Executive Summary

Background

The project's purpose is to determine if an algorithm designed to identify potential drug candidates can identify drug candidates that target complement factor C1s, a protein which causes tissue damage when under-regulated (Chen et al., 2018 p.1). The project consists of identifying potential candidates in a large compound database with predictive models and a series of experiments designed to test the different candidates' biological activity to confirm the performance of the algorithm. Since the algorithm has the capability to screen millions of compounds, it is beneficial to the process of medicine development as it allows a larger pool of compounds to be considered that would have otherwise been overlooked in a quicker manner than traditional methods (Chen et al., 2018 p.1). Once the performance of the algorithm is verified, the algorithm can be widely applied to target other biological systems and identify potential drug candidates for specific diseases. Furthermore, candidate structure can be analyzed to yield potential leads for further investigation.

Quantitative Results

In the first round of experimental testing, seven compounds were tested. Four of these compounds are considered active since they possess reported half-maximal inhibitory concentration (IC_{50}) values, meaning they inhibit the biological activity of the activated C1s by half (Aykul, 2016). The results from the first round of experiments are used to augment the existing data and retrain the models in the algorithms to improve predictive power for a higher hit rate. After retraining the models, fifty-two compounds were identified and ten were purchased for experimental validation (Chen et al., 2018 p. 11). Of the ten compounds that were tested, only five are considered active for an experimental validation hit-rate of 50%, which is lower than the first-round experimental validation hit-rate of 57%.

Conclusions

The 9 compounds showing activity are promising drug candidates as they have the capability to inhibit C1s. From observing the structures of the compounds that were experimentally tested in **Tables 4-5**, its apparent two core molecular structures or scaffolds are prominent, which can be seen in **Figures 3-4**. From comparing compounds, inferences can be made on how activity is related to both positions of large functional groups and the identity of functional groups, which is detailed in the **Discussion/Analysis** section (Chen et al., 2018 p.13). Moreover, in the PubChem Bioassay ID (AID) 787(Diamond, 2008) dataset, the fraction of actives predicted from the training set was 0.12, a fraction smaller than in other studies that contribute to the larger endeavor (Chen et al., 2018 p.13). These findings suggest the pipeline and models have the capability to be applied in cases with a limited amount of data on the desired active compounds (Chen et al., 2018 p.13).

Broader Implications of Work

The experimental results are of benefit to society due to the identification of nine potential drug candidates that have the capability to inhibit C1s, a protein which causes tissue damage when under-regulated (Chen et al., 2018 p.1). Developing medicine by using computer algorithms and virtual high-throughput screens is promising due to the larger number of compounds that can be considered in a quicker manner than traditional methods (Chen et al., 2018 p.1). Through the results of the studies in this series, the range of applications that the pipeline can be utilized is further defined in order to support and enrich future drug discovery efforts (Chen et al., 2018 p.18).

Technical/career skills that were obtained as the result of this research include technical aptitude, technical writing and editing skills in the process of publishing articles, problem solving, soft skills, data analysis, laboratory skills, and refined written and oral communication skills. Personal gains include exposure to concepts such as high-throughput computing, data mining and binning, structure-activity relationships, machine learning, and a deeper understanding of bioinformatics. Other personal gains include discipline, adaptability, open mindedness, and improved confidence in presenting research posters.

Recommendations

The 9 compounds showing activity are promising drug candidates as they have the capability to inhibit C1s. These compounds are recommended for further testing, such as at the cellular level, to see the effects on C1s inhibition (Chen et al., 2018 p.13). Applications that the pipeline can be applied to should continue to be explored as its potential to speed-up the process of developing medicine by using computer algorithms and virtual high-throughput screens is promising.

Introduction

At a time when viruses such as the coronavirus are becoming widespread, the importance of a reliable solution to efficiently find probable drug candidates to cure or manage symptoms is urgent. The greatest significance of this research is its contribution to the medicinal industry. Specifically, its potential to speed-up the process of developing medicine by using computer algorithms and virtual high-throughput screens. Using these methods, a larger number of potential compounds are considered in a quicker manner than traditional methods (Chen et al., 2018 p.1). As a result, the chances of finding a drug candidate that can successfully target a specific protein to debilitate the selected disease increase (Chen et al., 2018 p.1). This project is a part of a larger study which is comprised of several studies. The objective of the larger study is to determine the effectiveness and applicability of a series of models referred to as the pipeline (Chen et al., 2018 p. 5). The individual studies identify likely active compounds through utilizing computer algorithms and virtual high-throughput screens (Chen et al., 2018 p. 3). Through experimentation, compound activity with the targeted protein is verified. The focus of this study is to identify the drug candidates to treat human complement factor C1s, a protein which causes tissue damage when under-regulated (Chen et al., 2018 p.1).

Complement system functions as a part of the immune system in targeting and killing harmful bacteria in the body (Gani). The complement system is comprised of many proteins which circulate in blood and tissue fluids fulfilling a critical role in inflammation and defense against infections (Gani). The complement system can be activated via three pathways, one of which is the classical pathway that involves complement components C1, C2, and C4 (Gani). The classical pathway is activated by antibody-antigen complexes binding to C1(Gani). When proteins are activated, they have a cascade response. Meaning, the activation of one protein, for example C1, will consequently activate the next protein in the cascade (Gani). C1 is comprised of subcomponents C1q, C1r, C1s each of which are activated in a cascade manner to initiate responses in the immune system (Gani). Since the C1 complex is essential to the activation of the classical pathway, it is an imperative molecule to monitor (Gani).

The body's only mechanism to regulate C1 activation is through the production of the C1 inhibitor (Ratnoff & Lepow, 1957, as cited in Chen et al., 2018 p. 2). Alterations to the production of the C1 inhibitor can lead to the excessive activation of C1 which will consequently cause inadequate regulation of the classical pathway (Chen et al., 2018 p. 2). Excessive activation of C1 can lead to chronic inflammation and tissue damage (Chen et al., 2018 p. 2). Deficiencies in the components of the classical pathway including C1, C4, and C2 can lead to systemic lupus erythematosus, glomerulonephritis, and polymyositis, which are diseases characterized by tissue damage or inflammation (Gani). Deficiencies of the C1 inhibitor can lead to hereditary angioedema (Gani), a disease characterized by episodes of severe swelling (Hereditary). While treatment to these diseases could include a supplementary C1 inhibitor, it is a costly approach (Chen et al., 2018 p. 2). Another approach is to inhibit the activation of C1 directly versus utilizing the C1 inhibitor (Buerke et al., 2001, as cited in Chen et al., 2018 p. 2).

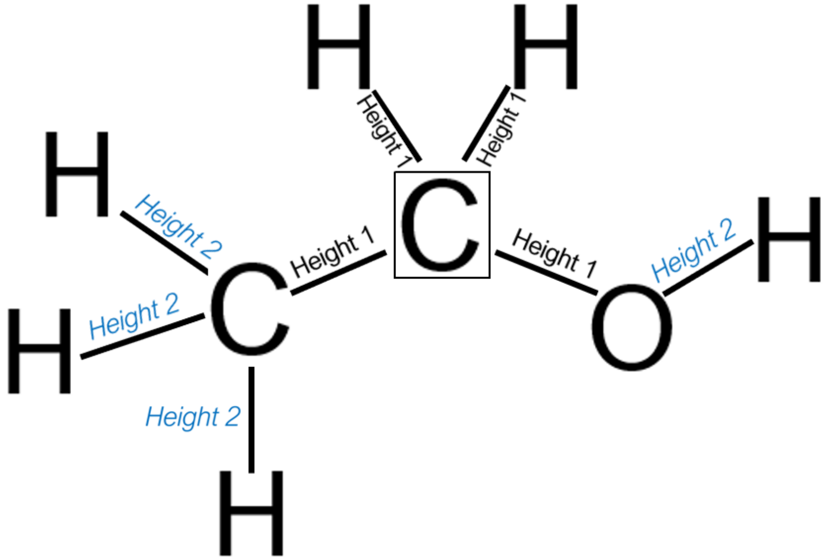
Specifically, targeting C1s is most promising as it initiates the classical pathway by activating subsequent steps C1r and C1q (Chen et al., 2018 p. 2&18). As a result, the focus of this work is to identify small molecule C1s inhibitors by using a virtual high-throughput screen (vHTS) and computer algorithms (Chen et al., 2018 p. 2). This is a promising approach as other researchers have used these methods to discover potential molecules for other components in the complement system (Vulpetti et al., 2017, as cited in Chen et al., 2018 p. 2). This study contributes to a larger endeavor which determines the effectiveness of the pipeline on varying protein/ligand systems, dataset sizes, and dataset active/inactive classification (Chen et al., 2018 p. 6). The pipeline refers to models through which structural feature patterns in compounds are correlated with experimental data to find new ligands as potential candidates (Chen et al., 2018 p. 6).

