Machine learning models identify molecules active against the Ebola virus in vitro [version 3; peer review: 2 approved]

Sean Ekins1-3, Joel S. Freundlich4, Alex M. Clark5, Manu Anantpadma6, Robert A. Davey6, Peter Madrid7

1 Collaborations in Chemistry, Fuquay-Varina, NC, 27526, USA
2 Collaborations Pharmaceuticals Inc, Fuquay-Varina, NC, 27526, USA
3 Collaborative Drug Discovery, Burlingame, CA, 94010, USA
4 Departments of Pharmacology & Physiology and Medicine, Center for Emerging and Reemerging Pathogens, UMDNJ, New Jersey Medical School, Newark, NJ, 07103, USA
5 Molecular Materials Informatics, Inc., Montreal, 94025, Canada
6 Texas Biomedical Research Institute, San Antonio, TX, 78227, USA
7 SRI International, Menlo Park, CA, 94025, USA

Abstract
The search for small molecule inhibitors of Ebola virus (EBOV) has led to several high throughput screens over the past 3 years. These have identified a range of FDA-approved active pharmaceutical ingredients (APIs) with anti-EBOV activity in vitro and several of which are also active in a mouse infection model. There are millions of additional commercially-available molecules that could be screened for potential activities as anti-EBOV compounds. One way to prioritize compounds for testing is to generate computational models based on the high throughput screening data and then virtually screen compound libraries. In the current study, we have generated Bayesian machine learning models with viral pseudotype entry assay and the EBOV replication assay data. We have validated the models internally and externally. We have also used these models to computationally score the MicroSource library of drugs to select those likely to be potential inhibitors. Three of the highest scoring molecules that were not in the model training sets, quinacrine, pyronaridine and tilorone, were tested in vitro and had EC\textsubscript{50} values of 350, 420 and 230 nM, respectively. Pyronaridine is a component of a combination therapy for malaria that was recently approved by the European Medicines Agency, which may make it more readily accessible for clinical testing. Like other known antimalarial drugs active against EBOV, it shares the 4-aminoquinoline scaffold. Tilorone, is an investigational antiviral agent that has shown a broad array of biological activities including cell growth inhibition in cancer cells, antifibrotic properties, α7 nicotinic receptor agonist activity, radioprotective activity and activation of hypoxia inducible factor-1.
Quinacrine is an antimalarial but also has use as an anthelmintic. Our results suggest data sets with less than 1,000 molecules can produce validated machine learning models that can in turn be utilized to identify novel EBOV inhibitors in vitro.

**Keywords**
Drug repurposing, Ebola Virus, Computational models, Machine learning, Pharmacophore, Pyronaridine, Quinacrine, Tilorone

This article is included in the Disease Outbreaks gateway.

This article is included in the Machine learning: life sciences collection.

**Corresponding author:** Sean Ekins (ekinssean@yahoo.com)

**Competing interests:** S.E. works for Collaborations in Chemistry, and Collaborations Pharmaceuticals, Inc. and S.E. and A.M.C. consult for Collaborative Drug Discovery Inc.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2017 Ekins S et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

**How to cite this article:** Ekins S, Freundlich JS, Clark AM et al. *Machine learning models identify molecules active against the Ebola virus in vitro* [version 3; peer review: 2 approved] F1000Research 2017, 4:1091 (https://doi.org/10.12688/f1000research.7217.3)

