CASE STUDY

CELERIS

Identification of Novel West Nile Virus NS2B/NS3 Protease Inhibitors

In this whitepaper we showcase the utilities of Xanthos Match Maker™ in rapidly and accurately predicting biomolecular interactions between drug/target pairs.

TECHNOLOGY:
Xanthos Match Maker™, Transfer Learning, Data Augmentation, Active Learning

ABSTRACT:

In this case study, we focus on identifying candidates in a library of DrugBank [1,2] compounds that feature high inhibitory activity against the NS2B/NS3 protease of the West Nile virus (WNV). Our employed technology is based on machine-learning-enabled virtual screening. To demonstrate the validity of our method, we augment the library of candidate compounds with a set of experimentally verified inhibitors of NS2B/NS3. These positive control compounds appear densely accumulated in the top ranks of our proposed lead candidates, substantiating the predictive power of our approach.

INTRODUCTION:

Proteases of *Flaviviridae* have gained considerable interest as potential drug targets to treat infectious diseases. However, particularly for WNV, nearly two decades of drug discovery campaigns were unable to yield the desired success. According to information published online by the CDC [3], neither a specific antiviral drug nor a vaccine exists at the time of this writing [4]. Adding in the recent spike of cases in Europe, as reported by the ECDC [5], the situation should be considered alarming.

Alongside vaccines, the development of antiviral drugs is therefore of significant importance. A promising strategy targets the viral reproduction process inside the cell. The genomes of *Flaviviridae* encode for a single polyprotein. This amino-acid strand is cleaved by a combination of host and viral proteases into individual proteins. Considering WNV, the viral NS2B/NS3 protease is responsible for the cleavage of the polyprotein at various sites. This protease is therefore essential for the viral replication process, rendering NS2B/NS3 a promising drug target for the treatment of West Nile virus infections [6].

METHODS:

The details of the technology we are employing are kept confidential at large [7]. Here, the general aspects of related deep-learning approaches are discussed. Canonically, such approaches rely on establishing a descriptive, machine-learned representation of the compounds as well as the protein. These representations are forwarded deeper into a neural network, where the information is processed into an increasingly high-level form. To give a somewhat crude example, this high-level information can be thought of as an educated opinion of a super-human WNV expert about a particular drug. At the end of the information processing pipeline, the opinion is expressed as a score for the given drug. This number indicates the NS2B/NS3 protease inhibition activity. Based on these predictions, we identify promising therapeutic compounds. A major advantage of such approaches is constituted by their considerable performance in terms of calculation time for predictions (i. e.

inference). This property enables scanning a library of significant size (i. e. on the magnitude of millions or billions of compounds), often even without the need for exhaustive computational resources.

Intrinsic to the application of virtual screening in a production environment, the following important aspect needs to be addressed: The 1492 DrugBank [1,2] compounds of our candidate set, i. e. the set of compounds we are primarily interested in investigating, have not yet been assayed experimentally for their inhibition activity. As a direct consequence, a quality control mechanism suggests itself as mandatory. We provide such a mechanism by including experimentally known inhibitors. These compounds serve as a positive control. If our employed approach features predictive power, then these compounds are expected to rank high in prediction scores. *Vice versa*, given the predictive power is demonstrated in this manner, highly scored compounds from our candidate set suggest themselves as promising for further investigation.

We seek to demonstrate the predictive power based on two methods. As already discussed, the first and more intuitive control mechanism is constituted simply by the experimentally known inhibitors showing up high in the rankings. Second, we provide average prediction scores and standard deviations for both our candidate and known inhibitor set. With the reasonable assumption that the number of actual inhibitors in our candidate set is small, the average scores of the known inhibitors should be significantly higher than the average scores of the candidates.

RESULTS:

Table 1 demonstrates the predictive power of our applied approach as detailed above in the Methods section.