Background

Virtual high-throughput screen (vHTS), which is a virtual approach to high-throughput screen, is a computational method enabling the organization and exploration of candidate libraries for drug testing (Chen et al., 2018 p. 3). Through vHTS, models are utilized to analyze molecules digitally. The algorithm that implements vHTS is comprised of the genetic algorithm (GA) [Whitley, 1994] and the support vector machine (SVM). This study and the other studies contributing to the larger endeavor utilize existing data through ligand-based approaches (Chen et al., 2018 p. 3). Ligand-based approaches consist of active/inactive classification, quantitative structure-activity relationship (QSAR) models, and similarity of known ligands (Chen et al., 2018 p. 3,6).

The methodology of this study and the other studies contributing to the larger endeavor follow the same general procedure consisting of four main steps including: (1) identifying a targeted dataset, (2) using the targeted data set to train predictive classification and quantitative structure-activity relationship (QSAR) models, (3) screen a compound library with the classification and QSAR models, and (4) experimentally validate model predictions (Chen et al., 2018 p. 4). To further elaborate on the first step, PubChem Bioassay ID (AID) 787 was selected for possessing the experimental and ligand structure data required for the pipeline while containing a minuet fraction (11.8%) of active compounds as Cls inhibitors (Diamond, 2008, as cited in Chen et al., 2018 p. 6). From the selected dataset, pan-assay interference compounds (PAINS)(Baell, 2010), which are compounds that undesirably interact with multiple proteins, are removed (Chen et al., 2018 p. 3). It was discovered that these compounds gave false-negative experimental results. Therefore, PAINS(Baell, 2010) were removed so the performance or predictability of the trained models are not affected (Chen et al., 2018 p. 3). After the PAINS(Baell, 2010) are removed, the resulting data set will be used to train models.

The second step involves applying the algorithms, which is used here as a blanket term for a bunch of codes including Signature molecular descriptor and model training (Chen et al., 2018 p. 4). These codes will be utilized to screen upwards of 72 million compounds in the PubChem Compound database for potential candidates/compounds (Chen et al., 2018 p. 4 & 6). Signature molecular descriptor is a technique that translates molecular structures into a code that can be understood and used by algorithms used for model training (Chen et al., 2018 p. 4). **Figure 1** depicts the molecular structure of ethanol, which includes its elements and bonds, and the resulting Signature molecular fragmentation that allows it to be used by the algorithms (Chen et al., 2018 p. 5). The root atom, which is designated by the carbon surrounded by borders in **Figure 1**, is a height of zero, and all directly bonded atoms have a height of one (Chen et al., 2018 p. 5). Atoms that are bonded to the atoms at a height of one, or the root atom’s secondary atomic neighbors, are designated by a height of 2 (Chen et al., 2018 p. 5). While the signature for the root atom is defined as the atomic structure, the combination of atomic Signatures for all atoms is defined as the molecular signature (Chen et al., 2018 p. 5).



Height = 1

Atomic signature for $\boxed{C}$:	C(C O H H)
Molecular Signature:	1C(C H H H) + 1C(C O H H) + 5H(C) + 1O(C H) + 1H(O)

Height = 2

Atomic signature for $\boxed{C}$:	C(C(H H H) O(H) H H)
Molecular Signature:	1C(C(O H H) H H H)+1C(C(H H H) O(H) H H)+2H(C(C O H))+1O(C(C H H)H)+3H(C(C H H))+1H(O(C))

Figure 1: Depicts the molecular structure of ethanol, which includes its elements and bonds, and the resulting Signature molecular fragmentation at heights one and two from the root atom which is designated by the carbon enclosed in a box. While the root atom has a height of zero, all directly bonded atoms to the root atom have a height of one (Chen et al., 2018 p. 5). Atoms that are bonded to the atoms at a height of one, or the root atom's secondary atomic neighbors, are designated by a height of 2 (Chen et al., 2018 p. 5). While the signature for the root atom is defined as the atomic structure, the combination of atomic Signatures for all atoms is defined as the molecular signature (Chen et al., 2018 p. 5).
Slightly modified from Chen, J and Visco, D.P. Identifying novel factor XIIa inhibitors with PCA-GA-SVM developed vHTS models. *European J. Med. Chem.*; 140:31-41. Copyright © 2017 Elsevier Masson SAS. All rights reserved. which is cited in Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.

After Signature translates the molecules, the Molecular Signatures are used as inputs to create and train models using algorithms including principal component analysis (PCA), support vector machine (SVM) and genetic algorithm (GA) (Chen et al., 2018 p.16). The principal component analysis (PCA) associates a weighted contribution of each atomic Signature so the atomic Signatures contributing the most to capturing variance are identified and used to create the GA-SVM Virtual high-throughput screen (vHTS) models (Chen et al., 2018 p. 16). The GA-SVM vHTS models create a positive feedback loop in which they depend on each other to yield an optimum model (Chen et al., 2018 p. 16).

The GA(Whitley, 1994) designs atomic Signature combinations, which the support vector machine (SVM) uses to create models. The SVM models are evaluated for cross-validation accuracy, which is used as a metric to measure the predictive power of the models and whether the compounds are active or inactive (Chen et al., 2018 p. 16). The cross-validation scores are reported back as scores for the atomic Signature combinations to the GA(Whitley, 1994) (Chen et al., 2018 p. 16). The GA(Whitley, 1994) can then ranks atomic Signature combinations by the cross-validation scores and uses the top percentage to make the next round of combinations which is then sent to the SVM. By removing unsuitable combinations through the iterations between GA-SVM, overfitting or the variance and noise in the data is minimized (Chen et al., 2018 p. 16). Through each iteration, the predictability relating Signatures to compound activity improves. The combination of Signature and GA-SVM models are then used in the next step to screen the PubChem Compound database for potential candidates/compounds (Chen et al., 2018 p. 6).

The third main step in the procedure involves screening a compound library with the classification and QSAR models (Chen et al., 2018 p. 4). The PubChem Bioassay ID (AID) 787[Diamond, 2008] is used to develop two models/filters: classification (active/inactive) and the half-maximal inhibitory concentration (IC_{50}) prediction to test for compound activity (Chen et al., 2018 p. 6). The trained GA-SVM vHTS models are then implemented to screen around 72 million compounds in the PubChem Compound database by filtering them based off classification and IC_{50} prediction to identify potential active candidates for experimental validation (Chen et al., 2018 p. 16). Candidates selected for experimental validation are further narrowed down by availability and financial feasibility. Step four involves experimentally validating the model predictions (Chen et al., 2018 p. 4). Experimental procedures for this study are detailed in the **Experimental Methods** section and are used to verify biological activity (Chen et al., 2018 p. 6). Experimental data is used to retrain the models, identify new C1s inhibitors, and evaluate model predictions and the pipeline (Chen et al., 2018 p. 6). After the models are retrained, the process is conducted again to determine whether model performance has improved (Chen et al., 2018 p. 6).

The components of the experiments are designed to replicate the enzyme-substrate interaction in cells within the body. Enzymes are complex molecules formed out of a chain of amino acids intended to serve a specific function or product within the cell (Biology…Editors, 2016). Enzymes possess active sites which are areas that can create weak bonds with the substrate (Biology…Editors, 2016). After the substrate fits into the active site/binding cavity, an enzyme substrate complex, which is a temporary molecule, is formed and the shape of the substrate is changed (Biology…Editors, 2016). Once the substrate no longer has its original shape, it can no longer bind to the enzyme and the products of the reaction are released (Biology…Editors, 2016). The enzyme will then repeat this process with another substrate molecule (Biology…Editors, 2016).