**First published:** 20 Oct 2015, 4:1091 (https://doi.org/10.12688/f1000research.7217.1)
Introduction

In 2014, the outbreak of the Ebola virus (EBOV) in West Africa highlighted the need for broad-spectrum antiviral drugs for this and other emerging viruses. Several groups had previously performed high throughput screens (HTS) and identified FDA approved drugs (amodiaquine, chloroquine, clomiphene and toremifene) with in vitro growth inhibitory activities against EBOV. It appears none of these molecules were tried during the epidemic in Africa, likely due to the lack of efficacy data in higher order species. We have previously summarized the numerous small molecules described in the literature as possessing antiviral activity that could be further evaluated for their potential EBOV activity alongside the few new antivirals. We have found that there is considerable prior knowledge regarding these small molecules possessing activity against EBOV in vitro or in animal models, and this includes a number of accessible FDA-approved drugs. Another recent study has shown three approved ion channel blockers (amiodarone, dronedarone, and verapamil) inhibited EBOV cellular entry. The drugs were given at concentrations that would be achieved in human serum, and were effective against several of the filoviruses. None of the FDA approved drugs described in these various studies were designed to target the Ebola virus. For example amodiaquine and chloroquine are well known antimalarials, clomiphene and toremifene are selective estrogen receptor modulators, while amiodarone, dronedarone, and verapamil are anti-arrhythmics. It may or may not be of importance but all of these compounds have a common tertiary amine feature. What is important is that they are all orally bioavailable and generally safe for humans at their approved doses. Some have suggested that G-protein-coupled receptors (GPCRs) may play a role in filoviral entry and receptor antagonists could be developed as anti-EBOV therapies. The compounds which are FDA-approved drugs for other diseases but with activity against EBOV in vitro or in vivo may represent useful starting points with the advantage that much is known regarding their absorption, distribution, metabolism and excretion (ADME) and toxicity properties. Thus, these repurposed drugs may represent a more advanced starting point for therapeutic development and approval compared with new chemical entities for preventing the spread and mortality associated with EBOV.

Beyond these early stage drugs, there are a number of other compounds that have also been identified as active against EBOV (summarized in a review). A thorough literature search identified 55 molecules suggested to have activity against EBOV in vitro and/or in vivo which were evaluated from the perspective of an experienced medicinal chemist as well as using simple molecular properties and ultimately 16 were highlighted as desirable. This dataset overlaps to some extent with another review that identified over 60 molecules. Two recent repurposing screens identified 53 and 80 compounds with antiviral activity which also overlap the earlier screens. Additional studies have identified small number of inhibitors. In total there may now be close to several hundred compounds identified with activity against EBOV in vitro.

Approaches with more capacity to screen compounds include using computational methods as a filter before in vitro testing. Computational models for anti-EBOV activity include one which used the average quasi valence number (AQVN) and the electron-ion interaction potential (EIIP), parameters determining long-range interaction between biological molecules for virtual screening of DrugBank and suggested hundreds of compounds to test. A follow up to this study proposed ibuprofen for testing. Others have also used computational docking studies to propose multi-target inhibitors of VP40, VP35, VP30 and VP24, inhibitors of VP40 or have suggested molecules to test in the absence of computational approaches. We are unaware of any validation of these compounds. A further computational approach used a pharmacophore that was generated from four FDA approved compounds resulting from the two earliest high throughput screens against EBOV. This pharmacophore closely matched the receptor-ligand pharmacophores for the EBOV protein 35 (VP35). Follow-up docking studies suggested that these compounds may also have favorable inhibitory interactions with this receptor. The pharmacophore was further used to screen several compound libraries. We proposed that if we could learn from the many compounds already screened for anti-EBOV activity, we could more efficiently find additional compounds and perhaps understand the key molecular features needed for antiviral activity. We speculated then that Laplacian-corrected Naïve Bayesian classifier models might be useful as they have been for M. tuberculosis and more recently for T. cruzi. To our knowledge machine learning approaches to identify EBOV inhibitors have not been attempted elsewhere. The current study extends the machine learning approach to EBOV and uses both commercially available Bayesian, Support Vector Machines (SVM) and recursive partitioning methods and open source Bayesian software for model generation and compound scoring. We report the identification of three novel EBOV inhibitors with nanomolar EC$_{50}$ values as validation of this approach.