Rank	Compound Type	Inhibition Score
1	Experimentally confirmed inhibitor	0,83
2	Proposed candidate (DrugBank compound)	0,80
3	Proposed candidate (DrugBank compound)	0,79
4	Proposed candidate (DrugBank compound)	0,78
5	Experimentally confirmed inhibitor	0,78
6	Proposed candidate (DrugBank compound)	0,78
7	Proposed candidate (DrugBank compound)	0,75
8	Proposed candidate (DrugBank compound)	0,75
9	Proposed candidate (DrugBank compound)	0,74
10	Proposed candidate (DrugBank compound)	0,74
11	Proposed candidate (DrugBank compound)	0,74
12	Experimentally confirmed inhibitor	0,73
13	Experimentally confirmed inhibitor	0,73
14	Experimentally confirmed inhibitor	0,73
15	Proposed candidate (DrugBank compound)	0,73
16	Proposed candidate (DrugBank compound)	0,73
17	Proposed candidate (DrugBank compound)	0,72
18	Experimentally confirmed inhibitor	0,72
19	Proposed candidate (DrugBank compound)	0,72
20	Experimentally confirmed inhibitor	0,72
21	Proposed candidate (DrugBank compound)	0,71
22	Proposed candidate (DrugBank compound)	0,71
23	Experimentally confirmed inhibitor	0,71
24	Proposed candidate (DrugBank compound)	0,71
25	Experimentally confirmed inhibitor	0,71

Table 1: Anonymized top 25-ranked compounds.

Table 1: The table shows a mixture of experimentally known inhibitors of WNV NS2B/NS3 protease inhibitors and DrugBank [1,2] compounds. Based on the inhibition activity score of our employed approach, these are the top-25 compounds identified. The proposed candidates

have not yet been characterized experimentally for their inhibition activity of NS2B/NS3 protease. Therefore, they represent promising candidates for further investigations in a follow-up drug-repurposing study.

The average inhibition activity scores of the 1492 DrugBank [1,2] compounds and the 144 experimentally confirmed inhibitors of WNV NS2B/NS3 protease are 0.413 ± 0.160 and 0.287 ± 0.177, respectively. The calculated p-value from an unpaired Student's t-test is $<10^{-4}$. Although this test assumes a Gaussian distribution (see forward Figure 2), given the above extremely significant p-value, the amount of inaccuracy introduced from this crude assumption can arguably be considered an academic discussion. Taken at face value, the result of the test indicates that a result of at least such high quality is obtained, by pure chance, in less than one per ten thousand cases. We would like to suggest that these results clearly demonstrate the predictive power of our employed approach.

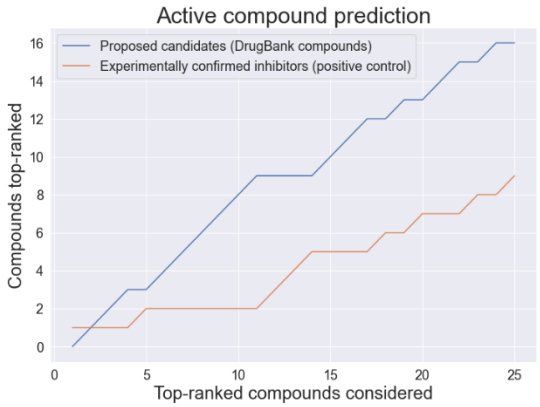

Figure 1: Graphical wrap-up of Table 1.

Figure 1: The x-axis describes the number of compounds considered, ordered by their ranking based on their inhibition activity score. The y-axis indicates how many compounds of the DrugBank [1,2] candidate set or the experimentally confirmed positive controls are found within the top ranks on the x-axis. E. g. at x = 5, there are two positive control and three candidate compounds, in accordance with the top-5 compounds in Table 1.