For a drug to be effective, the active compound must have the precise size and shape of the binding pocket of target enzyme (Klebe). In addition, the surface properties of the compound/ligand and enzyme must be compatible for interactions with cellular components to take place (Klebe). When the ligand is in the active site/receptor of the enzyme, it gets caught since the typical reaction between the substrate and protein does not occur (Biology…Editors, 2016). The compound, which are inhibitors in the case of this study, are then bound to the enzyme inhibiting its activity and produce the desired biological effect (Klebe). A visual representation of this can be seen in **Figure 2**. Compound activity with the enzyme is measured by obtaining data from a fluorescence scanner since the ligand-enzyme complex fluoresces. By experimentally validating compounds predicted from the pipeline, the predictive power of the models can be analyzed.

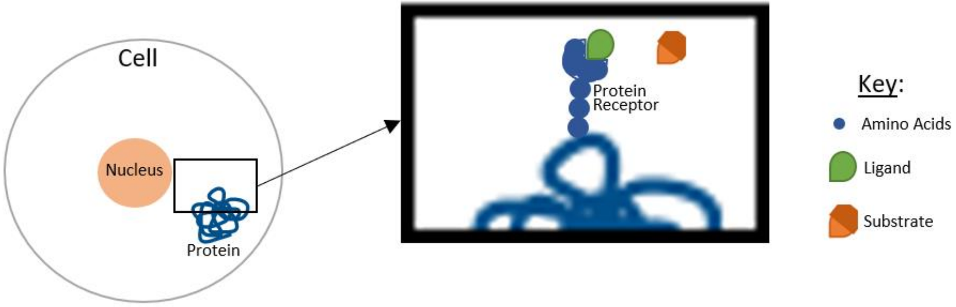


Figure 2: Symbolizes protein in a cell which is then magnified in order to see the protein receptor/active site interaction with a ligand/compound. The substrate is unable to interact with the protein since the ligand is stuck in the active site. Since this is a symbolic representation, colors, sizes, and shapes are not accurately depicted.

Experimental Methods

The purpose of the experiments is to identify compounds that are potential candidates for C1s inhibitors. Recall, C1s initiates the classical pathway so inhibiting C1s is essential to regulating/inhibiting C1 activation in diseases where there are variations in the production of C1 inhibitor C1q (Chen et al., 2018 p. 2). Without an adequate amount of the C1 inhibitor, the activation of C1 is uncontrolled leading to chronic inflammation and tissue damage (Chen et al., 2018 p. 2). The protocol in the PubChem Bioassay AID 787(Diamond, 2008) involves plating the activated human complex factor C1s and substrate solutions to test various potential drug compounds. Compounds that inhibit the biological activity of the activated C1s are deemed active and vice versa. In the experiments conducted, fluorescence is used as a measure of biological activity because the ligand fluoresces after it is processed by the protein. The results from experiments are used to augment the existing data and retrain the models in the algorithms to improve predictive power for a higher hit rate. The higher the hit rate, the more effectively drug candidates are identified. After retraining the models, a second round of compounds is tested to evaluate the pipeline’s ability to make accurate predictions (Chen et al., 2018 p. 11).

Before experimentation, procedure calculations were carried out utilizing the molarity equation (molar concentration = moles solute/liters of solution) and molecular weights to obtain the amount needed of each substance. From these calculations, solutions including assay buffer solution, substrate solution, and protein solutions were prepared. The assay buffer consisted of Millipore water, 50 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) from Sigma Aldrich (Prod. No. H4034), 200 mM sodium chloride (NaCl) from Chem-Impex (CatID 00829; Wood Dale, IL, USA), and 0.2% polyethylene glycol (PEG) from Sigma Aldrich (Prod. No. P 3390) (Chen et al., 2018 p. 17). Since the assay buffer needs to have a pH of 7.5, the pH of the solution was measured using litmus paper and was adjusted by hydrochloric acid (HCl) or sodium hydroxide to make the solution either more acidic or basic, respectively. Maintaining the proper pH is essential for the proper functioning of the protein. The enzyme solution consists of the activated human complement factor C1s from CalbioChem (CatID 204879; Billerica, MA, USA) at a final concentration of 0.02 mg/mL in the assay buffer (Chen et al., 2018 p. 17). The substrate solution is comprised of Boc-Leu-Gly-Arg-AMC from Bachem (CatID I-1105; Torrance, CA, USA) at a final concentration of 15 μM in the assay buffer (Chen et al., 2018 p. 17). Once the solutions are prepared, the remaining materials can be gathered.

All compounds to be tested were from Molport (Riga, Latvia) and were diluted at 50 times concentration in dimethyl sulfoxide (DMSO) (Chen et al., 2018 p. 17). Once the adequate amount of DMSO was added to each compound vile, they were placed on the Vortex machine until the compound completely dissolved. Once the compounds are ready, Corning black polystyrene 96-well, flat bottom plates, which are purchased from Sigma Aldrich (Prod. No. CLS3915), were obtained (Chen et al., 2018 p. 17). The plate has columns labeled 1-12 and rows labeled A-H. A multichannel pipette with twelve channels was used so the components are plated as close to the same time as possible to eliminate any possibility for discrepancies within the data. Once the mentioned materials are gathered, the experimentation process can begin.

The experiment begins with the substrate solution. Row A, columns 1-9 was filled with 66.67 μL of substrate solution while the remainder of the plate rows B-H, columns 1-9, 11-12 and row A, columns 11-12 received 50.0 μL of the substrate solution (Chen et al., 2018 p. 17). Next, 2.66 μL of the first compound was inserted into each well in row A, columns 1-3. The pipette tip was changed after each time the compound was injected into a well to prevent contamination of the compound vile. The next compound was inserted into wells in row A, columns 4-6, again changing the pipette tip after 2.66 μL of the compound was inserted into each well. Finally, 2.66 μL of the last compound was inserted into each well in row A, columns 7-9, again changing the pipette tip after each well. Next, the compounds in row A, columns 1-9 were diluted by eight four-fold dilutions from 2.5 mM to 152.6 nM to achieve the final testing concentrations from 50 μM to 3.05 nM (Chen et al., 2018 p. 17).

To achieve these concentrations, the wells in row A, columns 1-9 were lightly mixed by the multichannel pipette and 16.66 μL was removed from each of those wells and inserted into row B, columns 1-9, where the wells were again mixed with the pipette tip. This same procedure continues for the remaining rows C-H, where 16.66 μL was removed from each well in columns 1-9 from the prior row and inserted and mixed in the next row and so on. After 16.66 μL was removed from each well in row H, columns 1-9 the enzyme solution is ready to be added. The enzyme solution was added, 50 μL into each well, to wells in Rows A-H, columns 1-10 and 12. Next, assay buffer solution was added, 50 μL into each well, to wells in Rows A-H, columns 10 and 11. Then, the plate was covered in foil and incubated for 2.5 hours at room temperature before being scanned for florescence (excitation 355, emission 460) on a Tecan M200 (Chen et al., 2018 p. 17). This procedure was repeated to test 7 compounds in the first pass/round and 10 compounds in the second round.