Methods

Chemicals and materials

Quinacrine hydrochloride, pyronaridine tetraphosphate, and tilorone dihydrochloride (BOC Sciences, Shirley, NY), bafilomycin A1, and chloroquine diphosphate (Sigma, St. Louis, MO) were dissolved in either DMSO or water as 10 mM stock solutions and were stored at -20°C. The nucleus staining dye, Hoechst 33342, CellMask Deep™ Red cytoplasmic/nuclear stain, NHS-Alexa-488 dye, the Dual-Glo® Luciferase Assay System and CytoTox 96™ assay kit were purchased from Promega (Promega, Madison, WI). The modified MTT assay Cell Counting Kit 8 was procured from Dojindo Molecular Technologies (Dojindo Molecular Technologies, Gaithersburg, MD). The 96-well high-content imaging plates were from Corning (Corning Life Sciences, MA). The Opera QEHS confocal imaging plate was obtained from BD Biosciences, Franklin Lakes, NJ and the 96-well white-walled tissue culture plates were from BD Biosciences (BD Biosciences, Franklin Lakes, NJ) and the 96-well white-walled tissue culture plates were from Corning (Corning Life Sciences, MA). The Opera QEHS confocal imaging reader, Acapella™ and Definiens™ image analysis packages were purchased from PerkinElmer (PerkinElmer, USA). Image acquisition was done using Nikon TI eclipse high content imaging enabled microscope running NIS elements high content imaging software (version 4.30.02).
Machine learning
868 molecules from the viral pseudotype entry assay and the EBOV replication assay from a recent publication were made available as an sdf file. Salts were stripped and duplicates removed using Discovery Studio 4.1 (Biovia, San Diego, CA). For each assay, compounds with IC<sub>50</sub> values less than 50 μM were selected as actives. All other compounds were classed as inactives. Models were generated using a standard protocol with the following molecular descriptors: molecular function class fingerprints of maximum diameter 6 (FCFP<sub>6</sub>)<sup>17</sup>, AlogP<sup>27</sup>, molecular weight, number of rotatable bonds, number of rings, number of aromatic rings, number of hydrogen bond acceptors, number of hydrogen bond donors, and molecular fractional polar surface area. Models were validated using five-fold cross validation (leave out 20% of the database).

Bayesian, Support Vector Machine and Recursive Partitioning Forest and single tree models built with the same molecular descriptors in Discovery Studio were compared. For SVM models, we calculated interpretable descriptors in Discovery Studio and then used Pipeline Pilot to generate the FCFP<sub>6</sub> descriptors followed by integration with R<sup>28</sup>. RP Forest and RP Single Tree models used the standard protocol in Discovery Studio. In the case of RP Forest models, ten trees were created with bagging. Bagging is short for “Bootstrap AGGregation“. For each tree, a bootstrap sample of the original data is taken, and this sample is used to grow the tree. RP Single Trees had a minimum of ten samples per node and a maximum tree depth of 20. In all cases, 5-fold cross validation or leave out 50% x 100 fold cross validation was used to calculate the Receiver Operator Curve (ROC) for the models generated.<sup>29</sup>

Open Bayesian models
Open Bayesian models for the Ebola datasets were developed using open source software<sup>30-41</sup> and loaded into the Mobile Molecular Data Sheet (MMDS (http://molmatinf.com/)) and then the two models were used to score the three compounds selected by the earlier models. These two models are also openly accessible (http://molsync.com/ebola/) and can be uploaded into MMDS in order to score molecules of interest.

Pharmacophore mapping
Pyronaridine was mapped to the recently published pharmacophore<sup>36</sup> derived from Ebola in vitro inhibitors amodiaquine, chloroquine, clomiphene and toremifene in Discovery Studio Vers 4.1 and a fit score was generated.

In vitro testing
Recombinant, infectious Ebola virus encoding green fluorescent protein (GFP) was used for testing efficacy of compounds and was originally provided by Dr. Heinz Feldmann, Rocky Mountain Laboratories. The strain that was used has the GFP gene inserted between the VP30 and VP24 genes. All viral infections were done in the BSL-4 lab at Texas Biomedical Research Institute. Briefly, 4,000 HeLa cells per well were grown overnight in 384-well tissue culture plates, the volume of DMEM (Fisher scientific, Cat#MT10017CV) culture medium supplemented with 10% fetal bovine serum (Gemini Bio-Products, Cat#100106) was 25 μL. On the day of assay, test drugs were diluted to 1 mM concentration in complete medium. 25 μL of this mixture was added to the cells already containing 25 μL medium to achieve a concentration of 500 μM. All treatments were done in triplicates. 25 μL of medium was removed from the first wells and added to the next well. This type of serial dilution was done 12 times and treated cells were then incubated at 37°C in a humidified CO<sub>2</sub> incubator for 1 hour. Final concentrations of 250, 125, 62.5, 31.25, 15.62, 7.81, 3.9, 1.9, 0.97, 0.48, 0.24 and 0.12 μM were achieved upon addition of 25 μL of infection mix containing Ebola-GFP virus, Bafilomycin at a final concentration of 10 nM was used as a positive control drug. Infections were done to achieve a MOI of 0.05 to 0.15. Infected cells were incubated for 24 hours. 24 hours post-infection cells were fixed by fixing the plates in formalin for 24 hours at 4°C. Fixed plates were decontaminated and brought out of the BSL-4. Formalin from fixed plates was decanted and plates were washed thrice with PBS. EBOV-infected cells were stained for nuclei using Hoechst at 1:50,000 dilution and plates were imaged. Nuclei (blue) and infected cells (green) were counted using CellProfiler software (Broad Institute)- Version 2.1.1. Total number of nuclei (blue) was used as a proxy for cell numbers and a loss of cell number was assumed to reflect cytotoxicity. Concentrations where total cell numbers were 20% less than the control were rejected from the analysis.