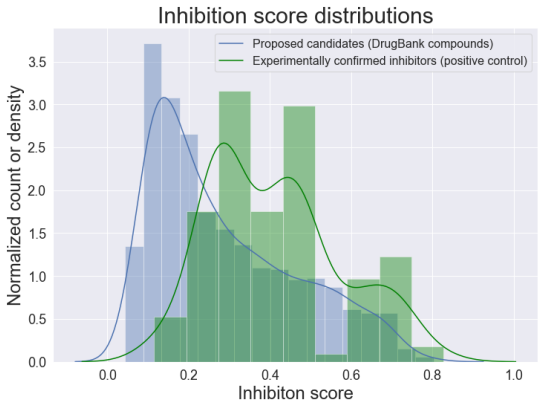

Figure 2: Rank-dependent prediction score.

Figure 2: The distribution of the inhibition activity scores is illustrated. The x-axis describes the inhibition scores as calculated by the applied approach. The histograms on the y-axis indicate how often we count an inhibition score in the respective bin. The smooth curves stem from Gaussian kernel density estimates. Both the histograms and the density estimates are normalized for the total area to equal one. The experimentally confirmed inhibitor compounds are scored significantly higher on average.

Figure 1 represents the information in Table 1 graphically. Note that there are more than 10 times as many compounds in the candidate set compared to the positive control set. Therefore, if predictions were made on a purely random basis, Figure 1 would show a ratio between the curves fluctuating around 10 along the x-axis. This is clearly not the case. It should be fair to say that the value of the ratio is rather a factor of two across a wide range of the x-axis. This fact can be considered a graphical demonstration of the predictive power. Furthermore, based on these approximate numbers and compared to random guessing, one might be tempted to assume an increase in predictive power by a factor of roughly five. However, this factor five increase in predictive power constitutes only a lower bound to

the true value. The exact factor is impossible to estimate; the proposed candidates do not pose true negatives. Instead, the proposed candidates could (and should) contain true inhibitors, i. e. true positives. Table 1 as well as the calculated p-values provide considerable evidence for this suggestion.

Figure 2 demonstrates a clear separation of the distributions of the two subsets, respectively. Except for the right-most peak, the applied approach rather seems to underscore the experimentally confirmed inhibitor set. This fact further emphasizes the top ranked DrugBank [1,2] compounds as highly promising candidates for further investigations. Furthemore, many compounds of the experimentally verified, positive control set feature more mediocre or even low scores. Here, we note that the applied approach has not yet been fine-tuned to the specifics of WNV NS2B/NS3 protease. The solid results of the present study hold out an encouraging prospect for the future performance of the applied machine-learning approach when fully engineered.

CONCLUSION AND OUTLOOK:

Based on a machine learning approach, the presented drug-repurposing case study has arguably demonstrated promising as well as competitive results. Two major aspects are left to be addressed. First, due to its nature of being a drug-repurposing study, only a small data set is screened. The timescale for screening this set of ~1600 compounds is on the order of seconds, given a single, affordable workstation. Scaling this up to a mediocre-sized High-Performance-Computing Cluster, billions of compounds can be scanned in a matter of days. Such a remarkable throughput seems unreached by classical virtual screening methods such as docking. Second, we have conducted the present study at the bleeding-edge process of embedding this exciting technology into our Xanthos Match Maker™ platform, leaving plenty of room for engineering and fine-tuning. With the solid and competitive results presented at this early stage, the true potential of this approach remains to be revealed.

REFERENCES:

[1] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.

[2] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

[3] https://www.cdc.gov/westnile/statsmaps/index.html

[4] https://www.cdc.gov/westnile/symptoms/index.html

[5] https://www.ecdc.europa.eu/en/news-events/epidemiological-update-west-nile-virus-transmission-season-europe-2018

[6] Nitsche C. Proteases from dengue, West Nile and Zika viruses as drug targets. Biophys Rev. 2019 Apr;11(2):157-165. doi: 10.1007/s12551-019-00508-3. Epub 2019 Feb 26. PMID: 30806881; PMCID: PMC6441445.

[7] Identification of Drug-Target Interactions using Xanthos Match Maker™ , https://celeristx.com/, 2021