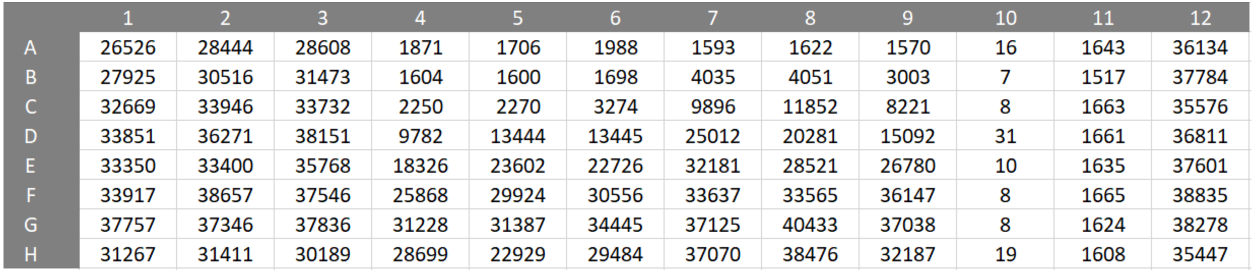
Data and Results

In the experiments conducted, fluorescence signal is used as a measure of biological activity because the ligand fluoresces after it is processed by the protein. Compounds that inhibit the biological activity of the activated C1s are deemed active. Quantitatively, compounds are considered active if they have reported half-maximal inhibitory concentration (IC_{50}) values signifying there is biological activity (Chen et al., 2018 p. 8). IC_{50} values measure a drug's efficacy and indicate the amount of the drug required to inhibit the C1s activity by half (Aykul, 2016). The importance of the reported/average IC_{50} values is that these values are used to retrain the quantitative structure-activity relationship (QSAR) models. The reported IC_{50} values are calculated by taking an average of the IC_{50} values, which are calculated by linear interpolating points straddling 50% inhibition (Chen et al., 2018 p. 18). Compounds that did not have 50% inhibition within the columns in which it was tested were considered inactive (Chen et al., 2018 p. 18). Inactive candidates are those which show limited biological activity and should not be considered for testing. Percent inhibition is calculated by utilizing **Equation 1**. To understand why the signal, blank, and control values are measured, the process of analyzing the data from the Tecan M200, which is a florescence scanner, needs to be understood.

$$\%\, inhibition = 1 - \frac{signal - \overline{blank}}{\overline{control} - \overline{blank}} * 100$$

Equation 1: depicts how to calculate percent inhibition. The signal refers to the fluorescence measurements seen in columns 1-9, rows A-H in **Table 1**. The blank value refers to the average of the substrate fluorescence signals in column 11 in **Table 1**. The control value is an average of the fluorescence signals in column 12 in **Table 1**.
The picture is adapted from Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.

Table 1: depicts unitless measurements from the Tecan M200 of the fluorescence emitted from each well of the 96-well, flat bottom plate. Three compounds are tested per plate. Compounds are labeled by their PubChem ID number (CID). In this table, compound CID 1107361 was plated in columns 1-3, rows A-H; compound CID 2986934 was plated in columns 4-6, rows A-H; and Compound CID 710644 was plated in columns 7-9, rows A-H.

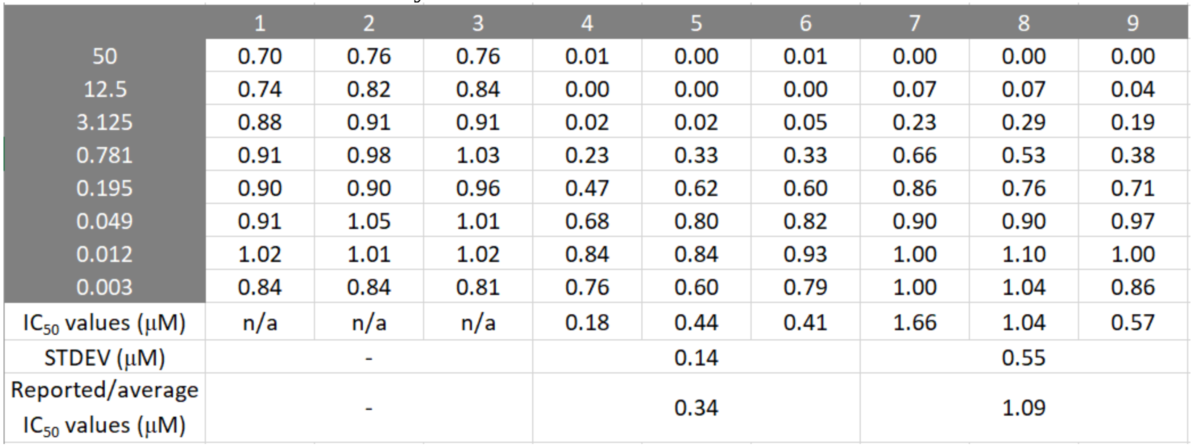


	1	2	3	4	5	6	7	8	9	10	11	12
A	26526	28444	28608	1871	1706	1988	1593	1622	1570	16	1643	36134
B	27925	30516	31473	1604	1600	1698	4035	4051	3003	7	1517	37784
C	32669	33946	33732	2250	2270	3274	9896	11852	8221	8	1663	35576
D	33851	36271	38151	9782	13444	13445	25012	20281	15092	31	1661	36811
E	33350	33400	35768	18326	23602	22726	32181	28521	26780	10	1635	37601
F	33917	38657	37546	25868	29924	30556	33637	33565	36147	8	1665	38835
G	37757	37346	37836	31228	31387	34445	37125	40433	37038	8	1624	38278
H	31267	31411	30189	28699	22929	29484	37070	38476	32187	19	1608	35447

Once the experimental Corning black polystyrene 96-well, flat bottom plate is scanned for florescence (excitation 355, emission 460) on a Tecan M200, its measured values are inputted into an Excel sheet (Chen et al., 2018 p. 17). The results from the first scan of the plate are used to find the maximum florescence signal value in columns 1-9, rows A-H. Once this value is located, the florescence scanner re-reads the plate by referencing the maximum value to give the values as seen in **Table 1**. The reason for referencing the maximum value is to scale the data so any nuances in the fluorescence signal can be detected. The values in **Table 1** are then utilized to calculate the percent inhibition (**Equation 1**), the results of which can be found in **Table 2**.

In **Table 1**, column 10 (rows A-H) serves as the protein check as it consists of 50 mL enzyme solution and 50 mL assay buffer in each well of this column (Chen et al., 2018 p. 17). Column 11 (rows A-H) or the blank is comprised of 50 mL substrate solution and 50 mL assay buffer in each well of this column (Chen et al., 2018 p. 17). The significance of removing the blank/substrate in **Equation 1** is to remove the background signal from analysis, so the fluorescence signal is purely from the ligand-protein interaction. Columns 10 and 11 were used as checks to make sure that neither the substrate nor the protein/enzyme are auto florescent. If either of them was auto florescent, it would be difficult to differentiate between the florescence signal coming from ligand after it is processed by the protein or from the protein and substrate themselves. By testing the substrate and protein/enzyme individually in columns 10 and 11 respectively, it guarantees that any florescence signal coming from wells in columns 1-9, rows A-H are from the ligand-protein interaction. Column 12 serves as the control as it consists of 50 mL substrate solution and 50 mL enzyme solution in each well of this column (Chen et al., 2018 p. 17). Since these same solutions are present in wells in columns 1-9, rows A-H, florescence signals higher than those found in column 12 can be attributed to the compound being auto florescent. Compounds that proved to be auto florescent were removed from analysis.

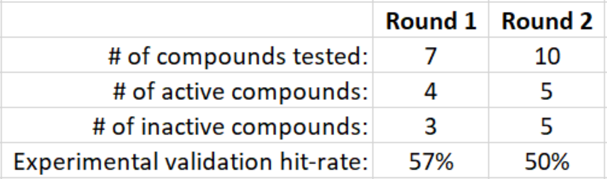
Table 2: shows inhibition fractions that were calculated using **Equation 1**. These calculations can be visualized in **Table 6** in the **Appendix**. From these values, the IC_{50} values are calculated by linear interpolating points straddling 0.5 inhibition and the corresponding dilutions, which are the eight four-fold dilutions starting with 50 micromolar listed in the column in gray. A sample of this calculation can be visualized in **Table 7** in the **Appendix**. The IC_{50} values for each compound are averaged to calculate the reported IC_{50} values. The standard deviations (STDEV) are also calculated from each of the IC_{50} values. For compound CID 1107361 in columns 1-3 there are no IC_{50} values since the measurements never achieved 50% inhibition activity.