Results

Machine learning
Using 5-fold cross validation the Bayesian approach (Data S1 and Data S2) performed the best for the EBOV replication data and was equivalent for the RP Forest approach (Table 1) and was better than SVM (Data S3 and Data S4) for the pseudotype data. The Open Bayesian models had ROC scores slightly lower than the Bayesian models built with Discovery Studio. A more exhaustive cross validation for the Bayesian models is the ‘leave out 50% repeated randomly 100 times’ which produced ROC values greater than 0.8 and were comparable to the 5-fold cross validation data. This indicated the models are stable. For the EBOV pseudotype assay, alkoxethylaminio was a common feature amongst active compounds in the training set, as were 1,3-diaminopropyl and saturated six-member heterocycles with an oxygen and perhaps an additional heteroatom in the ring (Figure 1A). Training set inactives commonly featured carboxylic acids, N,N’-disubstituted ureas, secondary and tertiary amides, pyrazoles, aromatic sulfonamides, tertiary cycloptanols, 1,2-mercaptoethanol, and penams (Figure 1B). For the replication assay training set, active features included piperazine, phenothiazine, tertiary amines, and alkoxethylaminio (Figure 2A).

<table>
<thead>
<tr>
<th>Models (training set 868 compounds)</th>
<th>RP Forest (Out of bag ROC)</th>
<th>RP Single Tree (With 5 fold cross validation ROC)</th>
<th>SVM (with 5 fold cross validation ROC)</th>
<th>Bayesian (with 5 fold cross validation ROC)</th>
<th>Bayesian (leave out 50% x 100 ROC)</th>
<th>Open Bayesian (with 5 fold cross validation ROC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebola replication (actives = 20)</td>
<td>0.70</td>
<td>0.78</td>
<td>0.73</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td>Ebola Pseudotype (actives = 41)</td>
<td>0.85</td>
<td>0.81</td>
<td>0.76</td>
<td>0.85</td>
<td>0.82</td>
<td>0.82</td>
</tr>
</tbody>
</table>
Figure 1. A. Active and B. Inactive features for the Discovery Studio pseudotype Bayesian model.
**Figure 2.** A. Active and B. Inactive features for the Discovery Studio EBOV replication model.
Inactive features included secondary amides, disubstituted amines, cyclopropylmethyl, carboxylic acids, 1,3-oxathiolanes, tertiary alcohols, phenethyl, and penams (Figure 2B). An actives feature common between both assays/models was alkoxyethylamino. Inactives features in common between both were carboxylic acids, secondary amides, penams and tertiary alcohols, which may relate to properties which prevent the molecules from accessing cellular sites of viral activity.

The MicroSource Spectrum set of 2320 compounds was then scored with both Bayesian models (Data S5). Predicted actives were quantified as to their chemical similarity, or distance, from molecules in the training set. When excluding compounds in the training set (as well as antipsychotics and other less desirable CNS active compounds), those scoring highly were considered most interesting and included the antiviral tilorone, the antimalarials quinacrine and pyronaridine (Figure 3). Perhaps not surprisingly, tertiary amines scored particularly well. These molecules were also scored with the open Bayesian models (Data S6) and all replication models scored the compounds highly (values close to or greater than 1). None of these three compounds has been described in recent reviews of small molecules with activity against EBOV14–16, to our knowledge.

Pharmacophore
The MicroSource set had previously been screened with the published Ebola common feature pharmacophore26,27, using the van der Waals surface of amodiaquine (which was more potent than chloroquine3) to limit the number of hits retrieved22–44. Two of the three selected – compounds quinacrine (fit score 2.59) and tilorone (fit score 3.65) – were retrieved previously. We therefore used the ligand pharmacophore mapping to map pyronaridine to the pharmacophore without the van der Waals surface (Figure 4, Fit score of 3.60 suggested this was a good match to pharmacophore features).