	1	2	3	4	5	6	7	8	9
50	0.70	0.76	0.76	0.01	0.00	0.01	0.00	0.00	0.00
12.5	0.74	0.82	0.84	0.00	0.00	0.00	0.07	0.07	0.04
3.125	0.88	0.91	0.91	0.02	0.02	0.05	0.23	0.29	0.19
0.781	0.91	0.98	1.03	0.23	0.33	0.33	0.66	0.53	0.38
0.195	0.90	0.90	0.96	0.47	0.62	0.60	0.86	0.76	0.71
0.049	0.91	1.05	1.01	0.68	0.80	0.82	0.90	0.90	0.97
0.012	1.02	1.01	1.02	0.84	0.84	0.93	1.00	1.10	1.00
0.003	0.84	0.84	0.81	0.76	0.60	0.79	1.00	1.04	0.86
IC_{50} values (μM)	n/a	n/a	n/a	0.18	0.44	0.41	1.66	1.04	0.57
STDEV (μM)		-			0.14			0.55	
Reported/average IC_{50} values (μM)		-			0.34			1.09	

In the first round of experimental testing, seven compounds were tested. Four of these compounds are considered active since they possess reported IC_{50} values. Compounds with no IC_{50} values are deemed inactive because the measurements never achieved 50% inhibition activity, meaning a concentration higher than what was tested is necessary. The highest compound concentration tested in the AID 787 protocol is 50 μM (Diamond, 2008, as cited in Chen et al., 2018 p. 6). Concentrations influence the effectiveness of a drug as it is absorbed into cells and intracellular sites (Steinberg, 1994). An example of an inactive compound is CID 1107361 in columns 1-3 of **Table 2**. As seen in **Table 2**, the measurements never achieved 50% inhibition activity so there are no IC_{50} values and the compound is deemed inactive. A summary of results can be seen in **Table 3**, and the compound structures, reported IC_{50} values, and standard deviations can be seen in **Table 4** for the first round of testing and **Table 5** for the second round. The results from the first round of experiments are used to augment the existing data and retrain the models in the algorithms to improve predictive power for a higher hit rate. After retraining the models, fifty-two compounds were identified and ten were purchased for experimental validation due to financial feasibility (Chen et al., 2018 p. 11). The second round of compounds is tested to evaluate the pipeline's ability to make accurate predictions (Chen et al., 2018 p. 11). Of the ten compounds that were tested, only five are considered active for an experimental validation hit-rate of 50%, which is lower than the first-round experimental validation hit-rate of 57%.

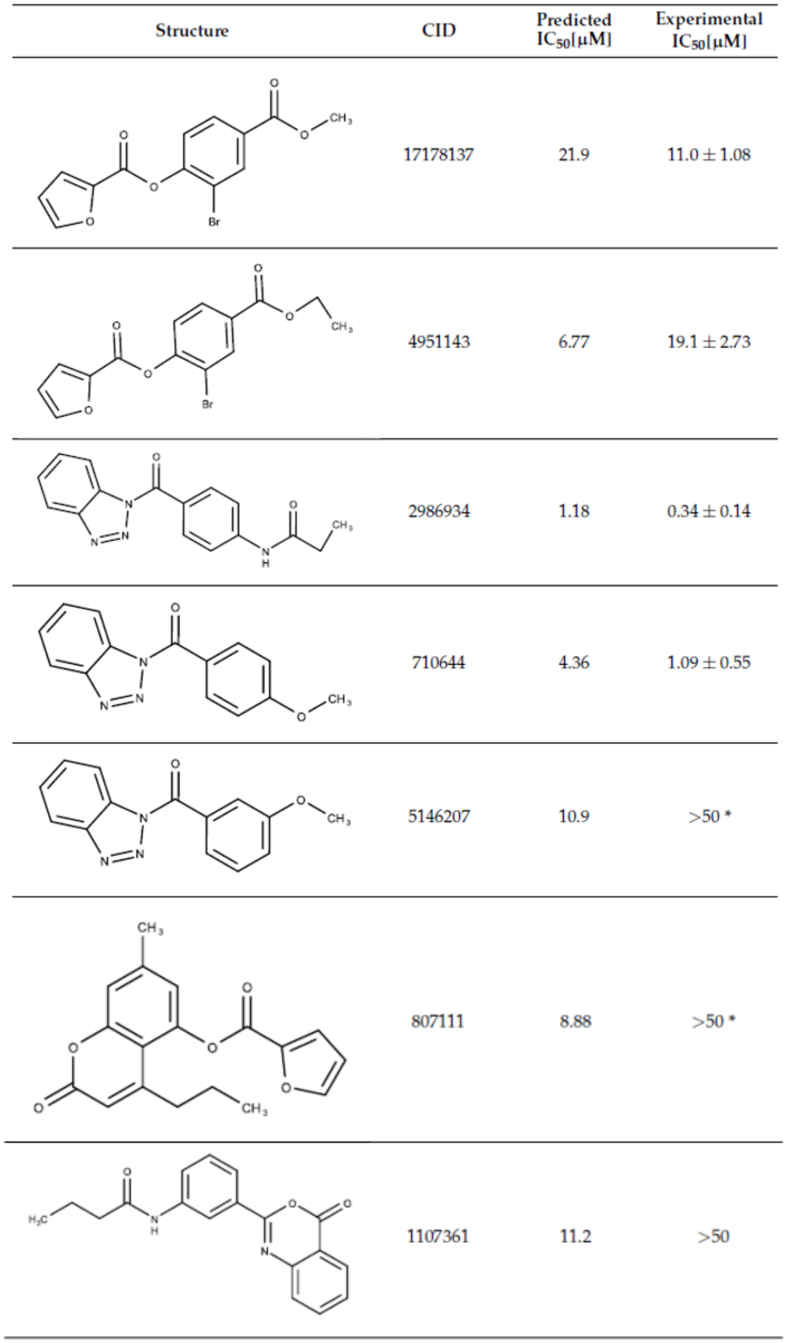
Table 3: depicts the summary of experimental results.



	Round 1	**Round 2**
# of compounds tested:	7	10
# of active compounds:	4	5
# of inactive compounds:	3	5
Experimental validation hit-rate:	57%	50%

Table 4: depicts the compounds that were tested in the first round of experiments. The compound structure, its PubChem ID number (CID), and predicted and experimental IC_{50} values are listed for each compound. Compounds with reported experimental IC_{50} values are considered active while those with ">50*" are deemed inactive. Compounds that require a concentration greater than 50 μM (>50*) are considered inactive and are not recommended for further testing. In this round, a total of 7 compounds were tested, 4 of which are active compounds.

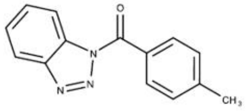
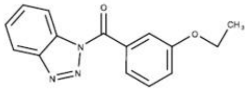
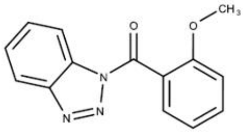
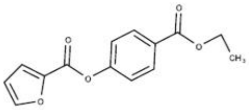
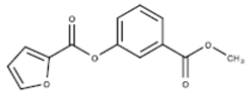
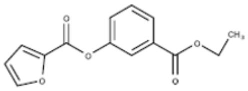
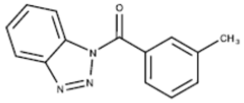
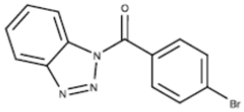
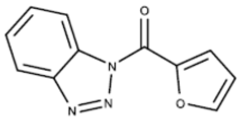
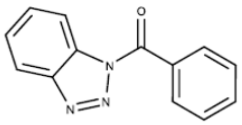
The table is adapted from Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.



Structure	CID	Predicted IC_{50}[μM]	Experimental IC_{50}[μM]
	17178137	21.9	11.0 ± 1.08
	4951143	6.77	19.1 ± 2.73
	2986934	1.18	0.34 ± 0.14
	710644	4.36	1.09 ± 0.55
	5146207	10.9	>50 *
	807111	8.88	>50 *
	1107361	11.2	>50

Table 5: depicts the compounds that were tested in the second round of experiments. The compound structure, its PubChem ID number (CID), and predicted and experimental IC_{50} values are listed for each compound. Compounds with reported experimental IC_{50} values are considered active while those with “>50*” are deemed inactive. Compounds that require a concentration greater than 50 μM (>50*) are considered inactive and are not recommended for further testing. In this round, a total of 10 compounds were tested, 5 of which are active compounds.

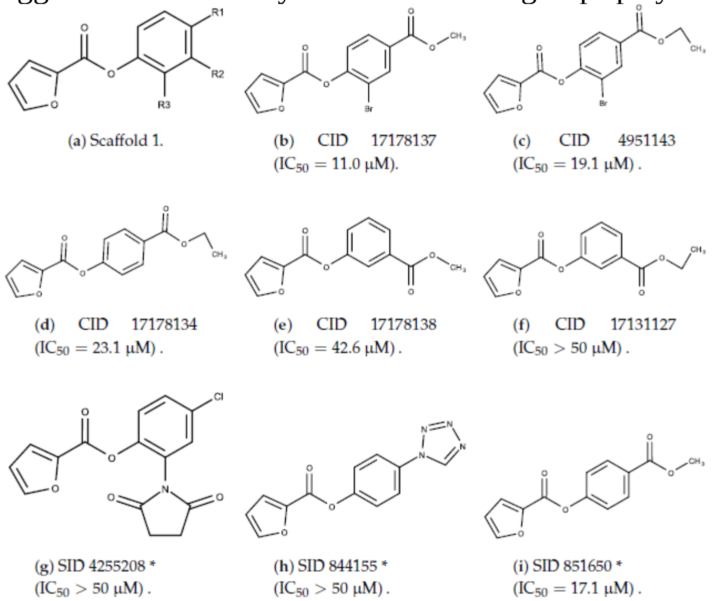
The table is adapted from Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.

Structure	CID	Predicted $IC_{50}[\mu M]$	Experimental $IC_{50}[\mu M]$
	827004	0.30	3.04 ± 1.24
	4957387	4.27	32.9 ± 3.04
	898930	26.01	5.54 ± 1.19
	17178134	7.21	23.1 ± 1.39
	17178138	33.44	42.6 ± 0.72
	17131127	23.71	>50 *
	834536	1.66	>50 *
	693001	0.43	>50 *
	792914	2.05	>50 *
	570059	4.20	>50 *

Discussion/Analysis

In total, 17 compounds were selected for experimentation based on pipeline outcomes, financial feasibility, and accessibility from suppliers (Chen et al., 2018 p.13). Experimental results validated the activity of 9 compounds in total from the two rounds of testing. Four active compounds were validated out of the seven tested for the first round for a hit rate of 57%. After the models were retrained, ten compounds were tested of which five were active for a hit rate of 50%. Ideally, the hit rate would improve (>57%) after the models are retrained reflecting that the drug candidates are more effectively identified. However, the reason for the potential drop is the result of the extrapolation required to determine compounds for the second round of testing (Chen et al., 2018 p.18). The 9 compounds showing activity are promising drug candidates as they have the capability to inhibit C1s. These compounds are recommended for further testing, such as at the cellular level, to see the effects on C1s inhibition (Chen et al., 2018 p.13). From observing the structures of the compounds that were experimentally tested in **Tables 4-5**, its apparent two core molecular structures or scaffolds are prominent.

For analysis, all tested compounds, including inactive compounds, are broken into two categories so inferences can be made on the effects functional groups have on activity. From examining the nine compounds containing the first scaffold two major inferences can be made. First, activity is largely impacted by the positions of large functional groups (Chen et al., 2018 p.13). This inference is supported by **Figure 3 (d),(f)**. The compound structures are the same besides one as an ester group in a para position (IC_{50} =23.1 μM) and the other has an ester group in a meta position (IC_{50} >50 μM) (Chen et al., 2018 p.13). Secondly, activity is largely impacted by the identity of functional groups as seen in **Figure 3 (h),(i)** (Chen et al., 2018 p.13). Both have identical structures besides the functional groups in the para positions. The one with an ester **Figure 3 (i)** has an IC_{50} =17.1 μM while the other (h) has an IC_{50} >50 μM (Chen et al., 2018 p.14). These results suggest that the identity of the functional groups plays a role in activity.



(a) Scaffold 1.

(b) CID 17178137 (IC_{50} = 11.0 μM).

(c) CID 4951143 (IC_{50} = 19.1 μM) .

(d) CID 17178134 (IC_{50} = 23.1 μM) .

(e) CID 17178138 (IC_{50} = 42.6 μM) .

(f) CID 17131127 (IC_{50} > 50 μM) .

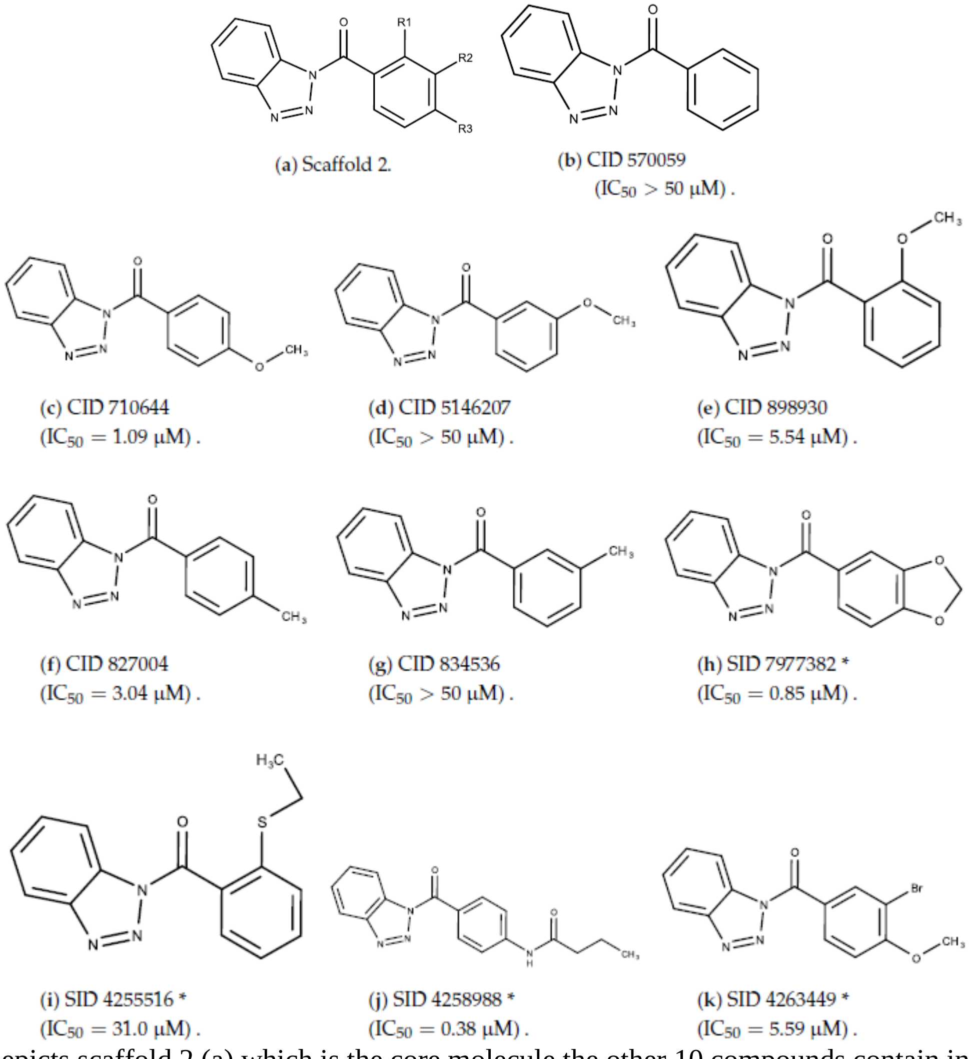
(g) SID 4255208 * (IC_{50} > 50 μM) .

(h) SID 844155 * (IC_{50} > 50 μM) .

(i) SID 851650 * (IC_{50} = 17.1 μM) .