In vitro testing
The three selected compounds were tested in vitro alongside the positive control chloroquine which gave an expected dose response curve (Figure 5, Table 2). Quinacrine, pyronaridine and tilorone, were tested in vitro and had EC₅₀ values of 350, 420 and 230 nM, respectively which were lower than for chloroquine 4.0 μM. Several images created in this study illustrate the results of this high content screen (Data S7).

Discussion
Our recent work on neglected diseases has shown that we can learn from existing assay datasets. Specifically we have previously analyzed large datasets for Mycobacterium tuberculosis to build machine learning models that use single point data, dose-response data43,45, combine bioactivity and cytotoxicity data (e.g. Vero, HepG2 or other model mammalian cells)28,29,46 or combinations of these sets47,48. These models in turn have been validated...
Table 2. Effect of drug treatment on infection with Ebola-GFP (n=3 per compound). The cytotoxicity of compounds are represented as a 50% cytotoxicity concentration (CC$_{50}$) estimated by the lowest concentration of drug that produced ≥ 50% loss in cell number by nuclei counting.

<table>
<thead>
<tr>
<th>Compound</th>
<th>EC$_{50}$ (μM) [95% CI]</th>
<th>Cytotoxicity CC$_{50}$ (μM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chloroquine</td>
<td>4.0 [1.0–15]</td>
<td>250</td>
</tr>
<tr>
<td>Pyronaridine</td>
<td>0.42 [0.31–0.56]</td>
<td>3.1</td>
</tr>
<tr>
<td>Quinacrine</td>
<td>0.35 [0.28–0.44]</td>
<td>6.2</td>
</tr>
<tr>
<td>Tilorone</td>
<td>0.23 [0.09–0.62]</td>
<td>6.2</td>
</tr>
</tbody>
</table>

Figure 5. Effect of drug treatment on infection with Ebola-GFP. Cells were treated and then challenged with Ebola virus encoding GFP. Infection efficiency was calculated as infected cells (expressing GFP)/total cells and normalized to infection efficiency seen in the untreated control. Shown is one representative experiment where each point is the average of 3 independent measurements of infection +/- standard deviation. Dose response curves were fitted by non-linear regression.

with additional non-overlapping datasets, demonstrating that it is possible to use publically accessible data to find novel in vitro active antituberculars. We have also applied the same approach recently to identify a molecule with in vitro and in vivo activity against T. cruzi$^{30}$. In the current study we found that different machine learning methods produced similar 5-fold cross validation data, although the Bayesian models had ROC values consistently above 0.80, which is preferable. One of the issues with computational models is that they are rarely accessible to others due to the commercial software licensing requirements. We have previously showed that models built with open source tools can produce validation statistics comparable to commercial modeling tools$^{39}$. We recently made “function class fingerprints of maximum diameter 6” (FCFP6) and “extended connectivity (ECFP6) fingerprints,” open source and have described their implementation with the Chemistry Development Kit (CDK)$^{40}$ components. In addition we described an open source Bayesian algorithm that can be used with these descriptors$^{38,40}$. One way to make such models more accessible is to use mobile devices for their delivery and we have developed cheminformatics mobile apps$^{31,33–35}$. Several of these apps combine Bayesian models and open source fingerprint descriptors to enable models that can be used within a mobile app (TB Mobile, MMDS, Approved Drugs and MolPrime). This enables a scientist to select a molecule and score it with models. In the current study we used the same training sets for the anti-EBOV activity using replication and pseudotype screening data to build open source models that we can share with the community (http://molsync.com/ebola/).

The Bayesian models allowed us to select three compounds from the MicroSource compound library that scored highly and were not in the model training sets. The Open Bayesian models also scored the three hits favorably, which bodes well for screening other
compounds of interest. Two of these molecules had also been identified with our earlier pharmacophore model which may be indicative of binding to VP35. When tested in vitro the three compounds possessed EC50 values 230–420 nM, much lower than the positive control chloroquine (EC50 4.0 μM) used in this study and identified previously. Tilorone is an investigational agent that has been known for over 40 years as an antiviral and is an inducer of interferon in mice. It has been shown to possess a broad array of biological activities including cell growth inhibition in PC3 CDK5dn prostate cancer cells (IC50 8–12 μM), inhibition of Primase DnaG from Bacillus anthracis (IC50 7.1 μM), in a mouse model of pulmonary fibrosis it decreased lung hydroxyproline content and the expression of collagen genes, α7 nicotinic receptor (nAChR) agonist activity (K, 56 nM), activated human alpha7 nAChR with an EC50 value of 2.5 μM, radioprotective activity, potent modulation of HIF-mediated gene expression in neurons with neuroprotective properties and induction of the accumulation of glycosaminoglycans, delay infectious prion clearance, and prolong prion disease incubation time. Quinacrine is an old antimalarial drug now more widely used as an antiprotozoal for the treatment of giardiasis and as an anthelmintic. Pyronaridine is a potent antimalarial (IC50 13.5 nM), has activity against Babesia spp., is active in vitro (EC50 225 nM) and in vivo (85.2% efficacy 4 days treatment at 50 mg/kg) against T. cruzi and is a P-glycoprotein inhibitor. Pyronaridine is used in combination with artemunate in the European Medicines Agency approved Pyramax which has performed well in clinical trials for malaria. As this molecule has already been approved this may have a more direct path to clinical testing if it is found to be active in standard animal models infected with the Ebola Virus.