Figure 3: depicts scaffold 1 (a) which is the core molecule the other 8 compounds contain in their structures. Compounds with an asterisk(*) were included in the first round of experimental testing. The table is adapted from Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.

The conclusions drawn from the structures in the first scaffold were also supported by the second scaffold pictured in **Figure 4 (a)**. The impact of functional group position on activity is again supported by **Figure 4 (c),(d),(e)** where the ether group is located at para(IC_{50} =1.09.1 μM), meta(IC_{50} >50 μM), and ortho(IC_{50} =5.54 μM) positions respectively (Chen et al., 2018 p.14). The para position had the highest activity and the meta position had the lowest, similar to the finding in the first scaffold. **Figure 4 (f),(g)** also follows this pattern with the methyl group located at para(IC_{50} =3.04 μM) and meta(IC_{50} >50 μM) positions (Chen et al., 2018 p.14). The impact of the identity of the functional group being important is reinforced by **Figure 4 (e),(i)** where differing functional groups in the ortho position resulted in differing IC_{50} values, 5.54 μM and 31.0 μM, respectively (Chen et al., 2018 p.14).



(a) Scaffold 2.

(b) CID 570059 ($IC_{50} > 50$ μM) .

(c) CID 710644 ($IC_{50} = 1.09$ μM) .

(d) CID 5146207 ($IC_{50} > 50$ μM) .

(e) CID 898930 ($IC_{50} = 5.54$ μM) .

(f) CID 827004 ($IC_{50} = 3.04$ μM) .

(g) CID 834536 ($IC_{50} > 50$ μM) .

(h) SID 7977382 * ($IC_{50} = 0.85$ μM) .

(i) SID 4255516 * ($IC_{50} = 31.0$ μM) .

(j) SID 4258988 * ($IC_{50} = 0.38$ μM) .

(k) SID 4263449 * ($IC_{50} = 5.59$ μM) .

Figure 4: depicts scaffold 2 (a) which is the core molecule the other 10 compounds contain in their structures. Compounds with an asterisk(*) were included in the first round of experimental testing. The table is adapted from Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018. Copyright © 2018 by the authors.

It is important to remember the study described, AID 787(Diamond, 2008), is a part of a larger study to test the predictability of the pipeline in vHTS as it is applied to various systems (Chen et al., 2018 p.18). Systems vary by data size, classification (active/inactive) distribution, and benefit of model retraining (Chen et al., 2018 p.18). In AID 787(Diamond, 2008), the fraction of actives predicted from the training set was 0.12, a fraction smaller than in other studies that contribute to the larger endeavor (Chen et al., 2018 p.13). These findings suggest the pipeline and models have the capability to be applied in cases with a limited amount of data on the desired active compounds (Chen et al., 2018 p.13). Through the results of the studies in this series, the range of applications that the pipeline can be utilized is further defined in order to support and enrich future drug discovery efforts (Chen et al., 2018 p.18).

Summary/Conclusions

The 9 compounds showing activity are promising drug candidates as they have the capability to inhibit C1s. These compounds are recommended for further testing, such as at the cellular level, to see the effects on C1s inhibition (Chen et al., 2018 p.13). From observing the structures of the compounds that were experimentally tested in **Tables 4-5**, its apparent two core molecular structures or scaffolds are prominent, which can be seen in **Figures 3-4**. From comparing compounds, inferences can be made on how activity is related to both positions of large functional groups and the identity of functional groups, which is detailed in the **Discussion/Analysis** section (Chen et al., 2018 p.13). Moreover, in the PubChem Bioassay ID (AID) 787(Diamond, 2008) dataset, the fraction of actives predicted from the training set was 0.12, a fraction smaller than in other studies that contribute to the larger endeavor (Chen et al., 2018 p.13). These findings suggest the pipeline and models have the capability to be applied in cases with a limited amount of data on the desired active compounds (Chen et al., 2018 p.13). Through the results of the studies in this series, the range of applications that the pipeline can be utilized is further defined in order to support and enrich future drug discovery efforts (Chen et al., 2018 p.18). Applications that the pipeline can be applied to should continue to be explored as its potential to speed-up the process of developing medicine by using computer algorithms and virtual high-throughput screens is promising.

References

Aykul, Senem and Erik Martinez-Hackert. “Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis.” *Analytical biochemistry* 508 (2016): 97-103.

Baell, J.B.; Holloway, G.A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* 2010, *53*, 2719–2740.

Biologydictionary.net Editors. “Enzyme Substrate Complex.” *Biology Dictionary*, Biologydictionary.net, 04 Dec. 2016, https://biologydictionary.net/enzyme-substrate-complex/.

Buerke, M.; Schwertz, H.; Seitz, W.; Meyer, J.; Darius, H. Novel small molecule inhibitor of C1s exerts cardioprotective effects in ischemia-reperfusion injury in rabbits. *J. Immunol.* 2001, *167*, 5375-5380.

Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018, https://doi.org/10.3390/biom8020024

Diamond, S.L. AID 787-Complement Factor C1s IC150 from Mixture Screen. 2008. Available online: https://pubchem.ncbi.nlm.nih.gov/bioassay/787 (accessed on 23 June 2017).

Gani, Zaahira. “Complement System.” *British Society for Immunology*, British Society for Immunology, www.immunology.org/public-information/bitesized-immunology/systems-and-processes/complement-system.

“Hereditary Angioedema - Genetics Home Reference - NIH.” *U.S. National Library of Medicine*, National Institutes of Health, ghr.nlm.nih.gov/condition/hereditary-angioedema.

Klebe G. (2013) Protein–Ligand Interactions as the Basis for Drug Action. In: Klebe G. (eds) Drug Design. Springer, Berlin, Heidelberg

Ratnoff, O.D.; Lepow, I.H. Some properties of an esterase derived from preparations of the first component of complement. *J. Exp. Med.* 1957, *106*, 327-343.

Steinberg, Thomas H., Cellular Transport of Drugs, *Clinical Infectious Diseases*, Volume 19, Issue 5, November 1994, Pages 916–921, https://doi.org/10.1093/clinids/19.5.916

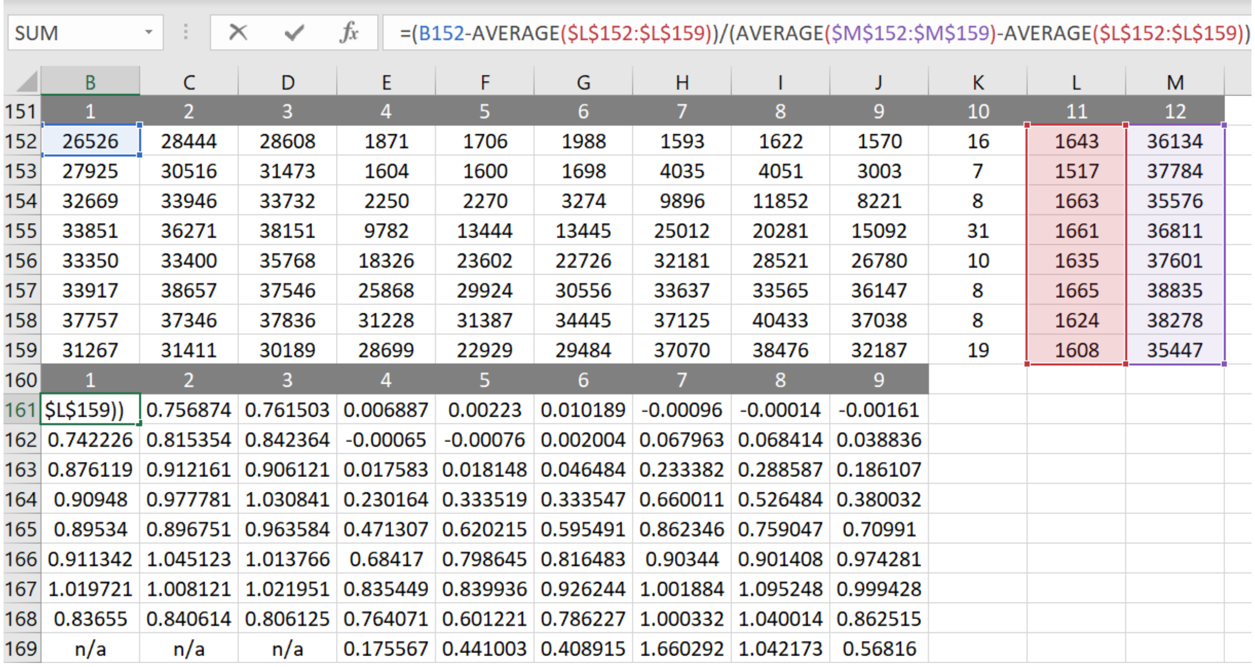
Vulpetti A., Randl S., Rüdisser S., Ostermann N., Erbel P., Mac Sweeney A., Zoller T., Salem B., Gerhartz B., Cumin F., et al. Structure-based library design and fragment screening for the identification of reversible complement Factor D protease inhibitors. *J. Med. Chem*. 2017;*60*:1946–1958. doi: 10.1021/acs.jmedchem.6b01684.

Whitley D. A genetic algorithm tutorial. Stat. Comput. 1994;4:65–85. doi: 10.1007/BF00175354.

Appendix

Sample Calculations

Table 6: shows how the inhibition fractions are calculated to get the values shown in **Table 2**. The cell bordered in green depicts how the inhibition values are calculated using **Equation 1** and the values in **Table 1**.

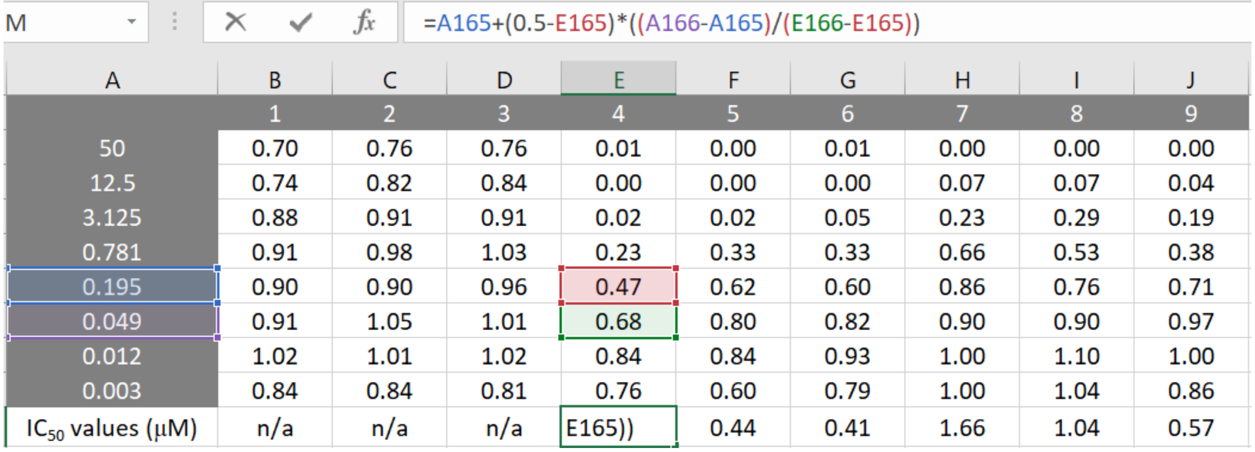


SUM

=(B152-AVERAGE($L$152:$L$159))/(AVERAGE($M$152:$M$159)-AVERAGE($L$152:$L$159))

	B	C	D	E	F	G	H	I	J	K	L	M
151	1	2	3	4	5	6	7	8	9	10	11	12
152	26526	28444	28608	1871	1706	1988	1593	1622	1570	16	1643	36134
153	27925	30516	31473	1604	1600	1698	4035	4051	3003	7	1517	37784
154	32669	33946	33732	2250	2270	3274	9896	11852	8221	8	1663	35576
155	33851	36271	38151	9782	13444	13445	25012	20281	15092	31	1661	36811
156	33350	33400	35768	18326	23602	22726	32181	28521	26780	10	1635	37601
157	33917	38657	37546	25868	29924	30556	33637	33565	36147	8	1665	38835
158	37757	37346	37836	31228	31387	34445	37125	40433	37038	8	1624	38278
159	31267	31411	30189	28699	22929	29484	37070	38476	32187	19	1608	35447
160	1	2	3	4	5	6	7	8	9			
161	$L$159))	0.756874	0.761503	0.006887	0.00223	0.010189	-0.00096	-0.00014	-0.00161			
162	0.742226	0.815354	0.842364	-0.00065	-0.00076	0.002004	0.067963	0.068414	0.038836			
163	0.876119	0.912161	0.906121	0.017583	0.018148	0.046484	0.233382	0.288587	0.186107			
164	0.90948	0.977781	1.030841	0.230164	0.333519	0.333547	0.660011	0.526484	0.380032			
165	0.89534	0.896751	0.963584	0.471307	0.620215	0.595491	0.862346	0.759047	0.70991			
166	0.911342	1.045123	1.013766	0.68417	0.798645	0.816483	0.90344	0.901408	0.974281			
167	1.019721	1.008121	1.021951	0.835449	0.839936	0.926244	1.001884	1.095248	0.999428			
168	0.83655	0.840614	0.806125	0.764071	0.601221	0.786227	1.000332	1.040014	0.862515			
169	n/a	n/a	n/a	0.175567	0.441003	0.408915	1.660292	1.042173	0.56816			

Table 7: shows how the IC$_{50}$ values are calculated by linear interpolating points straddling 0.5 inhibition and the corresponding dilutions, which are the eight four-fold dilutions starting with 50 micromolar listed in the column in gray. For compound CID 1107361 in columns 1-3 there are no IC$_{50}$ values since the measurements never achieved 50% inhibition activity.



M

=A165+(0.5-E165)*((A166-A165)/(E166-E165))

A	B	C	D	E	F	G	H	I	J
	1	2	3	4	5	6	7	8	9
50	0.70	0.76	0.76	0.01	0.00	0.01	0.00	0.00	0.00
12.5	0.74	0.82	0.84	0.00	0.00	0.00	0.07	0.07	0.04
3.125	0.88	0.91	0.91	0.02	0.02	0.05	0.23	0.29	0.19
0.781	0.91	0.98	1.03	0.23	0.33	0.33	0.66	0.53	0.38
0.195	0.90	0.90	0.96	0.47	0.62	0.60	0.86	0.76	0.71
0.049	0.91	1.05	1.01	0.68	0.80	0.82	0.90	0.90	0.97
0.012	1.02	1.01	1.02	0.84	0.84	0.93	1.00	1.10	1.00
0.003	0.84	0.84	0.81	0.76	0.60	0.79	1.00	1.04	0.86
IC$_{50}$ values (µM)	n/a	n/a	n/a	E165))	0.44	0.41	1.66	1.04	0.57

Honors Abstracts Addendum

A large study was conducted to identify possible drug candidates to treat various diseases using computer algorithms, virtual high-throughput screens, and experimental validation of activity. Since the algorithm has the capability to screen millions of compounds, it is beneficial to pharmaceutical development as it allows a larger pool of compounds to be considered that would have otherwise been overlooked. As part of this larger study, this project attempts to identify drug candidates to treat human complement factor C1, a protein which causes tissue damage when underregulated[1]. A series of designed experiments validate candidates and confirm the performance of the algorithm. After the first round of experiments, the compounds identified through the virtual high-throughput screening had a 57% hit rate of potential compounds and the second round after re-training the models was a 50% hit rate[1]. By analyzing results from the experiments, potential drug candidates targeting complement factor C1 were identified for additional study. Furthermore, structural analysis of the identified candidates can pinpoint certain features of the compounds resulting in potential leads for further investigation.

[1] Chen, J. J., Schmucker, L. N., & Visco, D. P. Pharmaceutical Machine Learning: Virtual High-Throughput Screens Identifying Promising and Economical Small Molecule Inhibitors of Complement Factor C1s. *Biomolecules*, *8*(2), 24. 7 May 2018, https://doi.org/10.3390/biom8020024