As stated before in perspectives by us and others, the fact that approved drugs may be repurposed for other diseases should not be viewed as a negative aspect of the small molecules, belying undesirable target promiscuity. Instead, we prefer to reference recently published crystallographic analyses demonstrating that small molecules may bind multiple proteins in different types of binding sites and with distinct conformations to ultimately facilitate molecular repurposing. While it would be most desirable to repurpose an approved drug and, thus, catapult a discovery effort into a Phase II trial, one should not ignore the significance of utilizing the discovery of a new use for an old drug to seed efforts in the lead optimization phase. Such an expedited program would be expected to have a high probability of producing novel small molecules, closely related to or inspired by the drug, with the opportunity to translate quickly to clinical trials.

In summary, this study has added to the previous work that identified several FDA approved compounds active against EBOV in vitro. Future work may include identification of targets using computational or experimental approaches. We propose that these three molecules may warrant further evaluation in vivo as they are significantly more active than chloroquine. Larger scale virtual screening could be performed on the millions of commercially available molecules or more complete sets of approved and older no longer used drugs than have already been screened. These computational efforts can then prioritize molecules for testing. Such an approach may be a useful way to leverage the HTS data that has already been developed at great cost. In this study we have focused on just the data from a single group but it may also be possible to combine this with the data from the other high throughput screens to provide a much larger training set. There is also the opportunity to apply many different computational approaches beyond those described here to identify whole cell active compounds against EBOV. Ultimately, we should be able to identify additional compounds that could be immediately useful to treat patients with the disease while we await the approval of a vaccine.

Data availability
Supplemental data contains results from Bayesian models and SVM models as well as the output of predictions with Bayesian models and open Bayesian models.

The training sets used in the models are available as SDF files (http://molsync.com/ebola/).

Author contributions
Conceived and designed the experiments: S.E., M.A., R.A.D., P.B.M
Performed the experiments: S.E., M.A., R.A.D., P.B.M
Analyzed the data: S.E., J.S.F., M.A., R.A.D., P.B.M
Contributed reagents/materials/analysis tools: S.E., A.M.C., M.A., R.A.D., P.B.M
Wrote the manuscript: S.E., J.S.F., A.M.C., M.A., R.A.D., P.B.M
All authors have seen and agreed to the final content of the manuscript.

Competing interests
S.E. works for Collaborations in Chemistry, and Collaborations Pharmaceuticals, Inc. and S.E. and A.M.C. consult for Collaborative Drug Discovery Inc.

Grant information
The author(s) declared that no grants were involved in supporting this work.

Acknowledgments
SE kindly acknowledges Biovia for kindly providing Discovery Studio and Dr. Megan Coffee and Dr. Christopher Southan for initially stimulating interest in this topic.
Supplementary materials

Supplemental data S1–S4 and S6, S7.


http://dx.doi.org/10.5256/f1000research.7217.s110159

Click here to access the data.

Supplemental data S5.

MicroSource predictions with Bayesian models xl file.

Click here to access the data.

References

30. Ekins S, de Siqueira-Neto JI, McCaI Li, et al.: Machine Learning Models and

Published Abstract | Publisher Full Text | Free Full Text


Open Peer Review

Current Peer Review Status: ✔️ ✔️

Version 3

Reviewer Report 18 January 2017

https://doi.org/10.5256/f1000research.11510.r19459

© 2017 Glisic S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sanja Glisic
Center for Multidisciplinary Research, Vinča Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 03 December 2015

https://doi.org/10.5256/f1000research.7773.r11173

© 2015 Chakraborty S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sandeep Chakraborty
Plant Sciences Department, University of California, Davis, CA, USA

Ekins et al. have presented a crisp and lucid manuscript on a very relevant topic. They have presented a methodology that implements machine learning techniques to learn from known active and inactive compounds (an ever increasing set, that will tend to provide improved results as time goes by), and score a larger set of compounds (MicroSource Spectrum set of 2320 compounds). The in silico methodology described here provides an excellent method to quickly screen known compounds for possible therapies (against Ebola in particular), and other viruses in general. Finally, they demonstrate (in vitro) the increased effectiveness of three compounds - the antiviral tilorone and the antimalarials quinacrine and pyronaridine
effectiveness of three compounds - the antiviral tilorone and the antimalarials quinacrine and pyronaridine - in comparison to the known active chloroquine (albeit at a higher cytotoxicity) in inhibiting viral infection of HeLa cells. Further, their efforts in ensuring open-access to such tools is commendable as the next pathogen caused humanitarian crisis looms in some nations.

The biggest open question is how good are the molecular descriptors, and how much of this is serendipitous. For example, I find it hard to believe that molecular weight and the number of rotatable bonds can be good predictors of drug-protein interactions (although I may be wrong). I have been investigating promiscuous ligand protein interactions for a couple of years on a molecular basis (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0032011, http://f1000research.com/articles/2-286/v3). One interesting example (unpublished) is suramin used in the treatment of African sleeping sickness (African trypanosomiasis) and river blindness (onchocerciasis), infections caused by parasites. Suramin binds eight non-homologous proteins in the PDB database, through different parts of the molecule and in binding sites that share little similarity in residues involved. Also, the molecule (in addition to the protein) undergoes conformational changes, underscoring the difficulty of computational methods to model such interactions. In the face of such data, the m/c learning models appear too simplistic.

Also, it is not completely clear why the 23, 31 and 34th compound was chosen from the Table S5, which is ordered on column H (all three have amines, don't the previous ones have it?).

Some minor comments:
1. It would be an interesting case study to evaluate how favipiravir (which I understand is yet to be clinically approved in the US, but approved in Japan) and BCX4430 would rank through the m/c learning methodology.

2. It would be good to have a set of images, and the corresponding nuclei counts obtained from CellProfiler. Is there a way to quantify the color green as a measure of viral infection?

3. AlogP as a molecular descriptor has not been explained (page 3).

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 19 Dec 2015

Sean Ekins, Collaborations in Chemistry, Fuquay-Varina, USA

We thank the reviewer for their constructive feedback.

The approach we have taken uses FCFP_6 fingerprints as well as 8 interpretable descriptors, and therefore the models do not depend on molecular weight and rotatable bond number. On the whole this approach has been remarkable useful for predicting whole cell activity as we described for mycobacterium tuberculosis, T. Cruzi and now Ebola. In all cases we are not considering a single target. This suggests the machine learning approach and descriptors used can discern active molecule features from those that are inactive and identify new molecules.

The three compounds were chosen as those above them in the list were either compounds in the training set or antipsychotics and other CNS acting compounds etc. which were deemed less
The model could certainly be used to predict additional molecules. These two suggested by the reviewer are structurally distinct from any of the actives in our current training set. We didn’t identify any of the classical antiviral polymerase-look ing compounds in our screening against Ebola. However we have previously collated and described many other diverse compounds active against Ebola in vitro as described in this manuscript. Perhaps the next step would be to utilize all of the different HTS screening data to build a combined model that considers this structural diversity and may overcome limitations in the current models.

Response: we have now added a new figure with cell images (S7).

Response: AlogP is a widely used measure of hydrophobicity and we have refered to its use along with the other descriptors in the previous machine learning papers.

**Competing Interests:** No competing interests were disclosed.
We thank the reviewer for their constructive feedback

Response: We are unsure which resources the reviewer is referring to and how these would help identify the antiviral target. We have tried to extensively describe the known activities of the three compounds against various targets outside of viruses. In addition we have previously suggested such antimalarials with Ebola activity may dock into VP35. Preliminary docking results suggest pyronaridine may dock into the same site which is also indicated by the pharmacophore provided already (fig 4). We have added the statement “Future work may include identification of targets using computational or experimental approaches.” to the discussion.

Competing Interests: No competing interests were disclosed.